AI Communications 28 (2015) 87–101
DOI 10.3233/AIC-140617
IOS Press

Automatic non-verbal communication skills analysis: A quantitative evaluation

Álvaro Cepero [a,*], Albert Clapés [a,b] and Sergio Escalera [a,b]
[a] *Departament Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via de les Corts Catalanes, Barcelona, Spain*
E-mails: acepero13@gmail.com, aclapes@cvc.uab.cat, sergio@maia.ub.es
[b] *Computer Vision Center, Campus UAB, Bellaterra, Barcelona, Spain*

Abstract. The oral communication competence is defined on the top of the most relevant skills for one's professional and personal life. Because of the importance of communication in our activities of daily living, it is crucial to study methods to evaluate and provide the necessary feedback that can be used in order to improve these communication capabilities and, therefore, learn how to express ourselves better. In this work, we propose a system capable of evaluating quantitatively the quality of oral presentations in an automatic fashion. The system is based on a multi-modal RGB, depth, and audio data description and a fusion approach in order to recognize behavioral cues and train classifiers able to eventually predict communication quality levels. The performance of the proposed system is tested on a novel dataset containing Bachelor thesis' real defenses, presentations from an 8th semester Bachelor courses, and Master courses' presentations at Universitat de Barcelona. Using as groundtruth the marks assigned by actual instructors, our system achieves high performance categorizing and ranking presentations by their quality, and also making real-valued mark predictions.

Keywords: Social signal processing, human behavior analysis, multi-modal data description, multi-modal data fusion, non-verbal communication analysis, e-Learning

1. Introduction and related work

Nowadays, the society is demanding new kinds of competences to its citizens and especially its professionals. With the implementation of the Bachelor degree in the European Higher Education Area, the concept of competences became even more important in the educational field. One of the main goals of this plan is to provide specific skills and competences to the student. Oral expression and communication are among the most relevant competences 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 within the top five skills required in potential hires. Nonetheless, 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 evaluate and provide the necessary feedback that can be used in order to improve these communication capabilities and, therefore, learn how to express ourselves better. However, we first should define the criteria and methods needed to measure our verbal and non-verbal communication and provide us the necessary feedback. In this sense, we propose an automatic system capable of providing both evaluation and feedback on communication skills within the e-Learning scope.

e-Learning (from Electronic Learning) is a very broad term. Typically, it refers to the usage of all kinds of information and communication technologies (ICT) in learning processes or, more particularly, in the educational field. In the literature, we find a fine-grained categorization of different e-Learning paradigms: Computer-based Instruction (CBI), Computer-based Training (CBT), Computer-assisted Instruction (CAI), Computer-assisted Training (CAT), and Internet-based Instruction (IBI) or Web-based Instruction (WBI) as extensions of CBI.

[*] Corresponding author: Alvaro Cepero, Departament Matemàtica Aplicada i Anàlisi, Universitat de Barcelona, Gran Via de les Corts Catalanes 585, 08007, Barcelona, Spain. E-mail: acepero13@gmail.com.

CBI and CBT have been studied for decades [45]. From both past studies [11,24,33] and more recent ones [10,18] researchers found that the application of these learning paradigms produces a positive effect in terms of attitude towards the learning processes and computers, and reduces the amount of instruction time. In the state-of-the-art, we find a wide range of application of these paradigms. For instance, in [36], CBI techniques are applied in order to improve general problems solving skills, whereas in [29,39], a more specific kind of instruction is provided: the training required by the medical staff dealing with patients affected by autism or terminal illnesses. Now, in the era of Internet, in which great amounts of information and knowledge can be instantly available, CBI is evolving to the more general IBI (or WBI) paradigm. Today, there are higher education courses completely online, accessible for everyone via multimedia formats such as videolectures; some of them are offered by universities (*MitOpenCourseware* or *Coursera*) and others not (*KhanAcademy* or *Udacity*). In fact, these online materials do not necessarily have to replace the face-to-face instruction. Many schools, high schools, and universities are making use of virtual education platforms, involving in them instructors and students. This emerged combination of face-to-face and online learning is known as *blended learning* [3].

Computer-assisted paradigms, unlike computer-based ones, give the initiative to the learner and the presence of the computer tends to be more unobtrusive, which in some cases turns out to be more convenient for the instruction goals' accomplishment. CAI/CAT can be used to train everyday life-related competences, for instance, in the beginning readers training [2] or in the prosody training in new language learners [21]. Also the instruction or training of professional competences can be assisted by a computer, like the surgery skills teaching [19]. People affected by particular conditions (elderly or illness) can also benefit from e-Learning. In [20], the authors intended to improve the life of elder people improving their cognitive performance, while in [44] the authors focused in the assistance of physically conditioned patients to train their wheelchair's moving ability.

Our proposed tool fits in the Computer-assisted Training e-Learning paradigm. The training of the non-verbal communicative competence could be aided providing auditory and/or visual feedback both during the non-verbal communication act and/or posteriorly. In this context, we found previous studies evaluating the non-verbal communication quality. In [22] medical students' are evaluated in terms of non-verbal communication during their clinical examinations exercises or patients' interviews. In fact, the indicators the authors of the study defined are quite similar to the ones defined and analyzed automatically in our proposal. However, in the work of [22] the evaluation of the non-verbal communication is not performed automatically, but by human observers.

Whereas the verbal communication can be evaluated as successful or not given some quite clear criteria – such as amount of knowledge transmission, richness of the language expression, discourse coherence and cohesion, and so forth – the non-verbal communication instead is often quite explicit to a human observer. The non-verbal signals are implicit in the human behavior and relatively subtle, despite the huge amount of information they provide about the communication and the communicating subject. Besides, the non-verbal signals 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 and its impact to the people. For instance, the study of [1] discussed the teacher's non-verbal behavior and its effects on higher education students.

In this context, and from the point of view of Artificial Intelligence (AI), we are at the beginning of a long way to go. While the verbal communication started to be automatically analyzed many years ago by the Natural Language Processing AI's subfield, the study of the non-verbal communication from the AI's point of view is relatively recent. In the past decade, one can find works in the field of social robotics building physical embodied agents capable of both emulating the humans' non-verbal behavior and perceiving information from the humans' non-verbal cue [5]. In [4], a social robot's head was emulating not only the infants' non-verbal behavior but also their perception abilities, the ones necessary to interpret the human non-verbal communication signals from visual and auditory cues. So much effort has been put in trying to provide to physical or virtual embodied agents the ability to emit meaningful and effective non-verbal communicative signals [8], since it has been shown this changes how the people interact with them [2]. On the other hand, we also need to know how to provide to the machines the ability to receive and interpret non-verbal communication signals emitted by a human beings.

In order to build machines capable of socially behave as we do and to be able to interact with us naturally – as we do with other people – we need in first place to better understand how the humans communicate to each other and develop methods for the automatic interpretation of these social signals. The Social Signal Processing field is the one that focuses on the automatic analysis of interactions among subjects exchanging communication signals (roles, signs of dominance, attitude towards the other subjects, and so on) by means of different sensors, such as microphones or cameras, applying some kind of pattern recognition strategies. Those tasks include the automatic analysis of non-verbal behavioral cues, that psychologists have grouped into five major classes [42], 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.

