Multimodal User Interfaces manuscript No.
(will be inserted by the editor)

Predicting Online Lecture Ratings Based on Gesturing and Vocal Behavior

Dong Seon Cheng · Hugues Salamin · Pietro Salvagnini · Marco Cristani · Alessandro Vinciarelli · Vittorio Murino

Received: date / Accepted: date

Abstract Nonverbal behavior plays an important role in any human-human interaction. Teaching - a inherently social activity - is not an exception. So far, the effect of nonverbal behavioral cues accompanying lecture delivery was investigated in the case of traditional *ex-cathedra* lectures, where students and teachers are co-located. However, it is becoming increasingly more frequent to watch lectures online and, in this new type of setting, it is still unclear what the effect of nonverbal communication is. This article tries to address the problem and proposes experiments performed over the lectures of a popular web repository ("*Videolectures*"). The results show that automatically extracted nonverbal behavioral cues (prosody, voice quality and gesturing activity) predict the ratings that "*Videolectures*" users assign to the presentations.

Keywords Video lectures · Generative modeling · Classification · Social signal processing

D.S. Cheng (✉)
Department of Computer Science & Engineering
Hankuk University of Foreign Studies, South Korea
Tel.: +82-10-2544-1974, Fax: +82-31-330-4120
E-mail: cheng_ds@hufs.ac.kr; cheng.dong.seon@gmail.com

H. Salamin[1] · A. Vinciarelli[1,2]
[1]School of Computing Science, University of Glasgow, UK
[2]Idiap Research Institute, Switzerland
E-mail: {hsalamin,vincia}@dcs.gla.ac.uk

P. Salvagnini · V. Murino
Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia
Via Morego 30, 16163 Genova, Italy
E-mail: {pietro.salvagnini,vittorio.murino}@iit.it

M. Cristani (✉)
Dipartimento di Informatica, Università di Verona
Strada Le Grazie 15, 37134 Verona, Italy
Tel.: +39 045 8027988, Fax: +39 045 8027068
E-mail: marco.cristani@univr.it

1 Introduction

During the last decade, advances in Information and Communication technologies had a major impact on teaching and learning practices. In the USA, as early as in the academic year 2000-2001, 90% of public 2-year and 89% of public 4-year institutions offered distance education courses, not to mention that almost three millions individuals were earning credits at college-level via online courses [41]. Social media as well, while not being specifically aimed at didactic activities, attracted the attention of teachers and students: in the USA, more than 60% of the faculties include social media in their teaching practices, whether this means to deliver content through social networking platforms (30% of the teachers) or to require access to social media in order to complete assessed exercises (40% of the teachers) [24].

Not surprisingly, the same period witnessed the birth of the most popular online repositories of lecture and presentation recordings like, e.g., "*Videolectures*"[1] (recognized in 2013 as a *World Summit Award Global Champion* by UNESCO) and the "*Khan Academy*"[2] (opened in 2006 and recognized with the *Microsoft Tech Award* in 2009). Furthermore, major institutions like the Stanford University[3] paved the way to *Massive Open Online Courses* (MOOCs), i.e. web delivered courses capable of attracting up to 160,000 individuals both watching lectures and solving assessed exercises [22].

More in general, the use of videos as a means of communication between people is becoming ubiquitous: at the moment this article is being written[4], Youtube users upload every day 12 years of video material and access the popular

[1] `http://www.videolectures.net`
[2] `http://www.khanacademy.org`
[3] `http://online.stanford.edu/courses`
[4] `http://www.youtube.com/yt/press/statistics.html`

on-line platform one trillion times per year, an average of 140 visits per person on Earth (the figure refers to 2011).

The evidence above suggests that watching online lectures is likely to become an important aspect of learning and teaching practices. While challenges and opportunities of such an evolution are widely debated (see, e.g., [11, 16, 17]), only minor attention was paid to the effect of nonverbal communication - well known to influence perception and outcomes of students in traditional teaching settings (see [40, 46] for extensive surveys) - during the consumption of online courses. The goal of this paper is to fill, at least partially, such a gap.

In particular, this work focuses on "*Videolectures*" (see above). One of the most interesting aspects of such a platform is that the online recordings are accompanied by ratings (one to five *stars*) that account for the overall appreciation of the users. Therefore, it is possible to investigate whether features automatically extracted from the lectures are predictive of the ratings or not. The experiments (performed over 90 lecture recordings) show that features accounting for gestures, voice quality and prosody (automatically extracted from the videos) allow one to predict with accuracy up to 69.4% whether a lecture is perceived to be of excellent quality (at least four stars) or not (less than four stars).

The rest of this paper is organised as follows: Section 2 surveys previous work on technological and psychological aspects of nonverbal communication, Section 3 describes the extraction of nonverbal behavioral cues, Section 4 reports on experiments and results and the final Section 5 draws some conclusions.

2 Previous Work

Nonverbal behavior is the subject of extensive investigation in Human and Computing Sciences. This section surveys some of the most important results of both areas with particular attention for works aimed at teaching related scenarios.

2.1 Human Sciences

Nonverbal communication (gestures, postures, facial expressions, vocalisations, etc.) is well known to play a major role in any form of human-human [15, 33] and human-machine interaction [26, 32]. Since teaching is a inherently social activity, education sciences investigated extensively the effect of nonverbal behavior during lectures or, more in general, oral presentations (see [40, 46] for extensive surveys). Two main aspects of the problem were explored: on one hand, the effect of teachers' nonverbal behavior on the impressions developed by students [2, 3] and, on the other hand, the impact of teachers' nonverbal behavior on students' outcomes [12, 36].

In [2], nonverbal behavioral cues (head nods, self-touches, yawns, frowns, downward gaze, etc.) observed in brief silent videos (10 seconds) were shown to be significantly correlated with ratings assigned by students. Similarly, students were shown to be consensual above chance about *warmth* and *flexibility* of teachers observed only for a few seconds in a video [3]. The way teachers communicate immediacy via nonverbal cues (e.g., the teacher is behind the desk or not, the teacher looks at the class or not, etc.) was shown to influence the cognitive performance of students in [36]. In the same vein [12], nonverbal immediacy was shown to have a statistically significant impact on student outcomes (grades, cognitive learning, etc.).

