Automatic Analysis of Non-verbal Communication

Víctor Ponce, Sergio Escalera, Xavier Baró and Petia Radeva

Department of Applied Mathematics and Analysis, Universitat de Barcelona.
Gran Via de les Corts Catalanes 585, 08007 Barcelona, Spain.
Centro de Visión por Computador, Campus UAB, Edificio O, 08193, Bellaterra, Barcelona.
E-mail: v88ponce@gmail.com, {sergio,xevi,petia}@maia.ub.es

Abstract

Oral expression and communication is one of the most important competencies for personal, academic, professional and civic life[3]. According to the American Society of Personnel Administrators [1], it is considered that a good oral communication skill is important for obtaining a job, and for a good efficiency at work [2]. The main objective of this project is to obtain a Software tool that is able to obtain a series of features of a subject from automatic audiovisual analysis. The extraction of the features obtained from the oral and nonverbal language is something of particular interest in the analysis of psychological factors that a subject presents. This analysis is useful to improve the quality of oral communication: presentations, job interviews, etc. This is the ultimate goal of the project. The system has been applied to 15 end career project videos and presentations of fourth course students. It has been created a version that analyzes a recording and other that makes it in real-time via WebCam.

1 Methodology

This section describes the technical part of the system for the analysis of oral and gestural expression of students. The modules of the system are shown in Figure 1.

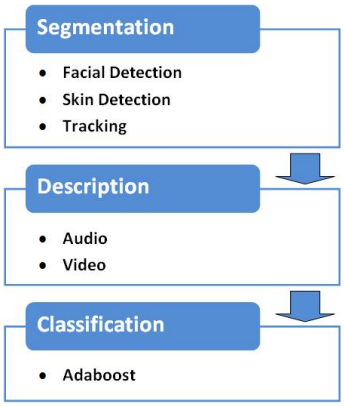

Figure 1: Schema of the system for nonverbal analysis.

1.1 Detecting interest regions

The first step corresponds to the segmentation of the person in order to isolate regions of an image which contain information of interest. In this version of the system, we focus on face and arms detection at the level of video features. For face detection we used one of the most popular methods, the face detection of Viola & Jones by cascade of classifiers [4]. This method extracts a feature set (Haar-like [4]) from images with frontal faces and background images, training a set of Adaboost classifiers that discriminates both object categories. This classifier is then tested on a multitude of the image regions at different scales and positions. The result is the detection of regions with high probability of containing a face.

Besides facial regions detected by the previous method, the pixels inside the face region are used to identify more precisely the exact skin color of the subject, and thus, finding the areas of highest probability to correspond to hands and arms [5].

Once we have found the candidate points belonging to a hand or a arm, the next step of segmentation is grouping. For this task, we define hands and arms as groups of nearby points that define a high density area. On the left side of Figure 2 we show the greater density group by squares, corresponding to the arms. On the right, we show the superposition of these squares on the original images. Once we have segmented image regions of interest, this process is repeated for all frames of the video. Since these regions are moving smoothly in the time, information on the regions of previous frames is used to strengthen future detections by a robust tracing process of regions.

1.2 Description of detected regions

Once we have identified the areas corresponding to the head, hands, and arms through the methods described in the previous section, we use the coordinates of these positions over time to extract a set of descriptors that codify information about the behavior of subject. In the work presented in [7], the authors define four general indicators that define the success of communication and evaluated them in environments of interest and dominance from social interactions. The four indicators are defined as follows:

◇ **Activity**: This is defined by the amount of speech if a subject in a dialogue.

◇ **Stress**: It is defined as the body agitation of subjects in the dialogue.

◇ **Involvement**: This includes patterns of behavior determining that a subject is "submerged" in the dialogue.

◇ **Backup mirror**: This defines the affinity between participants in a conversation from imitation gestures and speech patterns.

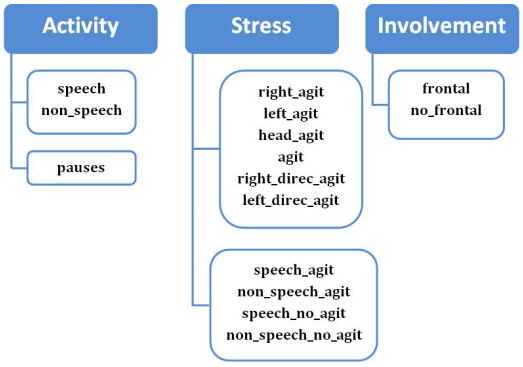

Figure 2: Grouping of features.

In our case, mirroring is not a valid indicator, since just one person appears in the video sequence. For the remaining cases we identified a set of different descriptors grouped by indicators, as shown in Figure 2. Next, we give a short description of each one.

1.2.1 Activity Descriptors

◇ *speech*: Percentage of time that has been speaking. To calculate this feature has been used a software which gets from a video with audio the vector of activation and activation of the voice over time.

◇ *non_speech*: Percentage of time that the subject has not been talking.

◇ *pauses*: Number of intervals of more than two seconds that the subject has not been talking.

1.2.2 Stress descriptors

◇ right_agit: Average of time shaking the right arm. The upheavals have been calculated from the accumulation of distances between the coordinates of regions in consecutive frames.

◇ head_agit: Average of time shaking the head.

◇ left_agit: Average of time shaking the left arm.

◇ agit: Average of time of general agitation.

◇ right_direc_agit: Quantity of rightwards shift.

◇ left_direc_agit: Quantity of leftwards shift.

◇ speech_agit: Percentage of captures with high shaking and speaking.