Several works have been recently published in the Social Signal Processing field in which the non-verbal communication of subjects group interactions are analyzed [41]. Most of these works focus on audio analysis, and main goals are based on detecting/identifying dominance, influence, and leadership. In the work of [30], the authors present the implementation of a platform for measuring and analyzing human behavior in organizational face-to-face settings using wearable electronic badges. The authors of [26] present the recognition of group actions in meetings with approaches based on Hidden Markov Models modeling audiovisual-featured observations. Other recent approaches for dominance analysis in group interrogations have been also proposed [14]. The work of [31] presents a Bayesian framework that models dominance skills based on audio input sources. In [35], the authors model a multi-modal audio-video system to recognize leadership in group interactions. The system is defined based on simple multi-modal features under controlled face-to-face interaction environments. In a similar scenario, the proposal of [25] defines multi-modal cues for the analysis of communication skills in an upper body setup. A more general purpose approach is presented in [28], where prosodic features are computed to define a set of relevant traits of subjects in oral communication settings. In our case, the main objective of the analysis is to be able, not to detect or segment social events, but to quantitatively evaluate the level of quality of the non-verbal communication and to, later, have the possibility to embed it in an e-Learning tool to assist in the training of this competence. Very few works have been reported on the analysis of non-verbal communication as a competence skill in e-Learning scenarios. The authors of [38] present 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. Therefore, to the best of our knowledge, it does not exist a tool capable of measuring automatically the level of quality of the non-verbal communicative act, similar to the one proposed in this paper.

In this work, we present a multi-modal RGB-Depth-Audio system for non-verbal communication analysis. The system, firstly, extracts low-level per-frame features on each of the modalities. In RGB and depth modalities, face detection and skeleton joints are tracked, whereas in the audio modality voice activity is detected. From the low-level features, a set of high-level behavioral indicators per-sequence are computed and used to describe each presentation. Once each sequence has been globally described, a quantifying evaluation of the quality level of the presentations is performed using different state-of-the-art statistical learning algorithms (binary classification, multi-class classification, ranking, and regression). In addition, a study of the most relevant features in order to evaluate the quality level is also presented. Quantitative results on a novel multi-modal dataset of university students' presentations show accurate measurements of the proposed system and its reliability to be used in the training routine of the non-verbal communication competence.

The rest of the paper is organized as follows. Section 2 presents the multi-modal RGB-Depth-Audio system for communication competence analysis. In Section 3, the proposed methodology is evaluated on a novel dataset of student presentations. Finally, Sections 4 and 5 conclude the paper and discusses future lines of research, respectively.

2. The multi-modal RGB-Depth-Audio system for communication competence analysis

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 RGB-Depth-Audio data which defines the input audio source, face tracking system and the skeletal body joints model, and second, the processing of these low-level features into high-level features to build the features that codify the user's behavior. These indicators are then used to train robust statistical classifiers able to predict the quality of the presentations given a groundtruth defined by experts. The different modules of our system are shown in Fig. 1 and described next.

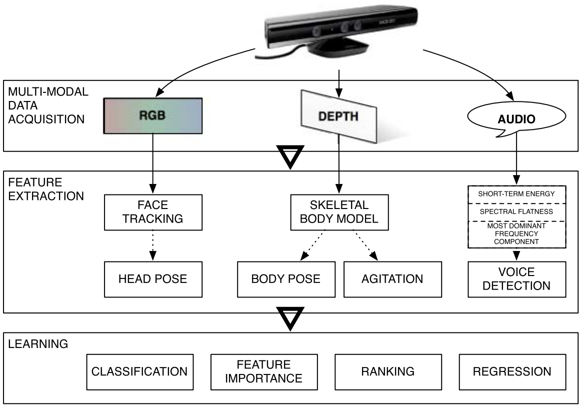

Fig. 1. System modules for non-verbal communication analysis. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.))

2.1. *Low-level features*

In order to analyze the communicative behavior of the user towards the audience we defined a set of low and high-level features. The low-level features are those characteristics that are extracted directly from the Microsoft® Kinect™ SDK, that is the RGB data, depth and the raw audio. The way Kinect™ provides multimodal data relies on three hardware components working together:

- *Color VGA video camera*. It is used on the face detection process and face tracking algorithm. This component acts as a regular camera and it captures images around 30 frames per second, and projects at a 640 × 480 pixels resolution.
- *Near-infrared light sensor and emitter*. The Kinect™ near-infrared light emitter projects a structured/codified matrix of points through the environment. Then, each depth pixel is computed by sampling the derivative of the higher resolution infrared image taken in the infrared sensor. This value is inversely proportional to the radius of each projected infrared dot, which is linearly proportional to the actual depth.
- *Microphone array*. The Kinect™ presents a microphone array that consists of four separate microphones spread out linearly at the bottom of the device, 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 addition, speech can be recognized in a large room where the speaker's lips are more than a few meters from the microphone array. In our system, the distance between the speaker and the device is less than 3.5 meters.

2.1.1. *RGB-depth features*

We use the color or RGB data to perform face detection using the well-known Viola and Jones method [43]. This method exhaustively searches for faces in an image using an incremental sliding window to cover all the different regions of different sizes. In each region, a set of Haar-like features[1] particularly selected for the task of face detection is very efficiently computed using integral images[2] and used to build a feature vector characterizing the region. Then, the feature vector extracted for that particular region is classified using the adaptive boosting classifier (AdaBoost) [16] to determine whether the region contains a face or not. AdaBoost is a meta-algorithm which builds a strong classifier from a cascade of weak classifiers; in other words, a set of simpler classifiers applied sequentially to each candidate face region. The first stages of the cascade separate the easier examples, and those more difficult are propagated as false positives to be dealt with in posterior and more constraining weak classifiers. In addition, the color is used together with the depth information in the face tracking algorithm that localizes 121 facial 3D landmarks based on the Active Appearance Models [12] from which we later compute the head pose.

On the other hand, we use the depth information to perform human body segmentation and skeleton joints detection and tracking. This skeletal model will yield the world coordinates of the user in real time. The Microsoft® Kinect™ SDK defines 20 keypoints to model a human skeleton. In our system, the coordinates of hands, wrists, arms, elbow hip, shoulders and head are the only ones considered. We use Random Forest of [37] to segment the human body and compute the skeletal model. The body part classification at pixel level is performed first by describing each depth pixel

[1] A Haar-like feature characterizes an image region by defining a disposition of rectangles within the region and calculating the difference of the sums of intensities in the rectangular subregions [32].

[2] The integral image can be used to compute Haar-like features in constant time [43]. Each pixel in the integral image is the accumulation of the intensities of all the upper-left pixels. In such a way, the sum of the intensities in a region can be calculated as: $I(A) - I(B) - I(C) + I(D)$, being $I(\cdot)$ the intensity of a pixel in the integral image, and A, B, C and D the bottom-right, bottom-left, upper-right, and upper-left pixels respectively.

x (in a dense depth image **D**) as follows:

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

where $\theta = (\mathbf{u}, \mathbf{v})$, and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^2$ are two random offsets, depth invariant. Thus, each θ determines two new pixels relative to **x**, the depth difference of which accounts for the value of $f_\theta(\mathbf{D}, \mathbf{x})$. Using this set of random depth features, a random forest of randomized decision trees is trained, where each tree consists of split and leaf nodes (the root is also a split node). From the classification of a pixel in each of the τ trees, a discrete probability distribution of body parts memberships is obtained. Finally, the discrete probability distributions are averaged 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 l_i is a certain body part, $P_j(l_i|\mathbf{D}, \mathbf{x})$ is the distribution of body parts of the pixels class 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 has been applied, a mean shift is used to estimate human joints and represent the body in skeletal form.

2.1.2. *Audio features*

From the raw audio obtained from the Kinect™ we compute three types of low-level features per frame [27]. The first feature is the widely used Short-term Energy. Energy is the most common feature for speech/silence detection. The second feature is the Spectral Flatness measure, which 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 voice and non-voice frames. These low-level features will be used later to compute the 'speaking' high-level behavioral indicator.

2.2. *High-level features*

The high-level features or meta-characteristics are computed globally for each sequence from the low-level features described in the previous section in order to define the speaker's communication indicators. Further we present the set of psychology-based behavioral indicators, which we consider in our framework. The values for the different parameters involved in the computation of the indicators are specified in Section 3.2.