2.2 Computing

In recent years, nonverbal behavior attracted significant attention in the computing community as well. In particular, domains like Affective Computing [31] and Social Signal Processing [43] adopted nonverbal behavioral cues as a physical, machine detectable evidence of emotional and social phenomena, respectively. Research efforts targeted a wide spectrum of problems, including conflict detection [28], communication dynamics [7, 25], mimicry measurement [10], early detection of developmental and cognitive diseases [37], role recognition [38], prediction of negotiation outcomes [9], videosurveillance [4, 5, 6, 8], etc. Furthermore, several works were dedicated to the automatic prediction of traits likely to be relevant in a teaching context like, in particular, personality [21, 23, 30] and dominance [13, 27, 34, 35].

In [30], Support Vector Machines (SVM) classify audio and video feature vectors (including mean of pitch, energy and spectral entropy, fidgeting, etc.) into classes accounting for two personality traits (extraversion and locus of control). The works in [21, 23] predict the way prosodic features influence the perception of personality, namely the way traits are perceived by others. Both works use machine learning algorithms (e.g., SVMs) to map basic prosodic features (e.g. pitch and energy) into personality assessments made in terms of the Big Five, the most important and common personality model [45].

In [34, 35], speaking activity (speaking time, number of turns, interruptions, etc.) and SVMs have been used to predict whether people are *low*, *normal* or *high* in dominance. Similar speaking related features were fed to a Dynamic Bayesian Network in [27], together with visual attention features (who looks at whom) in order to predict the most dominant person in a meeting. A similar multimodal approach was proposed in [13].

Nonverbal cues influence the perception of social phenomena not only when displayed by humans, but also when

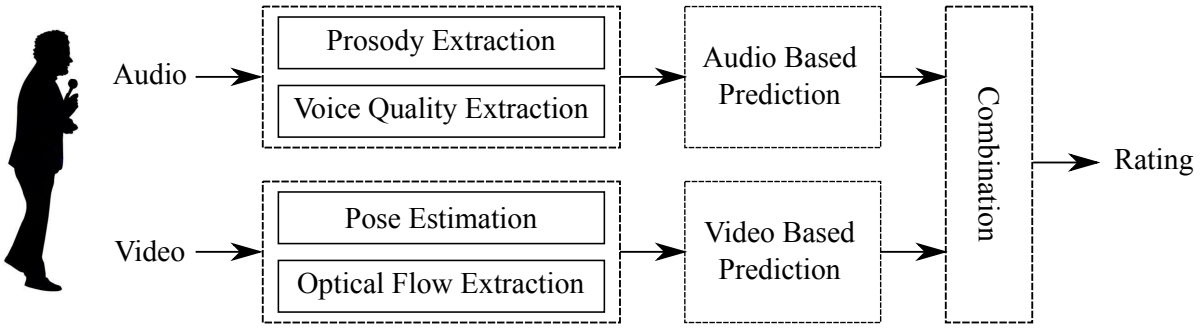

Fig. 1 The scheme shows tha rating prediction approach. Audio and video channels are processed separately to extract features accounting for nonverbal behavior. Rating predictions are made separately for the two modalities and then combined.

synthesized with machines [43]. This makes it possible to create artificial agents that replicate the results of social psychology [18] by interacting with humans like humans do [42]. In a teaching perspective, this was used, e.g., to develop agents helping students to learn procedural tasks [14] or instructing medical personnel on how to use web-based services [39].

3 Nonverbal Behavior Analysis

The goal of this work is to predict automatically whether *Videolectures* users rate a recording "*high*" (at least four *stars*) or "*low*" (less than four *stars*) based on the nonverbal behavior of the speaker. Figure 1 shows the overall architecture of the proposed approach. Video and audio channels are processed individually, the former to estimate pose and movement patterns, the latter to extract prosody and voice quality. The final prediction stage combines decisions made using individual modalities. The rest of this section shows how the approach analyses nonverbal behavioral cues and represents them in terms of feature vectors.

3.1 Pose Estimation

The approach estimates body pose using the *pictorial structures framework with flexible mixtures-of-parts* (see [47] for a full description). Such a model breaks a body image into K individually detectable *parts* that form the skeleton of a person (see Figure 2). In general, the *parts* correspond to recognizable limbs or joints (*e.g.*, head, arm, elbow, etc.) and are grouped into "*types*" T. These latter include different orientations (*e.g.*, vertical or horizontal) and semantic classes (*e.g.*, extended, bent, open, close, etc.), all captured through a mixture model.

Given an image I, $p_i = (x_i, y_i)$ denotes the pixel location of part i and t_i denotes its type. A *pose* corresponds to a configuration where the K *parts* co-occur while respecting mutual constraints (*e.g.*, the distance between upper and lower arm cannot exceed a certain value). The representation of a *pose* is a graph $G = (V,E)$ where vertices correspond to *parts* and links to constraints.

The model allows one to assign every pose a score that depends on three components:

$$S(I,p,t) = S(t) + S(I,p) + S(p). \tag{1}$$

The first component measures the compatibility of all part types:

$$S(t) = \sum_{i \in V} b_i^{t_i} + \sum_{i,j \in E} b_{ij}^{t_i,t_j}, \tag{2}$$

where b_i and b_{ij} are parameters learned from training data and $t = \{t_1, \ldots, t_K\}$ is a list of part types (see Section 4 for more details).

The second component is an appearance model that estimates the localization of each part in the image:

$$S(I,p) = \sum_{i \in V} w_i^{t_i} \phi(I, p_i), \tag{3}$$

where $p = \{p_1, \ldots, p_K\}$ is a list of part positions, and $w_i^{t_i}$ is a template for type t_i of part i. This latter is matched against the feature vector $\phi(I, p_i)$ extracted from location p_i (*e.g.*, Histogram of Gradients).

