Learning How to Teach from "Videolectures": Automatic Prediction of Lecture Ratings Based on Teacher's Nonverbal Behavior

Pietro Salvagnini*, Hugues Salamin†, Marco Cristani§,*, Alessandro Vinciarelli†,‡, Vittorio Murino*§

* Pattern Analysis & Computer Vision, Istituto Italiano di Tecnologia, via Morego 30, 16163 Genova, Italy
Email: {pietro.salvagnini, vittorio.murino}@iit.it
† Department of Computer Science, University of Glasgow, UK
Email: {hsalamin, vincia}@dcs.gla.ac.uk
‡ Idiap Research Institute - CP 592, 1920 Martigny, Switzerland
§ Dipartimento di Informatica, Università di Verona, Strada Le Grazie 15, 37134 Verona, Italy
Email: marco.cristani@univr.it

***Abstract*—Large repositories of presentation recordings (e.g., "Videolectures" and "Academic Earth") often provide their users with rating facilities. The rating of a presentation certainly depends on the content, but the way the content is delivered is likely to play a role as well. This paper focuses on the latter aspect and shows that nonverbal behavior (in particular arms movement and prosody) allows one to predict whether a presentation is rated as low or high in terms of quality. The experiments have been performed over 100 presentations collected from "Videolectures" and the accuracy is up to 66% depending on the techniques adopted. In other words, nonverbal communication actually influences the ratings assigned to a presentation.**

I. Introduction

Social media have become a common ritual in our everyday life if it is true that 75% of Internet users visit one or more social networking sites [3] each time they access the Internet. Hence, it is not surprising to observe that social media have permeated higher education practices as well: in the US, more than 60% of the faculties have used social media during their classes, 30% of them post content for their students and 40% include viewing or reading social media among their course assignments [6]. The multiplication of online lecture and presentation repositories, typically including social networking facilities, is one of the main traces of the trend above. Nowadays, it is possible to search and browse large collections of didactic presentations (more than $12,000$ recordings on the only "*Videolectures*") and even to get certificates from top universities by attending publicly available online courses (e.g., Stanford on "*Academic Earth*").

This paper aims at showing that such repositories are not only a useful source of information, but also a potential help towards developing better teaching and presentation skills. In fact, the repositories often provide rating mechanisms inspired by social media. Therefore, it is possible to rank the presentations according to ratings assigned by users and, in ultimate analysis, to identify the characteristics that discriminate between highly rated presentations and the others.

In the perspective above, this work investigates how nonverbal behavioral cues [10] displayed in online presentations influence the ratings assigned by the users. The reason is that the presentation content certainly plays a major role, but the delivery is still an important part of the process and it is likely to influence significantly the judgment of the users [11]. Furthermore, recent advances in Social Signal Processing show that nonverbal behavior is a reliable evidence for understanding human-human interactions [10]. Hence, nonverbal behavioral cues might be effective in understanding the interaction between speakers and audience as well.

The experiments have been performed over 100 presentations collected from "*Videolectures*"[1]. The users of the site have rated all presentations via a Likert scale ranging from one ("*poor quality*") to five *stars* ("*excellent quality*"). The results show that vocal behavior (in particular mean pitch) and gesturing (in particular arms movement) can predict whether a speaker is attributed at least four stars (high quality) or less than four stars (low quality). The prediction accuracy, up to 66%, is higher than chance to a statistically significant extent. Hence, the results seem to confirm that nonverbal communication influences, besides content and any other factors, the ratings assigned by the users. To the best of our knowledge, this is the first work that addresses this problem.

II. The Approach

The approach includes two main steps: the first is the extraction of features that account for the nonverbal behavior of speakers (in particular, pose, gestures, movement and prosody). The second is the automatic prediction of presentation ratings based on the features extracted at the first step.

A. Gestures and Pose Estimation

Gestures and pose have been estimated with the *pictorial structure framework* (see [12] for full details). In this model,

[1] www.videolectures.net

978-1-4673-5188-1/12/$31.00 ©2012 IEEE

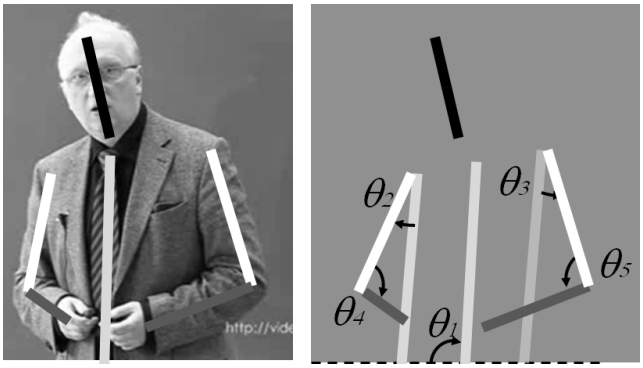

Fig. 1. The pose of the speaker is described collecting these 5 angles at each frame for which the lecturer is detected.