(1) *Frontal-facing*. The average number of frames in which the subject is facing the audience placed behind the acquisition device. Face detection and face tracking are performed in order to analyze whether the user is looking at the audience. If a frontal view of a face is detected, the face tracking algorithm computes the facial landmarks from which the nose's vector is computed. We consider the subject is looking at the audience if the nose's vector falls in a cone of sight (with a certain amplitude) directed to the audience. The following formula expresses this feature computation:

$$f_1 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\left\{\arccos\left(\frac{\hat{\mathbf{n}}^t_{nose}}{-\hat{\mathbf{z}}}\right) \leqslant \alpha\right\}, \tag{3}$$

where T is the total number of frames in the presentation, $\mathbb{1}\{\cdot\}$ is the indicator function, which takes the value 1 if the condition contained is fulfilled or 0 otherwise, $\hat{\mathbf{n}}^t_{nose}$ is a unit vector expressing the face toward-looking direction at time t, and $\hat{\mathbf{z}}$ is the unitary vector representing the Kinect's viewpoint direction. Hence, we consider the user is looking at the public if the angle formed by the two vectors, $\arccos(\hat{\mathbf{n}}^t_{nose}/-\hat{\mathbf{z}})$, is lower or equal than a certain angular distance α.

(2) *Crossed arms*. The average number of frames in which the user has his/her arms crossed. In order to determine if the arms are crossed, we check some precomputed distances among body joints expressing the length of some body limbs (Fig. 2). We consider the arms are crossed if both hands are closer to the opposite shoulder and if those distances are approximately the length of

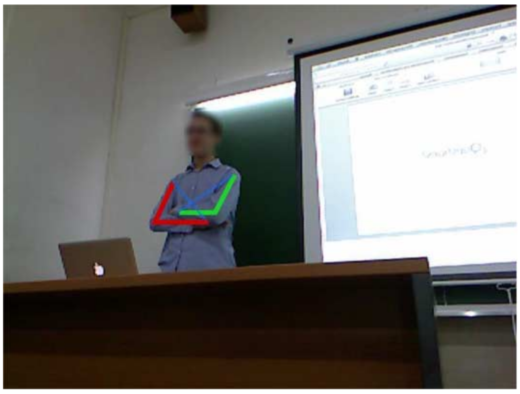

Fig. 2. Student with his arms crossed. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

the half upper arms:

$$
\begin{aligned}
f_2 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\big\{ &\big(d^t_{\text{hand}_\text{L},\text{shoulder}_\text{R}} \\
&< d^t_{\text{hand}_\text{L},\text{shoulder}_\text{L}}\big) \\
&\wedge \big(d^t_{\text{hand}_\text{R},\text{shoulder}_\text{L}} < d^t_{\text{hand}_\text{R},\text{shoulder}_\text{R}}\big) \\
&\wedge \big(d^t_{\text{hand}_\text{L},\text{shoulder}_\text{R}} < h_{\text{arm}_\text{R}}\big) \\
&\wedge \big(d^t_{\text{hand}_\text{R},\text{shoulder}_\text{L}} < h_{\text{arm}_\text{L}}\big)\big\}, \qquad (4)
\end{aligned}
$$

where $d^t_{\text{A,B}} = \|\mathbf{p}^t_\text{A} - \mathbf{p}^t_\text{B}\|_2$, is the Euclidean distance between two three-dimensional points corresponding to the joints A and B at time t, whereas h measures the constant length of a limb (indeed, a distance written in a more compact way), as in this case the length of the right upper half arm, $h_{\text{arm}_\text{R}} = d_{\text{shoulder}_\text{R},\text{elbow}_\text{R}}$, and the length of the left upper half arm, $h_{\text{arm}_\text{L}} = d_{\text{shoulder}_\text{L},\text{elbow}_\text{L}}$.

(3) *Pointing*. The average time the user is pointing towards the presentation screen. In order to know whether the user is pointing or not, we firstly discard those situations in which 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 audience, 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 in the real-valued interval ψ indicate that the user is pointing the presentation screen with high precision:

$$f_3 = \frac{1}{T}\sum_{t=1}^{T} \mathbb{1}\big\{P^t_{\text{hand}_L} \vee P^t_{\text{hand}_R}\big\}, \qquad (5)$$

$$
\begin{aligned}
&P^t_{\text{hand}_s} \\
&\quad = \mathbb{1}\big\{d^t_{\text{hand}_s,\text{body}} > d^t_{\text{elbow}_s,\text{body}}\big\} \\
&\qquad \times \mathbb{1}\Bigg\{\psi_a \\
&\qquad \leqslant \left(\frac{\|\mathbf{p}^t_{\text{hand}_s} - \mathbf{p}^t_{\text{hip}}\|}{\|\mathbf{p}^t_{\text{hand}_s} - \mathbf{p}^t_{\text{elbow}_s}\| \cdot |z^t_{\text{hand}_s} - z^t_{\text{hip}}|}\right)^{-1} \\
&\qquad \leqslant \psi_b\Bigg\}, \qquad (6)
\end{aligned}
$$

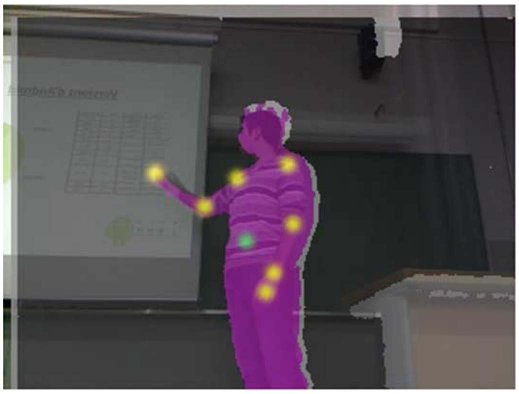

Fig. 3. Student pointing towards the presentation screen. Yellow points indicates the joints used to compute whether the student is pointing towards the presentation screen. The green point represents the hip center reference point. (The colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

where $P^t_{\text{hand}_s}$ is a binary variable indicating if the user's hand $s \in \{L, R\}$ is pointing to the presentation screen at time t. Figure 3 illustrates this situation. The different marks represent the inferred joints' positions, and concretely the green one indicates the hip center, used as reference point in the computation of the 'Pointing' feature.

(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 an implementation of the VAD algorithm [27] to detect voice activity in audio frames. We use the set of three low-level audio features A previously described and take 30 frames for threshold initialization. For each incoming audio frame, the three features are computed. The audio frame is marked as a voice frame ($v^t = 1$) if more than one audio feature $a \in A$ have a value over its precomputed threshold ρ_a. We consider that the subject is speaking ($s^t = 1$) only if there are M or more successive frames marked as voice activity:

$$f_4 = \frac{1}{T}\sum_{t=M}^{T} s^t_M, \qquad (7)$$

$$s^t_M = \mathbb{1}\left\{\left(\sum_{i=0}^{M-1} v^{t-i}\right) = M\right\}, \qquad (8)$$

$$v^t = \mathbb{1}\left\{\left(\sum_{a\in A} \mathbb{1}\{a^t > \rho_a\}\right) > 1\right\}. \qquad (9)$$

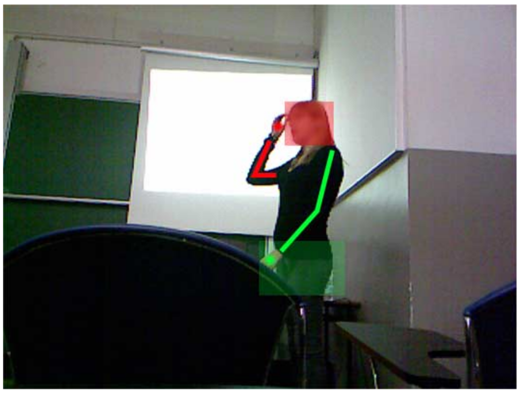

Fig. 4. Upper and bottom agitation. The red dot above the red square represents upper agitation, then the red lines are used to compute the magnitude, while the green dot indicates lower agitation. (The colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