The third component is a deformation model that evaluates the relative placement of all connected parts:

$$S(p) = \sum_{i,j \in E} w_{ij}^{t_i t_j} \psi(p_i, p_j), \tag{4}$$

where the parameters $w_{ij}^{t_i t_j}$ are learned from the data and $\psi(p_i, p_j) = [dx_{ij}, dx_{ij}^2, dy_{ij}, dy_{ij}^2]^T$ is the relative location of parts i and j, with $dx_{ij} = x_i - x_j$ and $dy_{ij} = y_i - y_j$.

The value of parameters w and b is learned by maximizing the difference between the score of positive and negative examples (the former are images where the exact position of each body part is available, the latter are images where the exact position is not available). In this way, the position of the person depicted in I is inferred by maximizing $S(I,p,t)$ over p and t. In other words, the model estimates the score above for all possible positions and types of the K parts and finds the case for which the score is maximum.

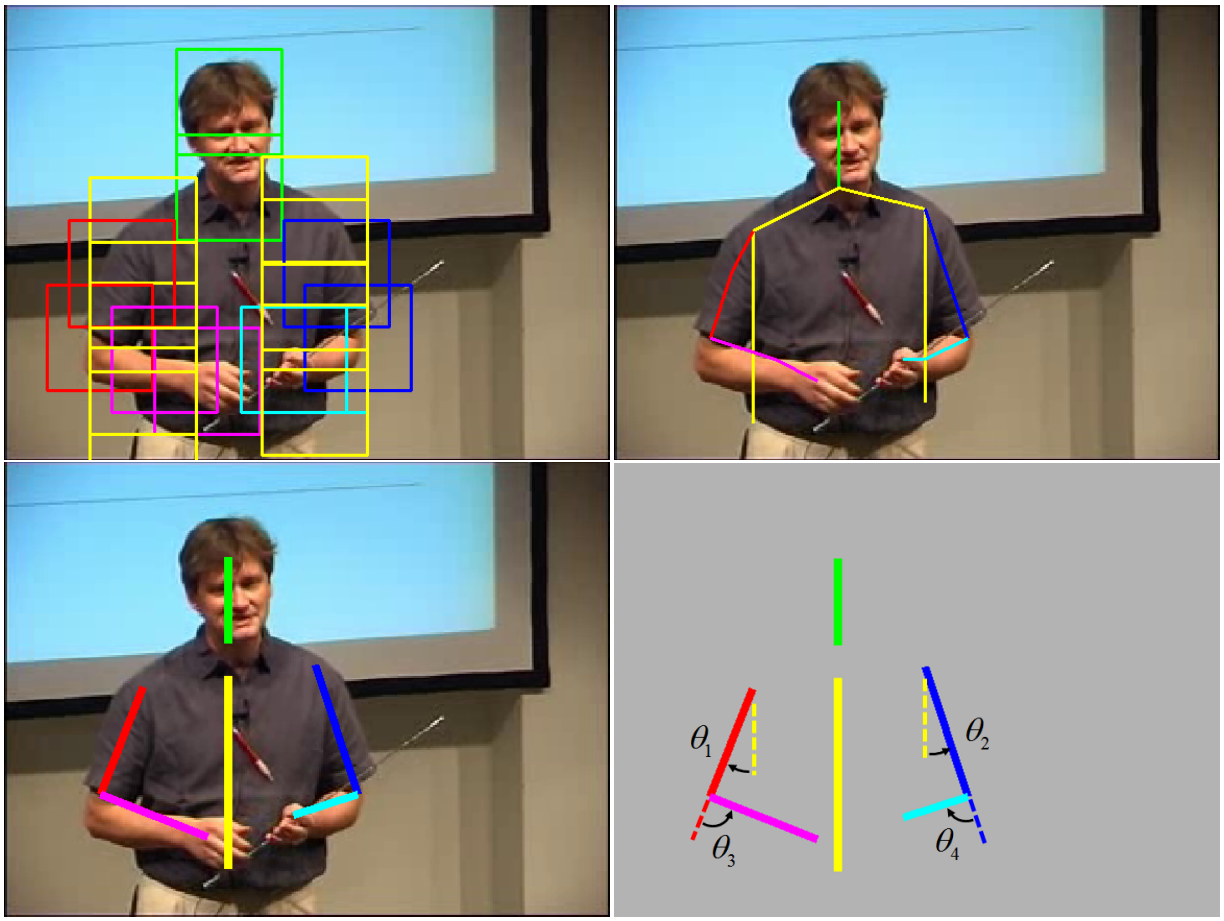

Fig. 2 The gesturing of the speaker is captured by tracking his body pose: (top-left) the pictorial structures framework of [47] detects the most plausible configuration of certain body parts; (top-right) these body parts connect to form a (partial) human skeleton; (bottom-left) we then calculate a "stick man" representation; and (bottom-right) we retrieve five angles of interest (θ_5, not shown, is the absolute deviation of the torso from the vertical position) at each frame for which the lecturer is detected.

The pose leading to the highest value of $S(I,p,t)$ allows the extraction of the "*stick-man*" of Figure 2 and, correspondingly, the measurement of five angles of interest: left shoulder angle θ_1, right shoulder angle θ_2, left elbow angle θ_3, right elbow angle θ_4, and torso inclination θ_5 (see Figure 2). The shoulder angles measure the relative change in orientation between the torso and the upper arms. The elbow angles measure the relative change in orientation between the upper arms and the lower arms. The torso angle measures the absolute deviation of the torso from the straight vertical orientation.

3.2 Movement from Optical Flow

The pose estimation algorithm adopts a simplified, skeleton-based representation of human body that can describe movements only partially. Therefore, the approach includes the extraction of low-level features that account for the optical flow and aim at capturing long-term movement patterns.

Raw optical flow vectors are extracted using the approach proposed in [20] and their dimensionality is then reduced via Principal Component Analysis (PCA). Due to memory constraints, the PCA is applied only to a subset of the optical flow vectors actually extracted from a given video. The number of vectors in such subset, denoted with N_{PCA}, is a hyper-parameter to be set experimentally (its value is supposed to be significantly smaller than the total number of raw vectors extracted from the video).