a person is composed of K individual parts that belong to L predefined categories (e.g., arm, head, etc.) and T types (e.g., extended, bent, etc.). Category and type of part i are denoted with p_i and t_i, respectively. The K parts respect physical constraints (they belong to a human body) represented with a graph $G = (V, E)$ where the K vertices correspond to the individual parts and the edges correspond to mutual constraints to be respected by the connected parts (e.g., the distance between upper and lower arm cannot go beyond a certain value). The graph G allows one to define a score that should be higher for configurations observed frequently where the K parts respect the constraints:

$$S(t) = \sum_{i \in V} b_i^{(t_i)} + \sum_{i,j \in E} b_{ij}^{(t_i,t_j)}. \quad (1)$$

The b parameters must be learned from training data (see Section III for more details).

If I is an image showing a person (see Figure 1 for an example), $p = (p_1, \ldots, p_K)$ contains the categories assigned to the K parts and $t = (t_1, \ldots, t_K)$ their types, the pose and gesture score is defined as follows:

$$S(I, p, t) = S(t) + \sum_{i \in V} w_i^{t_i} \phi(I, p_i) + \sum_{i,j \in E} w_{ij}^{t_i t_j} \psi(p_i, p_j) \quad (2)$$

where the w are coefficients to be learnt from the data (see Section III), $\phi(I, p_i)$ is a measurement (e.g. histogram of gradients) extracted from the image area where part i is supposed to be depicted and $\psi(p_i, p_j) = [dx, dx^2, dy, dy^2]^T$ ($dx = x_i - x_j$ and $dy = y_i - y_j$). The value of 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 find the case for which the score is maximum.

B. Movement from Optical Flow

The pose estimation algorithm provides a description of the position of the limbs but tends to be noisy and unstable, becoming unsuitable to properly describe the dynamic of the speaker's gesture. Moreover it can depict only partially the whole speaker motion, since only a simplified representation of the skeleton is adopted. Hence we employed the optical flow to integrate pose and gestures (as per estimated with the technique above) with low-level features accounting for long-term movement patterns.

The first step of the extraction process consists in reducing the dimensionality of the raw optical flow vectors (obtained as in [5]) using Principal Component Analysis (PCA). Due to memory constraints, the Principal Components are extracted from a subset of the data that includes only $NFPV$ raw optical flow vectors per video, where $NFPV$ (a hyper-parameter to be set experimentally) is much smaller than the total number of available frames.

The raw optical flow vectors are then projected onto the first N Principal Components obtained above, where N is set to retain a fraction ER of the total variance in the data (the value of ER is set experimentally). The projections so obtained are clustered with the kNN algorithm (the number NCC of clusters is a hyper-parameter to be set experimentally). In this way, each raw optical vector can be assigned to one of the clusters (after having been projected onto the Principal Components). Hence, a video can be represented with a histogram where each bin corresponds to one cluster and the value of each bin is the number of raw optical flow vectors assigned to that cluster (Bag of Words approach).

The raw optical flow vectors have been 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 extract the optical flow from 100 patches uniformly distributed on 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 (see [12]) has identified the speaker.

C. Prosody from Speech

The speech processing element of the approach focuses on prosody, i.e. on the way the speakers talk and not on what the speakers say. The reason is that the prosody is well known to influence, to a significant extent, the impression that listeners develop about a person [11]. The extraction of prosodic information includes two main stages: the first is the extraction of *low-level* features and the second is the estimation of statistics that account for the distribution of the low-level features.

In this work, the low-level features are *pitch* (oscillation frequency of vocal folds), *energy*, and *voicing probability*, i.e. the probability of voice being emitted at a given moment. The voicing probability is not used as such, but as the basis for extracting two low-level features, namely the length of voiced and unvoiced segments (time intervals during which there is emission of voice or not, respectively).