(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).

$$\begin{aligned} f_5 = {} & \frac{1}{T-1} \\ & \times \sum_{t=2}^{T} \biggl(\sum_{j \in \mathcal{J}_U} \bigl\| \bigl(\mathbf{p}_j^t - \mathbf{p}_{\mathrm{hip}}^t\bigr) \\ & - \bigl(\mathbf{p}_j^{t-1} - \mathbf{p}_{\mathrm{hip}}^{t-1}\bigr) \bigr\| \biggr) \\ & \times \mathbb{1}\bigl\{ \bigl(y_{\mathrm{hand_L}}^t > y_{\mathrm{head}}^t\bigr) \\ & \vee \bigl(y_{\mathrm{hand_R}}^t > y_{\mathrm{head}}^t\bigr) \bigr\}, \end{aligned} \tag{10}$$

where $\mathcal{J}_U$ is the set of upper body limbs' joints: both hands, both wrists, and both elbows. Figure 4 shows a case of upper agitation.

(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):

$$\begin{aligned} f_6 = {} & \frac{1}{T-1} \\ & \times \sum_{t=2}^{T} \biggl(\sum_{j \in \mathcal{J}_U} \bigl\| \bigl(\mathbf{p}_j^t - \mathbf{p}_{\mathrm{hip}}^t\bigr) \\ & - \bigl(\mathbf{p}_j^{t-1} - \mathbf{p}_{\mathrm{hip}}^{t-1}\bigr) \bigr\| \biggr) \\ & \times \mathbb{1}\bigl\{ \bigl(y_{\mathrm{hip}}^t < y_{\mathrm{hand_L}}^t < y_{\mathrm{head}}^t\bigr) \\ & \vee \bigl(y_{\mathrm{hip}}^t < y_{\mathrm{hand_R}}^t < y_{\mathrm{head}}^t\bigr) \bigr\}. \end{aligned} \tag{11}$$

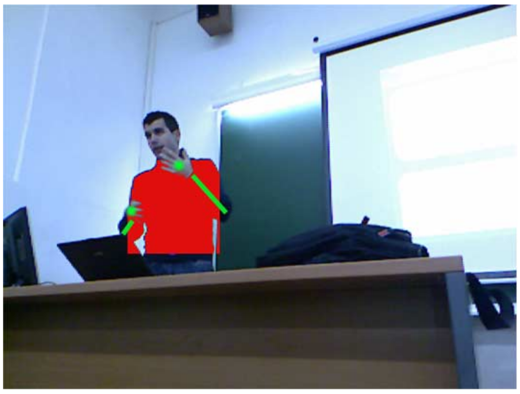

Fig. 5. Middle agitation. Green dots inside the red area indicates middle agitation. (The colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

In Fig. 5, middle agitation is illustrated.

(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)

$$\begin{aligned} f_7 = {} & \frac{1}{T-1} \\ & \times \sum_{t=2}^{T} \biggl(\sum_{j \in \mathcal{J}_U} \bigl\| \bigl(\mathbf{p}_j^t - \mathbf{p}_{\mathrm{hip}}^t\bigr) \\ & - \bigl(\mathbf{p}_j^{t-1} - \mathbf{p}_{\mathrm{hip}}^{t-1}\bigr) \bigr\| \biggr) \\ & \times \mathbb{1}\bigl\{ \bigl(y_{\mathrm{hand_L}}^t < y_{\mathrm{hip}}^t\bigr) \\ & \vee \bigl(y_{\mathrm{hand_R}}^t < y_{\mathrm{hip}}^t\bigr) \bigr\}. \end{aligned} \tag{12}$$

Figure 4 shows the bottom agitation with one hand, whereas the other was being agitated in the upper area.

(8) *Agitation while speaking*. The average of the displacement in real coordinates of arms, wrist and hands while the user is speaking combined with the response of speaking indicator

$$f_8 = \frac{1}{T-1} \times \sum_{t=2}^{T} \biggl(\sum_{j \in \mathcal{J}_U} \bigl\| (\mathbf{p}_j^t - \mathbf{p}_{\text{hip}}^t) - (\mathbf{p}_j^{t-1} - \mathbf{p}_{\text{hip}}^{t-1}) \bigr\| \biggr) \times \mathbb{1}\{v_M^t = 1\}. \tag{13}$$

(9) *Agitation while not speaking*. The average of the displacement in real coordinates of arms, wrist and hands while the user is not speaking combined with the response of speaking indicator

$$f_9 = \frac{1}{T-1} \times \sum_{t=2}^{T} \biggl(\sum_{j \in \mathcal{J}_U} \bigl\| (\mathbf{p}_j^t - \mathbf{p}_{\text{hip}}^t) - (\mathbf{p}_j^{t-1} - \mathbf{p}_{\text{hip}}^{t-1}) \bigr\| \biggr) \times \mathbb{1}\{v_M^t = 0\}. \tag{14}$$

Some examples of the detected low and high-level features are shown in Fig. 6. 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 groundtruth for that particular data sample. In the next section, before the presentation of the experimental results, we describe the novel dataset we recorded, and describe the different statistical classifiers we considered to validate our framework (mainly kernel machines, an adaptive boosting, two different neural networks, a randomized decision forest, a naive Bayes classifier, and lazy learning), which are used to validate the framework from the point of view of different learning paradigms: binary classification of two groups of quality, multi-class classification into several groups, analysis of feature selection of most relevant indicators, ranking prediction from classifiers capable of predicting this kind of structured output, and finally a real-valued quality prediction by means of regression.

3. Experimental validation of the communication competence analysis system

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 54 recorded videos, including 32 Bachelor thesis' defenses, 11 presentations from an 8th semester Bachelor course, and 11 presentations from a Master course at Universitat de Barcelona. All the videos were recorded with a Kinect™ device at a constant frame rate of 14 FPS, acquired on three different classrooms, and placing the

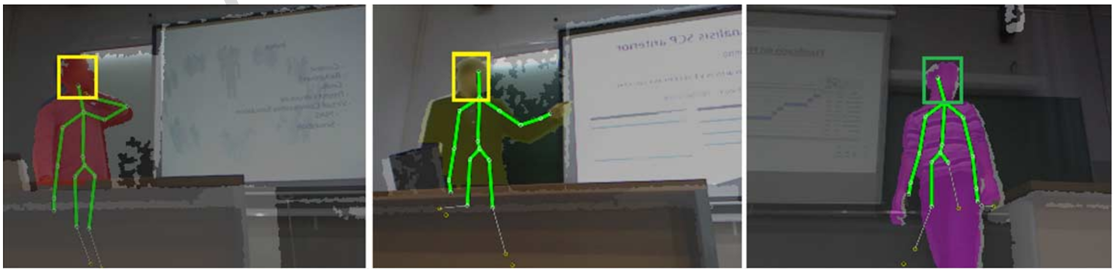

Fig. 6. Examples of low-level feature extraction. Depth maps and RGB images are superimposed with transparency. Color in the human body indicates user detection at pixel level using the approach of [37]. In this case, different colors indicate different user identifiers. Detected skeletons are drawn in green. Detected faces are also marked. Yellow color of faces indicates that the speaker is not looking at the audience/panel and green marked face indicates that the speaker is frontal-facing. (The colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