Optical flow vectors are projected onto the first N_{PC} Principal Components (N_{PC} is set to preserve a fraction E of the original data variance) and then grouped into N_C clusters using the *kNN* algorithm, where the value of N_C is set empirically. In this way, each optical flow vector can be assigned to a different cluster and a video can be represented with a histogram where each bin corresponds to the number of vectors assigned to a particular cluster.

The raw optical flow vectors were extracted with three different approaches: the first is to use the optical flow from the whole frame as a single vector. The second is to ex-

tract the optical flow from 100 patches uniformly distributed across the whole frame and consider each of them independently. The third is to extract the vectors only from the part of the image where the upper-body detector presented in [47] has identified the speaker.

3.3 Prosody and Voice Quality Extraction

Nonverbal aspects of speech influence, to a significant extent, the impression that listeners develop about speakers [44]. Therefore, the approach focuses on prosody (the way one talks) and voice quality (the spectral properties of one's voice). The extraction of such information includes two main steps: the first is the extraction of *low-level* features and the second is the estimation of statistics that account for the distribution of low-level features in a given recording.

The speech signal is first segmented into syllables using the approach in [29] because *syllable nuclei* are the segments where speech properties are more stable and can be extracted more reliably. The prosody related low-level features are *mean energy*, *mean pitch*, *mean* and *bandwidth* of the first three *formants*, *jitter* (pitch variation), *shimmer* (energy variation) and *syllable duration* measurements. The voice quality low-level features, extracted from the Long Term Average Spectrum of the syllable nuclei, are *harmonicity*, *spectral centroid*, *spectral tilt*, *spectral skewness* and *spectral kurtosis*. The feature set is completed by *glissando likelihood*, *Teager Energy Operator*, *Zero Crossing Rate* and the first 13 *Mel Frequency Cepstral Coefficients* (including their differences between consecutive frames).

The extraction of the features above is performed for each detected syllable. A speech sample is represented in terms of statistics that describe the distribution of the low-level features across individual syllables. In the case of this work, the statistics are mean, variance and 10^{th}, 25^{th}, 75^{th} and 90^{th} percentiles (the x^{th} percentile is the feature value below which lie x% of the observed feature values). The two main advantages of this approach are to ensure that the size of the feature vector is independent of the number of syllables and to reduce the dimensionality of the feature space. In total, a speech sample is represented with 216 features (6 statistics of 36 short-term features).

4 Prediction Experiments and Results

The next sections describe data and results obtained by using audio and video modalities both individually and in combination. All experiments were performed using a k-fold approach ($k = 9$) and the folds are the same for both modalities. The use of 9 folds makes it possible to have a test set with 10 samples (the dataset includes 90 videos) where both classes are represented 5 times. The experiments are speaker independent, the same subject never appears in both training and test set.

rating	total	analyzed	suitable	used
1	37	27	17	13
1.5	2	0	-	-
2	52	32	18	15
2.5	4	0	-	-
3	110	25	17	17
3.5	25	0	-	-
4	272	32	27	22
4.5	114	0	-	-
5	1067	30	26	23
total	1683	147	104	90

Table 1 Distribution of the ratings over the presentations of Videolectures. Around 150 presentations were analysed to be included in the corpus (column "analysed"), but only 104 of them show the speaker enough time to allow experiments (column "suitable"). The final number of presentations retained for the experiments (column "used") was set to ensure balance between the classes "low" (less than four) and "high" (more than three).

4.1 The Data

The experiments were performed over a corpus of 90 lecture recordings collected from *Videolectures*, one of the largest presentation repositories available on the web. The repository users have the possibility of assigning each lecture a score ranging between 1 ("*poor quality*") and 5 "*stars*" ("*excellent quality*"). As *rating* of a presentation, *Videolectures* posts the average of the scores assigned individually by all users.

Table 1 reports the distribution of the ratings across the presentations that were actually scored by the users (roughly 11% of the recordings available on *Videolectures*[5]). The table shows, from left to right, the total number of presentations assigned a given rating, the number of presentations analysed to be included in the corpus, the number of presentations identified as suitable for the experiments (meaning that the speaker is visible most of the time) and the number of presentations actually retained for the experiments. This latter was set to ensure a balanced distribution over the classes "*low*" (rating less than four) and "*high*" (rating greater or equal to four).

The experiments were performed over 3000 frames long segments (two minutes at 25 fps) extracted from each of the presentations[6].

[5] As of September 2011.

[6] The list of videos used in the experiment is available at `https://pavisdata.iit.it/data/salvagnini/RatingPrediction_VL/RP_VL90v_INFOCOM2013.pdf`

4.2 Experiments on Pose and Gestures

The pose and gesture estimator described in Section 3 was trained using four manually annotated videos of the corpus above as positive examples, and the videos of the *INRIA Person* corpus as negative examples (the same data as the work in [47]). The resulting estimator was used to extract the five angles depicted in Figure 2 for each frame of the 90 videos of the corpus (a total of 3000 angles per video). In case the speaker was not detected in a given frame, the angles were all set to null.

The angles were used as features in different combinations, only shoulders (S), only elbows (E), shoulders and elbows (SE), and all angles (SET). Table 2 reports the accuracy (percentage of correctly classified samples) for three different classifiers: Support Vector Machines (SVM) with radial basis kernel function, Logistic Regression (LR) and Hidden Markov Models (HMM) with Gaussian emission probabilities and different number of states.