The low-level features are extracted every 10 ms from 40 ms long analysis windows using the *Snack* toolkit [9]. Hence, if a speech segment is T seconds long, a low-level feature

is extracted $T \times 100 - 4$ times (around 12000 times for two minutes of speech). Therefore, it is necessary to use statistics that account for the distribution of the low-level features in order to represent speech segments with a vector of tractable dimension. In this work, the statistics are *mean*, *variance*, *minimum*, *maximum*, *quantiles* (10%, 25%, 75% and 90%) and *entropy*. This latter is expected to capture the predictability of the low-level features and is defined as:

$$h = -\lim_{n\to\infty} \sum_{s_1,\ldots,s_n} \mathrm{P}(s_1,\ldots,s_n) \log \mathrm{P}(s_n \mid s_1,\ldots,s_{n-1}) \tag{3}$$

where $(s_1,\ldots,s_n)$ is a sequence of values of length n. The entropy is a good indicator of how difficult it is to predict the next symbol in the sequence given all the previous symbols. The sequences are assumed to come from a stationary Markov chain. Hence, the entropy can be computed with Lempel-Ziv based estimators [8]. As a result of the process above, a speech segment is represented with 36 features (9 statistics for each of the 4 low-level features).

III. Experiments and Results

The next sections describe the data, the results obtained with pose, optical flow and prosody separately, and the results obtained by combining all features with different approaches. All experiments were run using a 50-fold validation scheme. For each fold, one clip of each class was selected for the test set and the rest was used to train the model.

A. The Data

The experiments of this work have been performed over a corpus of 100 lecture recordings collected from *Videolectures*, one of the largest presentation repositories available on the web. The presentations of the corpus have been rated by the users of *Videolectures* (the rating is the average of the scores assigned individually by all users that have assessed the lecture). The value of the ratings ranges between 1 ("*poor quality*") and 5 ("*excellent quality*").

Table I reports the distribution of the scores across the presentations that have actually been rated by the users (roughly 11% of the total). The table reports the number of times each score has been assigned, the number of presentations that were assigned a given score and were considered for inclusion in the corpus, and the number of presentations actually included in the corpus, set to ensure a balanced distribution over the classes *low* (rating less than four) and *high* (rating higher or equal to four). In building the dataset we tried to select the videos in which the speaker is kept in the field of view most of the time.

The experiments were performed over one segment of two minutes extracted from each of the presentations[2]. In total, this corresponds to 3×10^5 frames.

[2]The list of videos used in the experiment is available at https://pavisdata.iit.it/data/salvagnini/RatingPrediction_VL/RP_VL_COGINFOCOM2012.pdf

rating	total	analyzed	suitable	used
1	37	27	17	16
1.5	2	0	-	-
2	52	32	18	17
2.5	4	0	-	-
3	110	25	17	17
3.5	25	0	-	-
4	272	32	27	25
4.5	114	0	-	-
5	1067	30	26	25
total	1683	147	104	100

TABLE I
DISTRIBUTION OF THE SCORES OVER VIDEOLECTURES AND THE CORPUS USED FOR THE EXPERIMENTS.

Body Parts	Angles	SVM	LR
Shoulders (S)	θ_2, θ_3	52%	51%
Elbows (E)	θ_4, θ_5	52%	42%
Shoulders+Elbows (SE)	$\theta_2, \theta_3, \theta_4, \theta_5$	**60%**	58%
SE+Torso (SET)	$\theta_1, \theta_2, \theta_3, \theta_4, \theta_5$	**60%**	**60%**

TABLE II
RESULTS FOR DIFFERENT BODY PARTS AND CLASSIFIERS. VALUES ARE TYPED IN BOLDFACE WHEN THE DIFFERENCE WITH RESPECT TO CHANCE IS STATISTICALLY SIGNIFICANT.

B. Experiments on Pose and Gestures

The pose and gesture estimator described in Section II has been trained using four videos of the corpus above manually annotated as positive examples, and the videos of the *INRIA Person* corpus as negative examples (the same data as the work in [12]). The resulting estimator has been used to extract the five angles depicted in Figure 1 for each frame of the 100 videos of the corpus (a total of 3000 angles per video). In case the speaker was not detected in a given frame the angles are all set to null value.

The angles have been used as features in different combinations, only shoulders (S), only elbows (E), shoulders and elbows (SE), and all angles (SET). The prediction results, obtained using a k-fold approach ($k = 50$) are reported in Table II. The performance is measured in terms of accuracy (percentage of samples that have been assigned to the correct class) for two different classifiers (Support Vector Machines, with radial basis kernel function, and Logistic Regression). The difference with respect to chance is statistically significant when using shoulders and elbows (only with SVM), and when using all angles (with both classifiers). On the contrary the classification performance considering only elbows angles are quite poor, this could be also due to the fact that the lower arms are the most difficult to be detected correctly by the pose estimation algorithms (see [12]). As a result these experiments seems to suggest that position and movement of arms and torso influence the attribution of the scores on Videolectures.