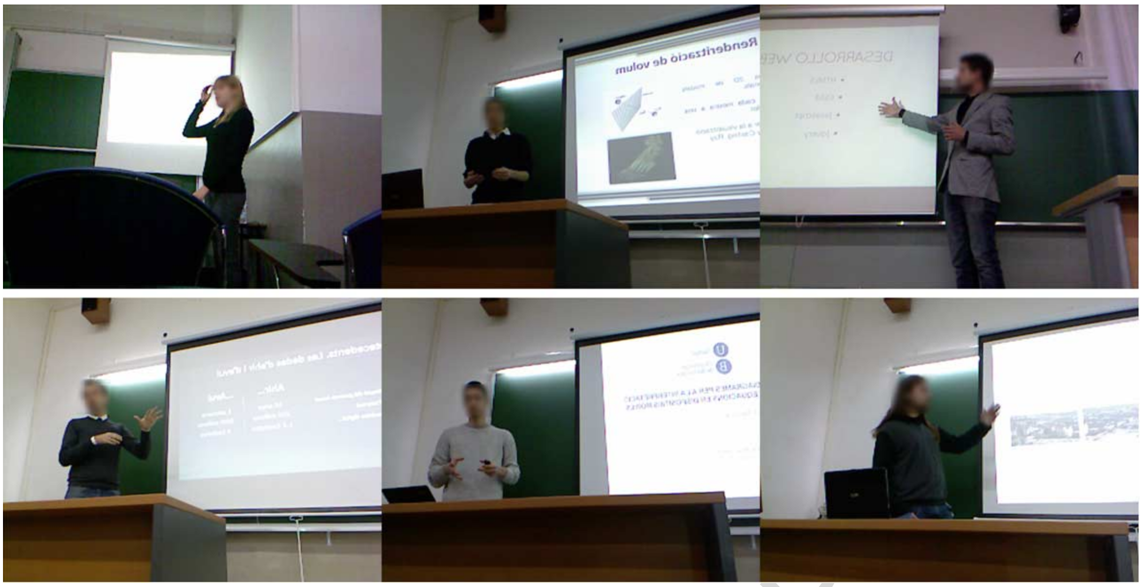

Fig. 7. Some examples of the presentations of our dataset. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

capturing device in front of the audience (thus, looking at them implies looking at the acquisition device). In total, with a mean duration of about 15 minutes per presentation, 768,600 frames were acquired for further processing. In addition, some examples of the recorded scenarios are also shown in Fig. 7.

Each presentation was rated by three different instructors, each of them providing a real-valued mark in the range $[0, 10]$, though in practice the minimum mark assigned is 6. It is interesting to notice that in higher levels of education, there is an increment in the students' marks, being significantly better the Master course's presentations and in a close second place the Bachelor thesis' defenses. In order to define the groundtruth, we averaged the marks assigned by the different instructors to each presentation; though, we first extracted a measure of agreement among their samples of marks based on the Pearson correlation coefficient r, which is computed as follows:

$$r = \frac{\sum_{i=1}^{R}(x_i - \bar{x})(y_i - \bar{y})}{\sqrt{\sum_{i=1}^{R}(x_i - \bar{x})^2}\sqrt{\sum_{i=1}^{R}(y_i - \bar{y})^2}}, \qquad (15)$$

where R is size of the sample of ratings.

Since the pairwise agreements of the raters are greater than the critical value[3] (0.2681 for 54-2 degrees of freedom at a significance level of 0.05), we considered the average of the marks could be safely computed. In Table 1, the pairwise correlation coefficients among the three raters are shown.

Table 1
Pearson correlation coefficient among raters

	Rater 1	Rater 2	Rater 3
Rater 1	1	0.8353	0.4204
Rater 2	0.8353	1	0.4518
Rater 3	0.4204	0.4518	1

3.2. *Settings*

Regarding the multi-modal feature extraction, some parameters were experimentally set. These values are summarized in Table 2.

In order to train the multi-modal features and evaluate the quality of the presentations, we use different statistical learning algorithms. Specifically, and for the classification scenario, we selected the following classifiers: a Support Vector Machine with Radial Basis Function kernel (SVM-RBF) [7], a Gentle AdaBoost with decision stumps [17], a Radial-basis Function Neural Network (RBFNN) [7], a Multi-Layer Perceptron (MLP) [34], a Random Forest (RF) [6], a k-Nearest Neighbor [15], and a Naive Bayes classifier. In the ranking experiment, we used Ranking Support Vector Machine (RankSVM) [23]. Finally, and for regression purposes, epsilon-Support Vector Regressor (e-SVR) [40] has been used.

[3] http://capone.mtsu.edu/dkfuller/tables/correlationtable.pdf.

Table 2
Feature extraction parameters

Parameter	Value	Description
α	$\frac{\pi}{6}$	Angular distance to frontal-facing direction
ψ	$[0.0039, 1]$	Pointing's threshold range
M	5	No. of successive voice activity frames
ρ	$(0.3, 0.5, 0.2)$	Audio features' thresholds tuple

Table 3
Application of learning algorithms in the different scenarios

Scenario	Learning algorithms
Classification (2–6 classes)	*SVM-RBF*, *AdaBoost*, *RBFNN*, *MLP*, *RF*, *K-NN*, *Naive Bayes*
Feature selection (2 classes)	*SVM-RBF*, *AdaBoost*
Ranking	*RankSVM*
Regression	*e-SVR*

The kernel machines used in this paper are implementations from the *LibSVM* library [9]. The AdaBoost is a self-made implementation and generalized to multi-class using a one-versus-one ECOC design [13]). The rest of the classifiers are from the *caret* package,[4] implemented in R language.

AdaBoost is used in three ways, first to obtain a classifier which is able to separate between two differentiated groups: "good" versus "bad" presentations, to perform multi-class classification among different presentation ranges of marks, and also as a feature selection method analyzing the weights assigned by the classifier to each of the high-level indicators. We also analyzed the weights assigned to the features in the case of SVM-RBF to analyze the most relevant indicators. Moreover, SVM-RBF classifiers are tested in four additional scenarios: binary classification, multi-class classification, ranking, and regression. Besides, the rest of the classifiers are used in binary and multi-class categorization as well. Finally, two additional variations of SVM, RankSVM and e-SVR, are used to predict rankings of presentations and to make real-valued mark predictions respectively. The application of the learning algorithms to the different scenarios is summarized in Table 3.

3.3. Experimental methodology and validation measures

The parameters α, ψ, M and ρ have been selected in order to maximize the performance of the system. This selection has been done by means of different grid searches: one for α, another one for the two extremes of the ϕ interval, and the last one to optimize together M and the values of the 3-dimensional tuple ρ. For this purpose, a small number of examples of "Frontal-facing", "Pointing" and "Speaking" actions were manually annotated in the sequences, and lately a detection accuracy measure was computed to assess the goodness of the different combinations of parameters for the automatic detection. Concretely, 5 examples per action and per subject were labeled. Then, having those parameters fixed, the system can be validated as it is explained next.

In order to measure the generalization capability of the proposed system, we perform a leave-one-out cross validation (LOOCV) in classification and regression problems: that is, a single observation from the original sample is separated so as to be the test data, and the remaining observations, the training data, are used to train a parameterized model. This process is repeated as many times as observations we have, and the average number of hits are divided by the total number of elements in the sample to get an accuracy measure. And, for the regression problem, the performance is measured with the root-mean-square error (RMSE).

In the case of the ranking classification, because the prediction is a structured output from the instances in the test sample, it has no sense to have just one sample in it. Thus, a k-fold cross validation is used to measure the performance in this case, instead of LOOCV. In the test sample, the error in the prediction is calculated as the ratio between how many positions did the classifier fail predicting the correct position and the maximum number of displacement errors. This metric is detailed in the ranking experiment section.

In all cases, within the training partitions of the different LOOCVs or k-fold CVs, an internal 5-fold CV is performed in order to parameterize the learning algorithm with the best selection of the learning parameters and to obtain the best model. Furthermore, in the case of the ANNs, this internal step has to be repeated many times because of the stochastic process implied by the random neuron weights' initialization and to keep the best model, the one that performed better in average in a particular validation set.