In the case of SVM and LR, the sequence of the angle values extracted from a video was represented as a single feature vector and treated as a high-dimensional point. In the case of the Hidden Markov Models, it was possible to process the sequence as such and, therefore, to capture the gesture dynamics. The high visual appearance variability of the data does not allow one to train one HMM per class and then assign a test sequence $\bar{x}$ to class $\hat{c}$ according to the following expression:

$$\hat{c} = \arg\max_{c \in C} p(\bar{x}|\bar{\theta}_c) \tag{5}$$

where $\bar{\theta}_c$ is the parameter set of the HMM trained over the sequences of class c, $p(\bar{x}|\bar{\theta}_c)$ is the likelihood of such HMM and C is the set of the classes. The approach is rather to train one HMM per training sequence and then find the set $\mathscr{M}$ of the M models with the highest likelihood for the test sequence $\bar{x}$. Each of the models in $\mathscr{M}$ is trained over a sample of a certain class and, therefore, it is possible to find the class most frequently represented in the set. Such a class is assigned to the test sequence $\bar{x}$ as well. Such an approach corresponds to a *kNN* approach where the likelihoods act as distances (even if likelihoods do not induce a correct distance metric). In this way, the highly heterogeneous typologies of gestures can be locally learned, and taken into account in a non-parametric form by the *kNN* policy. The value of M was set through cross-validation for each HMM topology and set of angles (see above). The best performances were obtained with M ranging between 11 and 19. The fraction of models in $\mathscr{M}$ that are trained over the most represented class is used as a confidence score.

Table 2 reports the results obtained using a k-fold approach ($k = 9$) and three values for the number H of states in the HMMs ($H = \{2,3,4\}$). Overall, the results are better than in the case of SVMs and LR. In particular, elbows angles seem to be the most informative, while they give low accuracy when considered by the other approaches. The reason is likely that the states of the HMMs are the realization of a clustering process separating diverse configurations of the joint angles. Furthermore, the transition matrix accounts for the dynamics of such configurations. SVM and LR do not capture such aspects.

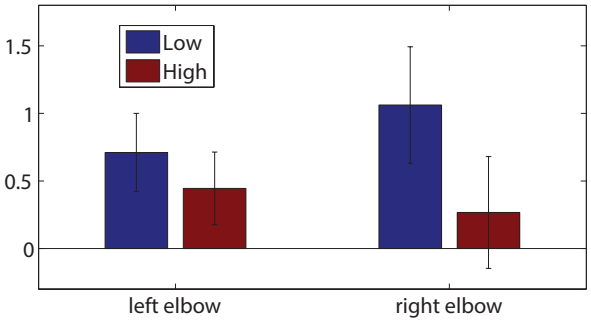

Fig. 3 Amplitude of gestures considering the elbow angles. As visible, liked lessons have the lecturer which presumably is steadier.

Besides achieving higher performances, HMMs allow one to understand the difference between presentations rated *high* and *low*. If HMMs with two states ($H = 2$) and only elbows are considered (case E above), each state of the HMM encodes a particular configuration of the elbows, i.e., a pair of mean angles with the associated variances. The similarity between the two states of a given HMM can be computed as the linear distance between the related mean angles of the right elbow, and the same for the left elbow. The higher this distance, the higher the difference between the two states, i.e., the higher the angular span between the two configurations. For example, this will happen if people keep their arms and forearms along the body in some frames, and in other frames the forearms are orthogonal to the arms. In other words, higher diversity among the states means a more pronounced gesture dynamics.

This reasoning can be translated into quantitative measurements. It is possible to compute the distances between the states of each HMM, then it is possible to obtain the median of the distances for the models trained over videos of class *high* and videos of class *low*. Figure 3 shows the medians for both classes (left and right elbow are considered separately). The plot shows that distances, hence amplitude of gestures, tends to be higher for the presentations of the *low* class. In other words, when the speaker gestures too much and too widely, the presentation is less appreciated.

Finally, it is worth noting that HMMs cannot be applied to the motion flow or the audio data (see below), since in both cases we have a histogram for a single video, i.e. temporal information is absent.

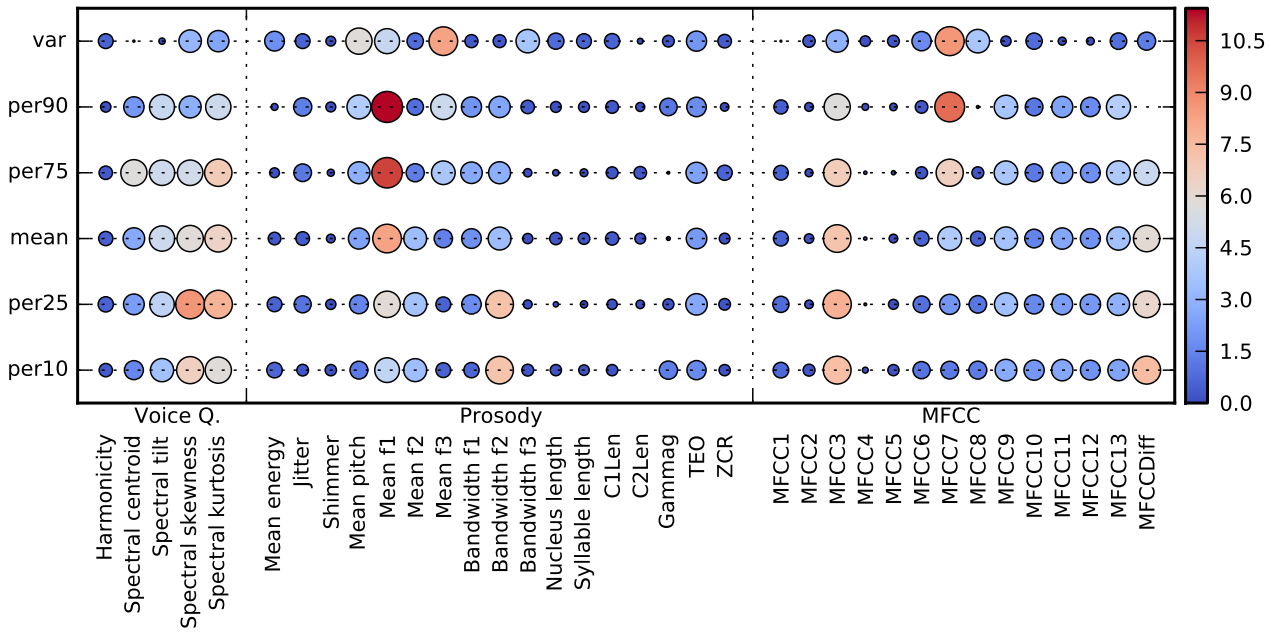

Fig. 4 Size and color of the bubbles account for the F-score of the features (the larger the bubble, the more discriminant the feature). The features are grouped into Voice Quality, Prosody and MFCC. The abbreviations "*var*", "*mean*" and "*perNN*" stand for variance, mean and NNth percentile respectively.