C. 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 II for more details): number of clusters (NCC), number of frames per video (NFPV), and

Preprocessing	SVM	LR
whole frames	55.47%	**59.69%**
100 patches whole frame	57.04%	57%
100 patches only speaker	55.33%	**60%**

TABLE III
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: NFPV, ER, NCC.

criteria	max	min	average	median	maj. vote
acc.	58.17 %	58.17 %	61%	**65.67%**	**66.17%**

TABLE IV
RESULTS FOR THE FUSION AT DECISION LEVEL. THE SCORES OUTPUT WITH CLASSIFIERS APPLIED TO DIFFERENT FEATURE VECTORS ARE COMBINED WITH DIFFERENT CRITERIA.

amount of variance to be retained when applying PCA to video frames (ER). Table III reports the performances obtained by averaging over all combination of parameters above for the following values: NCC$\in \{50, 100, 200\}$, ER$= 99\%$, NFPV$\in \{50, 100\}$ when using only part of the frames, and NPFV$\in \{15, 20, 30\}$,ER$\in \{95\%, 99\%\}$ when computing the optical flow on the whole frame.

The results are above chance only twice and this seems to suggest that the overall movement on the frame does not influence the attribution of the scores, but some information could be retrieved from the area around the speaker. It must be remembered 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 probably why no major effects are observed.

D. Experiments on Prosody

Prosody based models were trained in two phases. A first phase of feature selection was conducted using the CFS algorithm [2]. This method selects subsets of features highly correlated with the label and weakly correlated with the other features in the set. The process has retained the pitch mean as the only feature in 49 cases (out of 50 folds).

The classification was then performed with a Support Vector Machine with radial basis kernel function as implemented in libSVM [1]. The classification accuracy was 59%, statistically significantly better than chance. The result suggests that Videolectures raters are influenced by the pitch of the speakers when they assign a score.

E. Combination

The three approaches have been combined at both features and decision level. In the first case we concatenated the feature vectors from each approach; in the latter, six standard combination approaches,see [4], have been used on the confidence values from the classification experiments. The results have been obtained using the parameter values that have given the best individual performances: the mean pitch feature for the audio, SET angles for pose estimation, $NCC = 200$, $ER = 99$ and $NFPV = 50$ for optical flow measurement. The accuracies obtained by combining at the feature level are 61% and 51% for SVM and Logistic Regression, respectively. For the decision level combination we tried all the optical flow experiments around the speaker and then averaged, the accuracies are reported in Table IV. The results show that the performance can be improved with respect to the best individual classifier and, overall, it is possible to predict correctly the rating assigned by Videolectures users up to 66% of the times. The results suggest that prosody, movement, pose and gestures are diverse, i.e. carry different information. Furthermore, the results seem to suggest that nonverbal communication influences the rating assigned by the users.

IV. CONCLUSIONS

This work has 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 have focused on the nonverbal cues most important in an oral presentation, namely pose, gestures, movements and prosody. The results have been obtained over a corpus of 100 recordings collected from "Videolectures" and show that mean pitch and position of arms allow one to predict, to a statistically significant extent, whether a presentation is rated as high or low 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 above 66%.

The findings above 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 [7], 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 should be investigated.

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.

REFERENCES

[1] C. Chang and C. Lin. LIBSVM: a library for Support Vector Machines. *ACM Transactions on Intelligent Systems and Technology*, 2(3):1–27, 2011.

[2] M. Hall and L. Smith. Practical feature subset selection for machine learning. *Computer Science*, 98:181–191, 1998.

[3] A. Kaplan and M. Haenlein. Users of the world, unite! the challenges and opportunities of social media. *Business Horizons*, 53(1):59–68, 2010.

[4] L. Kuncheva. A theoretical study on six classifier fusion strategies. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 24(2):281–286, 2002.

[5] C. Liu. *Beyond Pixels: Exploring New Representations and Applications for Motion Analysis*. PhD thesis, Massachusetts Institute of Technology, 2009.

[6] M. Moran, J. Seaman, and H. Tinti-Kane. Teaching, learning, and sharing: How today's higher education faculty use social media. Technical report, Pearson Education, 2011.

[7] B. Reeves and C. Nass. *The media equation*. Cambridge University Press New York, NY, USA, New York (USA), 1996.

[8] T. Schurmann and P. Grassberger. Entropy estimation of symbol sequences. *Chaos*, 6(3):414, 1996.

[9] K. Sjölander. The snack sound toolkit. *KTH Stockholm, Sweden*, 2004.

[10] A. Vinciarelli, M. Pantic, and H. Bourlard. Social Signal Processing: Survey of an emerging domain. *Image and Vision Computing Journal*, 27(12):1743–1759, 2009.

[11] T. Wharton. *The Pragmatics of Non-Verbal Communication*. Cambridge University Press, 2009.

[12] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 1385–1392, 2011.