3.4. Experimental results and discussion

In order to validate the proposed system, we performed five different analyses: (a) binary classification into "good" or "bad" quality presentations, (b) multi-

[4] http://caret.r-forge.r-project.org/index.html.

class classification into three, four five and six categories of quality, (c) analysis of feature relevance and feature selection and classification with different feature subsets, (d) ranking of presentations based on quality, and finally, (e) regression.

3.4.1. Binary classification

In order to train our model in the binary classification problem, we consider a "good" presentation if its mark (in a scale from 6 to 10) is greater or equal than 8.0, and "bad" otherwise. The first bars of Fig. 8 shows the results for binary classification using the different classifiers. In particular, one can see the successfulness of SVM-RBF, RBFNN, and AdaBoost for automatically splitting the presentations in those two levels of quality. In the case of SVM-RBF, the achieved accuracy is approximately of 80% correctly categorized presentations.

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, four, five, and six qualification groups. The discretization in categories is done equally partitioning in width the range [6, 10] with the number of desired partitions.

The results applying multi-class one-versus-one SVM, ECOC multi-class AdaBoost, and the rest of the multi-class classifiers are shown in Fig. 8. In general, SVM outperforms the rest of classifiers. RBFNN and AdaBoost perform similarly to the SVM in the 2-class, but suffered a more severe drop of performance in the 3-class categorization. It is interesting to note in the case of the 6-class discretization that RBFNN outperforms SVM-RBF and also the maintenance of performance of K-NN. The results of the Naive Bayes for more than 4 classes could not be computed because some categories were not represented by the minimum number of examples necessary to train the classifier.

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 approximately 60% with the opinions of the instructors in the case of three and four qualification categories, and 50% and 40% in the five and six categories respectively.

3.4.3. Feature correlation, importance and selection

The first experiment in this section is intended to explain how the features are related to each other and to the output. In order to do this, we simply computed a matrix of pair-wise linear correlation coefficients. The results got are illustrated in Fig. 9. Red tones and blue tones indicate negative and positive correlations respectively.

From those calculations, we found some statistically significant correlations (testing the hypothesis of no correlation against the alternative that there is a non-zero correlation, at a significance level of 0.05), among them: the strong negative correlation between "Frontal-facing" and "Pointing" (−0.46), which is quite straightforward to analyze. When pointing, the subject must assert visually the pointing direction towards the region of interest in the presentation screen, thus not being able to face the audience at the same time. Another negative correlation to point out is the one between "Frontal-facing" and "Crossed arms" (−0.34). On the other hand, we found obvious positive correlations among agitation-like features. In addition, an interesting fact to point out is the null correlation

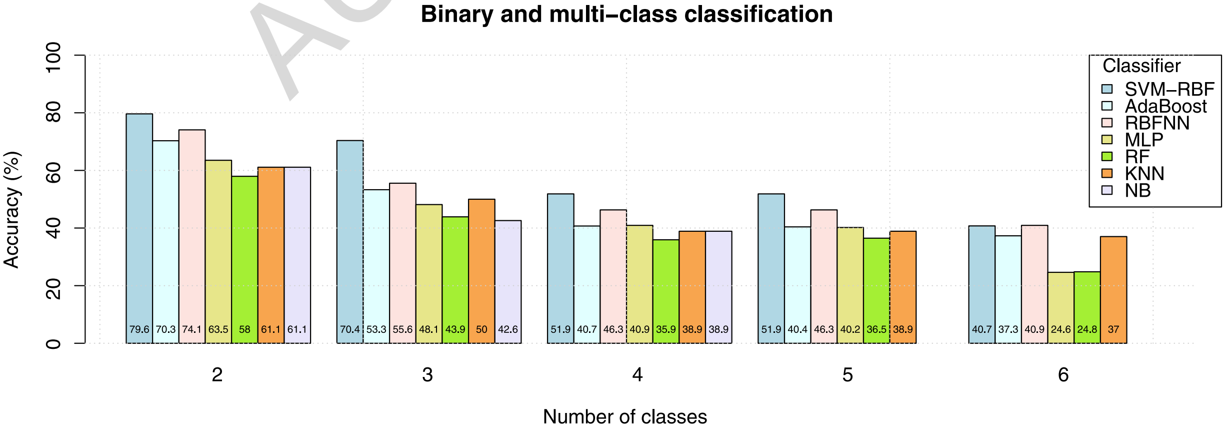

Fig. 8. Binary and multi-class classifiers performance comparison. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

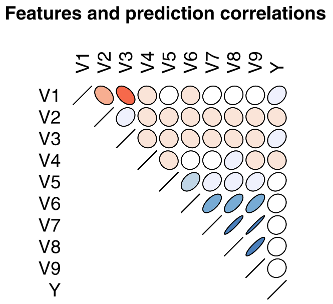

Fig. 9. Features (V1–V9) and prediction (Y) pair-wise linear correlations. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

between speaking and agitation-like features. Possibly, a feature expressing the voice intensity while speaking, instead of indicating the fluency in the speech ("Speaking"), would be positively correlated to the agitation-related indicators. Finally, when looking at the correlation measurements among the output and all the input variables, we find the strongest correlations are with "Frontal-facing", "Pointing" and "Upper agitation" (0.29, 0.23 and 0.17, respectively). However, the only input variable we found to be significantly correlated to the output is the "Frontal-facing" feature.

We also perform a feature importance analysis based on the weights assigned by AdaBoost and SVM classifiers to the 9 high-level indicators when discriminating among categories of presentations; later, these results are used in the feature selection. For all the iterations of the leave-one-out evaluation, the alpha weights assigned by AdaBoost and the weights assigned by SVM are saved, averaged from the different iterations, and normalized by their sums. These computed normalized values, presented in Table 4, express relative importance among features. SVM-RBF and AdaBoost were selected for this experiment because their good results and the easy-to-handle extraction of the weights, at the contrary to the RBFNN, that was omitted for this experiment.

For each classifier, the four features selected with the highest score are in bold. One can see that both classifiers correlate in the relevance of 'Frontal-facing' and 'Pointing', selected with high scores by both classifiers. Additionally, AdaBoost gives high scores to the 'Middle agitation' and 'Crossed arms' meanwhile SVM declares as relevant the 'Speaking' and 'Agitation while speaking' indicators. Therefore, agitation indicators become relevant for both classifiers. Whilst in the case of the AdaBoost it is quite obvious from the results, in the SVM-RBF this is not so evident. Nonetheless, considering the weight assigned to the agitation-involved indicators in relation to the third and fourth more important indicators, we can assert their importance.

Table 4

Feature relevance

Feature no.	Description	Classifier	
		AdaBoost	SVM-RBF
1	Frontal-facing	**28.92**	**50.88**
2	Crossed arms	**14.08**	0.00
3	Pointing	**22.14**	**7.01**
4	Speaking	8.46	**16.28**
5	Upper agitation	2.17	5.75
6	Middle agitation	**16.91**	3.42
7	Bottom agitation	6.54	6.41
8	Agit. while speaking	0.62	3.28
9	Agit. while not speaking	0.13	**6.92**

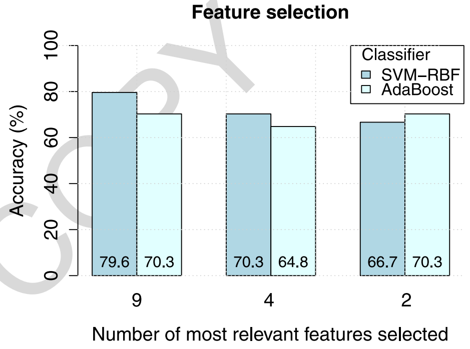

Fig. 10. Binary classification with different feature subsets. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

Finally, in order to analyze the generalization capability of the most relevant features, we reproduced the binary classification of presentations with subsets 2, 4 and 9 high-level features. Figure 10 shows the results. The first pair of bars corresponds to the previous results using the complete set of nine high-level behavioral indicators. Second and last sets of bars show 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 decreases due to the use of a reduced feature set, we are able to correlate almost 70% of the times with the instructors' marks only considering the four, and even the two, most discriminate features.