Body Parts	SVM	LR	HMM$_{H=2}$	HMM$_{H=3}$	HMM$_{H=4}$
S	52.0%	51.0%	64.4%	55.5%	56.6%
E	52.0%	42.0%	**66.7%**	61.1%	65.5%
SE	60.0%	58.0%	**66.7%**	62.2%	62.2%
SET	61.1%	60.0%	64.4%	64.4%	55.5%

Table 2 Results for different body parts and classifiers. All values above 50% are statistically significant and the accuracies written in boldface are the highest ones.

Preprocessing	SVM	LR
whole frames	54.9%	57.3%
100 patches whole frame	54.8%	54.8%
100 patches only speaker	50.1%	50.9%

Table 3 Classification based on the optical flow extracted from each frame of the videos. For each experiment we report the average on all the tests for the values of the parameters: N_{PCA}, E, N_C.

4.3 Experiments on Movement

The overall movement of the speaker is captured with optical flow measurements. The approach requires one to set several parameters (see Section 3 for more details), namely number of clusters N_C, number of frames N_{PCA} used to extract the Principal Components from a video, and amount of variance E to be retained when applying PCA to video frames. Table 3 reports the performances obtained by averaging over all combinations of the parameters above for the following values: $N_C \in \{50, 100, 200\}$, $E = 99\%$, $N_{PCA} \in \{50, 100\}$ when using only part of the frames, and $N_{PCA} \in \{15, 20, 30\}$, $E \in \{95\%, 99\%\}$ when computing the optical flow on the whole frame.

The results are not above chance and this seems to suggest that the overall movement on the frame does not influence the attribution of the scores. The reason is probably that the optical flow captures not only the movement of the speaker, but also the movement of the camera and/or of the background. This is likely to introduce a confounding effect.

4.4 Experiments on Prosody and Voice Quality

In the case of speech, the prediction is made after selecting the features most likely to account for the differences in ratings, i.e. after identifying the fraction V of features with the highest F-score:

$$F(x) = \frac{(E_l[x] - E[x])^2 + (E[x] - E[x])^2}{\frac{1}{n_l - 1} E_l[x - E[x])^2] + \frac{1}{n_h - 1} E_h[(x - E[x])^2]} \quad (6)$$

where $E_l[x]$ and $E_h[x]$ are the averages of feature x over class *low* and *high*, n_l and n_h are the number of samples in class *low* and *high*, and $E[x]$ is the average of feature x over the entire dataset. After the feature selection, the samples are classified with a linear kernel SVM. The fraction V (see above) and the SVM regularization parameter C are set via cross-validation: the values of V and C that lead to the highest

Feature set	SVM
Prosody	57.8%
Voice Quality	53.3%
MFCC	64.5%
all	**66.7%**

Table 4 Classification based on prosody, voice quality, MFCC and their combination. The values above 60% correspond to statistically significant differences with respect to chance ($p < 5\%$). The highest accuracy is written in boldface.

performance over the training set are retained for the test. The values of V explored during the cross-validation range between 10% and 50%, those of C in $\{10^n, -5 \leq n \leq 1\}$.

The experiments were performed using separately prosody, voice quality and MFCC features as well as using all the features together (see Section 3 for details on how features are grouped). The results (see Table 4) show that prosody and voice quality features do not work better than chance when used alone, but improve the performance of the MFCC when used in combination.

Figure 4 shows the F-scores of all features. Each column corresponds to a low-level feature while each row corresponds to one of the statistics extracted over the whole recording (the larger is the bubble the more the feature is likely to discriminate between *low* and *high* class). The values are measured over the entire dataset, but they are representative of those observed for each of the 9 folds adopted in the experiments. The features that seem to be the most discriminant are spectral skewness and kurtosis (related to the concentration of the energy on one side of the spectral centroid), MFCC coefficients (related to the distribution of the energy over the frequencies) and the mean of the first formant (related to quality of the vowels and to the phonemes being uttered).

4.5 Combination

The experiments presented so far consider audio and video separately. This section shows that the combination of the two modalities, performed at both feature and decision level (see below), leads to an improvement of the best results obtained using only audio or only video.

Combination (or fusion) of the modalities at the feature level means that that the feature vectors extracted from audio and video are concatenated. This makes it possible to apply such a combination approach only in the case of motion flow and audio. Pose estimation cannot be included because it does not produce a feature vector, but a a sequence of angle values. The combination was performed using the parameters that were most effective in the unimodal experiments: $N_C = 100$, $E = 95$, $N_{PCA} = 30$ and extraction of optical flow from the entire frame for the optical flow, the 91 features with the highest F-score in the case of audio. The concatenation of the feature vectors is fed to SVM an LR and the resulting accuracies are 66.7% and 63.3%, respectively. This does not correspond to an improvement with respect to the best unimodal approach (see above).

criteria	max	min	average	median	maj. vote
acc.	61.7 %	61.7 %	62.3%	**69.4%**	69.3 %

Table 5 Performance of the combination (decision level). The scores from the classifiers on the three different signals are combined according to the criteria proposed in [19] (see the text for more details). The highest accuracy is typed in boldface.

Combination (or fusion) at the decision level means that the classification outcomes obtained with individual modalities (only audio or only video) are used jointly to reduce the number of errors. By "classification outcome" it is meant here the score (typically a probability) that classifiers associate to a given class. In the case of SVM and LR, the score is an estimate of $p(c|\bar{x})$, where c is a class and $\bar{x}$ is the feature vector extracted from the data. In the case of the HMMs, the outcome is the confidence score described in Section 4.2. All outcomes used in this work range between 0 and 1 and are thus comparable to each other.