Surprisingly, AdaBoost is outperforming SVM in the most extreme case of feature selection. It turns out that, in its case, all but the two most discriminative features were acting like noise in the classification task.

3.4.4. Ranking

The goal of RankSVM [23] is to predict multivariate or structured outputs. In this case, we use the groundtruth mark value to order all the presentations by score, and generate pair-wise preference constraints. For this experiment, we defined different number of splits of our data, namely 2, 3 and 5-fold cross-validation, so that the instances in the different test samples are ordered by quality. In this case, a ranking error E_ε is computed, 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:

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

where m is the number of missed positions, S is total of test samples at each iteration of a k-fold experiment, and d is the number of different marks within the test samples. Then, the classification performance C is defined as $C = 100 - E_\varepsilon$. The results of these experiments are shown in Table 5. One can see that for different values of $k \in \{2, 3, 5\}$, corresponding to rank at each iteration of the cross validation 27, 18 and 10 test samples, respectively, high performance rates were achieved, approximately in the range of 75–90%.

Table 5

Ranking of presentation results

k	E_ε (%)	C (%)
2	24	76
3	12	88
5	19	81

3.4.5. Regression for mark prediction

Finally, we performed a regression analysis using epsilon-SVR (from LibSVM) to estimate the relationships among variables, between a dependent variable (the mark) and the independent variables (the feature vector). The estimation target is a function of the independent variables used to predict a mark. The results are shown in Fig. 11. For each presentation, the groundtruth mark and the automatically computed one by means of regression are illustrated. The predictions are represented with green bars, whereas the groundtruth marks are drawn as red dots. Note the high correlation among the estimations and the real marks. In this case, the root-mean-square error of the leave-one-out regression evaluation is computed:

$$\mathrm{RMSE} = \sqrt{\frac{\sum_{i=1}^{N} y_i - \hat{y}_i}{N}},$$

where y_i and $\hat{y}_i$ are the predicted and groundtruth values respectively, and N the total number of sequences (54). The result was 1.26 points.

4. Conclusion

An automatic system for the quantitative evaluation of the quality of oral presentations was presented, performing multi-modal human behavior analysis from RGB, depth, and audio data. A reliable set of high-level behavior indicators was defined. Moreover, a novel dataset of multi-modal oral presentations' sequences was recorded; for them, a groundtruth of marks was defined based on the scores assigned by three actual instructors from the Universitat of Barcelona. Then,

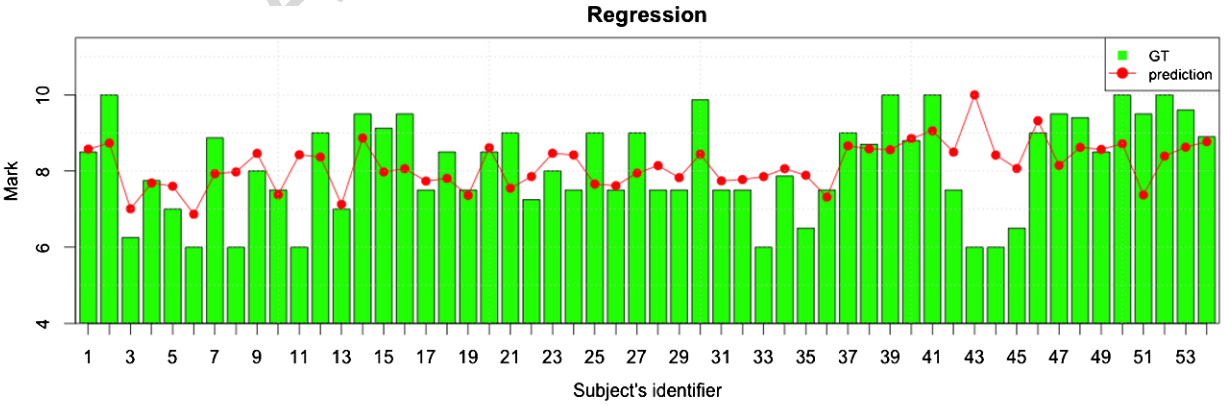

Fig. 11. Regression analysis. (Colors are visible in the online version of the article; http://dx.doi.org/10.3233/AIC-140617.)

trained binary, multi-class, and ranking classifiers, together with a regressor, evaluated the performance of the proposed system. In addition, feature relevance and feature selection analyses determined which were the most discriminative features, that correlates to the observers opinion, achieving classification rates of approximately 80% of accuracy categorizing two levels of presentation quality, and upon 60% three and four categorization, and 50% and 40% in the case of five and six qualities respectively.

One of the main issues we found in our setup is the wide range of movement of the speaker; this causes a great number of potential occlusions of body parts (with tables or other furniture) or even the possibility of not having the subject in the view frustum, which increase the bad estimations of the indicators involving the body pose estimation, yielding in a decrease of performance. In any case, since we assume the system would be used as a CAT application in a very controlled environment, the former problems will not occur. On the other hand, the feature vectors representing the presentations are global summaries that do not take into account changes in the quality of the presentation throughout time. In many cases, the speaker starts more nervous and performs worse than normal, but as the time goes by, the speaker calms down and recovers his usual non-verbal communication quality level. This fact affects the systems' prediction, since the indicators measure the average performance of the speaker. However, our hypothesis is that the quality level observed in later track of the presentation tends to have more impact in the observers' evaluation than the one in the initial track.

The results of this work show the feasibility of the system to be applied as an automatic tool useful for both evaluation purposes, and for providing user feedback in training scenarios as well.

5. Future work

Given the reliability of our system, as future work, we first plan to increase the amount of behavioral patterns including temporal information (the quality of the presentation may vary during time, for instance, being worse at the beginning and better at the end). Then, we plan to recognize facial expressions. Moreover, we are also planning to extend the number of samples so that the learners can have more data to learn and the correct quality level, rank position, or more a precise mark in regression for each presentation could be computed. Finally, we plan to apply the methodology in real scenarios to define a useful protocol for user feedback and include the framework as an e-Learning tool in the training routine of non-verbal communication competence.

References

[1] E. Babad, Teachers' nonverbal behavior and its effects on students, in: *The Scholarship of Teaching and Learning in Higher Education: An Evidence-Based Perspective*, Springer, 2007, pp. 201–261.

[2] H. Blok, R. Oostdam, M.E. Otter and M. Overmaat, Computer-assisted instruction in support of beginning reading instruction: a review, *Review of Educational Research* **72**(1) (2002), 101–130.

[3] C.J. Bonk and C.R. Graham, *The Handbook of Blended Learning: Global Perspectives, Local Designs*, Wiley, 2012.

[4] C.L. Breazeal, Sociable machines: Expressive social exchange between humans and robots, PhD thesis, Massachusetts Institute of Technology, 2000.

[5] C.L. Breazeal, *Designing Sociable Robots*, MIT Press, 2004.

[6] L. Breiman, Random forests, *Machine Learning* **45**(1) (2001), 5–32.

[7] D.S. Broomhead and D. Lowe, Radial basis functions, multi-variable functional interpolation and adaptive networks, Technical report, DTIC Document, 1988.

[8] J. Cassell, *Embodied Conversational Agents*, MIT Press, 2000.

[9] C.-C. Chang and C.-J. Lin, LIBSVM: a library for support vector machines, *ACM Transactions on Intelligent Systems and Technology (TIST)* **2**(3) (2011), 27.