For the decision level combination as well, the parameters of the individual classifiers are set to the values that gave the best performances in the unimodal experiments (see above). The classification outcomes are combined using the rules descibed in [19]: *max* (the output of the classifier with the highest score is the output of the combination), *min* (the output of the classifier with the lowest score is the output of the combination), *average* (the class for which the average of the scores is the highest is the output of the combination), *median* (the class for which the median of the scores is the highest is the output of the combination) and *majority vote* (the class that has the highest score for the majority of the individual classifiers is the output of the combination).

Table 5 reports the combination results when using all individual classifiers (pose estimation, optical flow and audio). All combination rules lead to statistically significant results and the best accuracy is 69.4%, obtained with the median rule. For comparison purposes, the combination was performed using the max rule for all possible pairs of individual classifiers as well, but the results were inferior. Overall, nonverbal communication appears to influence the ratings to an extent sufficient to be modeled and predicted with statistical approaches.

5 Conclusions

This work investigated the effect of nonverbal communication on the ratings assigned to presentations posted on large online repositories such as "*Videolectures*" and "*Academic Earth*". The experiments focused on the nonverbal

cues most important in an oral presentation, namely pose, gestures, movements and prosody. The results were obtained over a corpus of 90 recordings collected from"*Videolectures*" and show that it is possible to predict, to a statistically significant extent, whether a presentation is rated as high or low in terms of overall quality (less than four or at least four stars, respectively). Furthermore, the experiments show that the combination of different cues, especially when performed at the decision level, leads to an accuracy close to 70%.

The findings confirm that the way speakers deliver their content to the audience influences the overall appreciation of a presentation. In line with the "*Media Equation*", the tendency to react to people portrayed in videos like if we meet them in person [32], the effect is observable even in the case of recordings watched through a web interface. The technical quality of the videos available on "*Videolectures*" changes significantly depending on the cases. While certain recordings are of professional quality, others barely manage to show the speaker. This limits significantly the effectiveness of pose, gesture and movement estimators. Hence, the performance achieved in the experiments is likely to be a lower bound of what it can be obtained with data of higher quality. Furthermore, editing, compression and overall video quality might influence the ratings as well and their effect should be investigated. Using only an excerpt of the videos might be a limitation even if psychologists suggest that a thin slice of behavioral evidence is sufficient to develop an impression about an individual [1].

To the best of our knowledge, this work is the first attempt to predict lecture ratings using an automatic approach, with the exception of the preliminary results presented earlier by the authors [**?**]. While showing that the task can be performed with accuracy well beyond chance, the work leaves open at least two important issues. The first is the role of the verbal content that, in this work, was fully neglected. Separating the effect of verbal and nonverbal communication is difficult. However, the safest interpretation of the results obtained in this work is probably that nonverbal behavioral cues can make an important difference (in terms of ratings) when the content of several presentations is of comparable quality.

The second issue is the role of culture on production and perception of nonverbal behavior. Web based repositories include contributions from any culture and provide access to users of any cuktural background. Therefore, it is difficult to address the problem in detail. On the other hand, the performances achieved over the data of this work seem to suggest that cultural differences (both among the speakers and the raters) still allow effective automatic processing.

The most promising applications of the approach proposed in this paper are, on one hand, the automatic assessment of material to be included in online repositories and, on the other hand, the training of speakers and teachers towards better delivery practices. Future work will explore both directions after significantly increasing the size of the corpus.

Acknowledgements This work was supported in part by Hankuk University of Foreign Studies Research Fund of 2013, in part by the European Commission through the Network of Excellence SSPNet and in part by the Swiss National Science Foundation vi the NCCR IM2.

References

1. Ambady N, Rosenthal R (1992) Thin slices of expressive behavior as predictors of interpersonal consequences: a meta-analysis. Psychological Bulletin 111(2):256–274
2. Ambady N, Rosenthal R (1993) Half a minute: Predicting teacher evaluations from thin slices of nonverbal behavior and physical attractiveness. Journal of Personality and Social Psychology 64(3):431–431
3. Babad E, Bernieri F, Rosenthal R (1991) Students as judges of teachers' verbal and nonverbal behavior. American Educational Research Journal 28(1):211–234
4. Bazzani L, Cristani M, Tosato D, Farenzena M, Paggetti G, Menegaz G, Murino V (2013) Social interactions by visual focus of attention in a three-dimensional environment. Expert Systems 30(2):115–127
5. Cristani M, Murino V, Vinciarelli A (2010) Socially intelligent surveillance and monitoring: Analysing social dimensions of physical space. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops (CVPRW 2010), pp 51–58
6. Cristani M, Paggetti G, Vinciarelli A, Bazzani L, Menegaz G, Murino V (2011) Towards computational proxemics: Inferring social relations from interpersonal distances. In: Proceeedings of IEEE International Conference on Social Computing, pp 290–297
7. Cristani M, Pesarin A, Drioli C, Tavano A, Perina A, Murino V (2011) Generative modeling and classification of dialogs by a low-level turn-taking feature. Pattern Recognition 44(8):1785–1800
8. Cristani M, Pesarin A, Vinciarelli A, Crocco M, Murino V (2011) Look at who's talking: Voice activity detection by automated gesture analysis. In: AmI Workshops, pp 72–80
9. Curhan J, Pentland A (2007) Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. Journal of Applied Psychology 92(3):802–811
10. Delaherche E, Chetouani M, Mahdhaoui A, Saint-Georges C, Viaux S, Cohen D (2012) Interpersonal synchrony: A survey of evaluation methods across disciplines. IEEE Transactions on Affective Computing 3(3):349–365

11. Fini A (2009) The technological dimension of a massive open online course: The case of the cck08 course tools. The International Review Of Research In Open And Distance Learning 10(5):1–20
12. Harris M, Rosenthal R (2005) No more teachers' dirty looks: Effects of teacher nonverbal behavior on student outcomes. In: Riggio R, Feldman R (eds) Applications of nonverbal communication, Lawrence Erlbaum, pp 157–192
13. Jayagopi D, Hung H, Yeo C, Gatica-Perez D (2009) Modeling dominance in group conversations from nonverbal activity cues. IEEE Transactions on Audio, Speech and Language Processing 17(3):501–513
14. Johnson WL, Rickel J (1997) Steve: An animated pedagogical agent for procedural training in virtual environments. ACM SIGART Bulletin 8(1-4):16–21
15. Knapp M, Hall J (1972) Nonverbal Communication in Human Interaction. Harcourt Brace College Publishers
16. Kop R (2011) The challenges to connectivist learning on open online networks: learning experiences during a Massive Open Online Course. The International Review of Research in Open and Distance Learning, Special Issue-Connectivism: Design and Delivery of Social Networked Learning 12(3):19–38
17. Kop R, Fournier H, Mak J (2011) A pedagogy of abundance or a pedagogy to support human beings? participant support on massive open online courses. International Review of Research in Open and Distance Learning 12(7):74–93
18. Krämer NC, Bente G (2010) Personalizing e-learning. the social effects of pedagogical agents. Educational Psychology Review 22(1):71–87
19. Kuncheva L (2002) A theoretical study on six classifier fusion strategies. IEEE Transactions on Pattern Analysis and Machine Intelligence 24(2):281–286
20. Liu C (2009) Beyond pixels: Exploring new representations and applications for motion analysis. PhD thesis, Massachusetts Institute of Technology
21. Mairesse F, Walker MA, Mehl MR, Moore RK (2007) Using linguistic cues for the automatic recognition of personality in conversation and text. Journal of Artificial Intelligence Research 30:457–500
22. Martin F (2012) Will Massive Open Online Courses change the way we teach? Communications of the ACM 55(8):28–30
23. Mohammadi G, Vinciarelli A (2012) Automatic personality perception: Prediction of trait attribution based on prosodic features. IEEE Transactions on Affective Computing 3(3):273–284
24. Moran M, Seaman J, Tinti-Kane H (2011) Teaching, learning, and sharing: How today's higher education faculty use social media. Tech. rep., Pearson Education
25. Morency LP (2010) Modeling human communication dynamics. IEEE Signal Processing Magazine 27(5):112–116
26. Nass C, Brave S (2005) Wired for speech: How voice activates and advances the Human-Computer relationship. MIT Press
27. Otsuka K, Takemae Y, Yamato J (2005) A probabilistic inference of multiparty-conversation structure based on Markov-switching models of gaze patterns, head directions, and utterances. In: Proceedings of ACM International Conference on Multimodal Interfaces, pp 191–198
28. Pesarin A, Cristani M, Murino V, Vinciarelli A (2012) Conversation analysis at work: detection of conflict in competitive discussions through semi-automatic turn-organization analysis. Cognitive processing 13(2):533–540
29. Petrillo M, Cutugno F (2003) A syllable segmentation algorithm for English and Italian. In: Proceedings of Eurospeech, pp 2913–2916
30. Pianesi F, Zancanaro M, Not E, Leonardi C, Falcon V, Lepri B (2008) Multimodal support to group dynamics. Personal and Ubiquitous Computing 12(3):181–195
31. Picard R (2000) Affective computing. MIT press
32. Reeves B, Nass C (1996) The media equation: How people treat computers, television, and new media like real people and places. Cambridge University Press
33. Richmond V, McCroskey J (1995) Nonverbal Behaviors in interpersonal relations. Allyn and Bacon
34. Rienks R, Heylen D (2006) Dominance Detection in Meetings Using Easily Obtainable Features. In: Lecture Notes in Computer Science, Springer, vol 3869, pp 76–86
35. Rienks R, Zhang D, Gatica-Perez D (2006) Detection and application of influence rankings in small group meetings. In: Proceedings of the International Conference on Multimodal Interfaces, pp 257–264
36. Rodriguez T Jand Plax, Kearney P (1996) Clarifying the relationship between teacher nonverbal immediacy and student cognitive learning: Affective learning as the central causal mediator. Communication education 45(4):293–305
37. Saint-Georges C, Cassel R, Cohen D, Chetouani M, Laznik MC, Maestro S, Muratori F (2010) What studies of family home movies can teach us about autistic infants: A literature review. Research in Autism Spectrum Disorders 4(3):355–366
38. Salamin H, Vinciarelli A (2012) Automatic role recognition in multiparty conversations: An approach based on turn organization, prosody, and conditional random fields. IEEE Transactions on Multimedia 14(2):338–345

39. Shaw E, Johnson WL, Ganeshan R (1999) Pedagogical agents on the web. In: Proceedings of the third annual conference on Autonomous Agents, ACM, pp 283–290
40. Smith H (1979) Nonverbal communication in teaching. Review of Educational Research 49(4):631–672
41. Tallent-Runnels M, Thomas J, Lan W, Cooper S, Ahern T, Shaw S, Liu X (2006) Teaching courses online: A review of the research. Review of Educational Research 76(1):93–135
42. Thalmann NM, Thalmann D (2012) Virtual humans: back to the future. In: Proceedings of Graphics Interface, pp 1–8
43. Vinciarelli A, Pantic M, Heylen D, Pelachaud C, Poggi I, D'Errico F, Schroeder M (2012) Bridging the gap between social animal and unsocial machine: A survey of social signal processing. IEEE Transactions on Affective Computing 3(1):69–87
44. Wharton T (2009) The Pragmatics of Non-Verbal Communication. Cambridge University Press
45. Wiggins J (ed) (1996) The Five-Factor Model of Personality. Guilford
46. Woolfolk AE, Brooks D (1983) Nonverbal communication in teaching. Review of Research in Education 10:103–149
47. Yang Y, Ramanan D (2011) Articulated pose estimation with flexible mixtures-of-parts. In: IEEE Conference on Computer Vision and Pattern Recognition, pp 1385–1392