[10] R.C. Clark, *Developing Technical Training: A Structured Approach for Developing Classroom and Computer-Based Instructional Materials*, Wiley, 2011.

[11] J.E. Coulson, Programmed learning and computer-based instruction, in: *Proceedings of the Conference on Application of Digital Computers to Automated Instruction*, Wiley, 1962.

[12] G.J. Edwards, C.J. Taylor and T.F. Cootes, Interpreting face images using active appearance models, in: *Proceedings of the Third IEEE International Conference on Automatic Face and Gesture Recognition*, IEEE, 1998, pp. 300–305.

[13] S. Escalera, O. Pujol and P. Radeva, Error-correcting output codes library, *Journal of Machine Learning Research* **11** (2010), 661–664.

[14] S. Escalera, O. Pujol, P. Radeva, J. Vitria and M.T. Anguera, Automatic detection of dominance and expected interest, *EURASIP Journal on Advances in Signal Processing* **2010** (2010), 39.

[15] E. Fix and J.L. Hodges Jr., Discriminatory analysis-nonparametric discrimination: consistency properties, Technical report, DTIC Document, 1951.

[16] Y. Freund and R.E. Schapire, A decision-theoretic generalization of on-line learning and an application to boosting, in: *Computational Learning Theory*, Springer, 1995, pp. 23–37.

[17] J. Friedman, T. Hastie and R. Tibshirani, Additive logistic regression: a statistical view of boosting, 1998, 7(7.1), available at: citeseer.ist.psu.edu/friedman98additive.html.

[18] R.M. Gagné, *Instructional Technology: Foundations*, Routledge, 2013.

[19] P.J. Gorman, A.H. Meier and T.M. Krummel, Computer-assisted training and learning in surgery, *Computer Aided Surgery* **5**(2) (2000), 120–130.

[20] V. Günther, P. Schäfer, B. Holzner and G. Kemmler, Long-term improvements in cognitive performance through computer-assisted cognitive training: a pilot study in a residential home for older people, *Aging & Mental Health* **7**(3) (2003), 200–206.

[21] D.M. Hardison, Generalization of computer-assisted prosody training: quantitative and qualitative findings, *Language Learning & Technology* **8**(1) (2004), 34–52.

[22] H. Ishikawa, H. Hashimoto, M. Kinoshita, S. Fujimori, T. Shimizu and E. Yano, Evaluating medical students' non-verbal communication during the objective structured clinical examination, *Medical Education* **40**(12) (2006), 1180–1187.

[23] T. Joachims, Training linear SVMS in linear time, in: *Proceedings of the 12th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, ACM, 2006, pp. 217–226.

[24] C.-L.C. Kulik and J.A. Kulik, Effectiveness of computer-based instruction: an updated analysis, *Computers in Human Behavior* **7**(1) (1991), 75–94.

[25] A. Marcos-Ramiro, D. Pizarro-Perez, M. Marron-Romera, L. Nguyen and D. Gatica-Perez, Body communicative cue extraction for conversational analysis, in: *10th IEEE International Conference and Workshops on Automatic Face and Gesture Recognition (FG)*, IEEE, 2013, pp. 1–8.

[26] L. McCowan, D. Gatica-Perez, S. Bengio, G. Lathoud, M. Barnard and D. Zhang, Automatic analysis of multimodal group actions in meetings, *IEEE Transactions on Pattern Analysis and Machine Intelligence* **27**(3) (2005), 305–317.

[27] M. Moattar, M. Homayounpour and N.K. Kalantari, A new approach for robust realtime voice activity detection using spectral pattern, in: *IEEE International Conference on Acoustics Speech and Signal Processing (ICASSP)*, IEEE, 2010, pp. 4478–4481.

[28] G. Mohammadi and A. Vinciarelli, Automatic personality perception: prediction of trait attribution based on prosodic features, *IEEE Transactions on Affective Computing* **3**(3) (2012), 273–284.

[29] M.R. Nosik, W.L. Williams, N. Garrido and S. Lee, Comparison of computer based instruction to behavior skills training for teaching staff implementation of discrete-trial instruction with an adult with autism, *Research in Developmental Disabilities* **34**(1) (2013), 461–468.

[30] D.O. Olguín, B.N. Waber, T. Kim, A. Mohan, K. Ara and A. Pentland, Sensible organizations: technology and methodology for automatically measuring organizational behavior, *IEEE Transactions on Systems, Man, and Cybernetics, Part B: Cybernetics* **39**(1) (2009), 43–55.

[31] W. Pan, W. Dong, M. Cebrian, T. Kim, J.H. Fowler and A. Pentland, Modeling dynamical influence in human interaction: using data to make better inferences about influence within social systems, *Signal Processing Magazine, IEEE* **29**(2) (2012), 77–86.

[32] C.P. Papageorgiou, M. Oren and T. Poggio, A general framework for object detection, in: *Sixth International Conference on Computer Vision*, IEEE, 1998, pp. 555–562.

[33] J.M. Roschelle, R.D. Pea, C.M. Hoadley, D.N. Gordin and B.M. Means, Changing how and what children learn in school with computer-based technologies, *The Future of Children* (2000), 76–101.

[34] F. Rosenblatt, Principles of neurodynamics: perceptrons and the theory of brain mechanisms, Technical report, DTIC Document, 1961.

[35] D. Sanchez-Cortes, O. Aran, D.B. Jayagopi, M.S. Mast and D. Gatica-Perez, Emergent leaders through looking and speaking: from audio-visual data to multimodal recognition, *Journal on Multimodal User Interfaces* **7** (2012), 1–15.

[36] O. Serin and N.-N. Cyprus, The effects of the computer-based instruction on the achievement and problem solving skills of the science and technology students, *The Turkish Online Journal of Educational Technology* **10**(1) (2011), 183–202.

[37] J. Shotton, T. Sharp, A. Kipman, A. Fitzgibbon, M. Finocchio, A. Blake, M. Cook and R. Moore, Real-time human pose recognition in parts from single depth images, *Communications of the ACM* **56**(1) (2013), 116–124.

[38] H. Tanaka, S. Sakti, G. Neubig, T. Toda, N. Campbell and S. Nakamura, Non-verbal cognitive skills and autistic conditions: an analysis and training tool, in: *IEEE 3rd International Conference on Cognitive Infocommunications (CogInfoCom)*, 2012, pp. 41–46.

[39] J.A. Tulsky, R.M. Arnold, S.C. Alexander, M.K. Olsen, A.S. Jeffreys, K.L. Rodriguez, C.S. Skinner, D. Farrell, A.P. Abernethy and K.I. Pollak, Enhancing communication between oncologists and patients with a computer-based training program: randomized trial, *Annals of Internal Medicine* **155**(9) (2011), 593–601.

[40] V. Vapnik, *Statistical Learning Theory*, 1998.

[41] A. Vinciarelli, M. Pantic and H. Bourlard, Social signal processing: survey of an emerging domain, *Image Vision Comput.* **27**(12) (2009), 1743–1759.

[42] A. Vinciarelli, H. Salamin and M. Pantic, Social signal processing: understanding social interactions through nonverbal behavior analysis, in: *IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, IEEE, 2009, pp. 42–49.

[43] P. Viola and M.J. Jones, Robust real-time face detection, *Int. J. Comput. Vision* **57**(2) (2004), 137–154.

[44] J.S. Webster, P.T. McFarland, L.J. Rapport, B. Morrill, L.A. Roades and P.S. Abadee, Computer-assisted training for improving wheelchair mobility in unilateral neglect patients, *Archives of Physical Medicine and Rehabilitation* **82**(6) (2001), 769–775.

[45] T. Yoshida, A perspective on computer-based training, CBT, in: *Proceedings of the 2013 IEEE 37th Annual Computer Software and Applications Conference*, IEEE Computer Society, 2013, pp. 778–783.