◇ non_speech_agit: Percentage of captures with high shaking, but not speaking.

◇ speech_no_agit: Percentage of captures with no shaking, but speaking.

◇ non_speech_no_agit: Percentage of captures with no shaking and no speaking.

1.2.3 Involvement descriptors

◇ frontal: Front capture Percentage (those frames where the subject looks to the public or the court). Although the color model can track facial lossless, we apply frontal face detector [4] to determine the percentage of frames where the subject addressed to the public.

◇ no_frontal: Percentage of catches front (Those in which the subject is not looking to the public/court).

1.2.4 Classification

The objective of our tool is to extract those patterns that distinguish the better quality of presentations of those with fewer quality, quantifying the relevance of each of them. For thus, once the regions of interest have been detected, tracked, and described, we use statistical classifiers to analyze the data regarding the quality of the presentation. In particular, Adaboost has been used for training [6]. Using Adaboost, we train a classifier which combines different simple decisions to obtain a strong hypothesis of the conversation. This method not only makes a selection of the most relevant hypothesis, but also provides a rule combination based on a weighted sum of the characteristics. Details about this algorithm can be found in [6]. In the evaluation part of the system, this method is used to find a classifier which separates between two main groups of conversations, those of higher ”quality” from those of fewer ”quality”. Moreover, it has also been used to analyze the order in which the characteristics are selected from higher to less relevance (ranking).

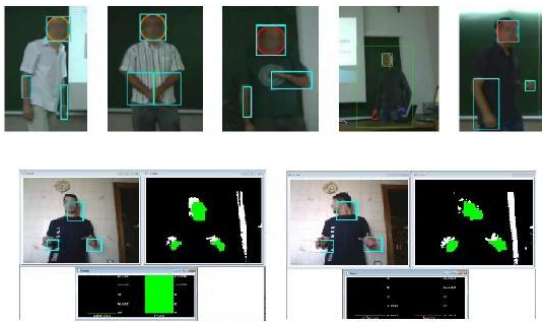

Figure 3: Examples of detected regions for different students. GUI: Detected regions on the left, segmentation on the right and statistics bars on the bottom.

2 Results

Next, we describe the data analyzed, methods and evaluation of the proposal.

◇ Data: The data analyzed consist on 15 videos recorded in presentations of final year project and 15 defenses of projects of a fourth-year elective course in Computer Science Degree from the University of Barcelona. All sequences have been recorded in frontal position towards the subject along with the tribunal, so he could capture the deviation regarding the frontal position and fixation of the subject’s gaze.

◇ Methods: we have used the system described in the previous section to analyze the videos. Some results and the final GUI are shown in Figure 3. All regions have been normalized regarding the detected facial area in order to make comparable the values of the characteristics obtained by all students. This step is important because depending on the distance from the student to the camera, the displacement of the pixels can be larger or smaller even when the agitation rate between different subjects is the same. Regarding the classifier, we allow it to do a selection of the eight most relevant characteristics.

◇ Evaluation: There have been two types of evaluations. The first consist of finding those features which correlate better the punctuation of students with behavior patterns. Although this final note is influenced by other aspects such as quality of

work and the writing of the memory, we analyze if there exists some relevant communicative part which influences the final score. The result of Adaboost classification is an ordering of the eight more relevant features that allow the best separation of the 15 presentations with top marks of the 15 presentations with poorer marks. With the three first features selected by Adaboost: *head_agit, left_direc_agit, right_agit*, both groups of presentations are correctly split based on the score obtained by the subjects.

In the second evaluation, 30 different observers evaluated the quality of presentations without knowing the contain of the presented work. These data will serve to detect if there exists any combination of features that Adaboost is able to find to agree with the observer's opinion. In this second experiment, we found that seven of the eight features selected features by adaboost match with the ones selected in the first experiment, but giving more weight to agitation features. In this evaluation, it is also possible to separate the two partitions of 15 videos by combining the values of the first three features selected by Adaboost. Both experiments were carried out considering binary problems, ergo, analyzing the characteristics that best separate within two groups of presentations.

3 Conclusion

We presented a tool for automatic analysis of oral and gestural communication of students in public presentations. The system is able to automatically detect the regions corresponding to face, hands and arms, extracting a set of features that are analyzed by statistical classifiers. Results obtained on 30 videos showed the viability and usability of the system to obtain assessments of oral and gestural expression of the students, offering a "feedback" that can be useful to improve the quality of their presentations.

The most immediate future work is to increase the discretization of the presentations score, increasing from two to N "quality" categories, in order to obtain a more accurate description of oral and gestural communication. We also want to include more accurate features for agitation and speech in order to differentiate between nervousness or involvement situations. These situations can be attacked directly by combining characteristics instead of individual indicators, for example: The student speaks continuously but he agitates without paying attention to the public.

References

[1] D.B. Curtis, J. L. Winsor, and R.D. Stephens. *National preferences in business and communication education*. Communication Education, Vol. 38 (1), pp. 6-14. 1989.

[2] J. L. Winsor, D.B. Curtis, and R.D. Stephens. *National preferences in business and communication education: A survey update*. Journal of the association of Comunication Administration, Vlo. 3, pp. 170-179. 1997.

[3] T. Allen, *Charting a communicative pathway: Using assessment to guide curriculum development in a re-vitalized general education plan.*. Communicative Education, 51(1) 26-39. 2002.

[4] P. Viola and M. J. Jones, *Robust Real-Time Face Detection*, Int. J. Comput. Vision, volumen 57, nmero 2, pginas 137-154, 2004.

[5] M. Jones and J. Rehg, *Statistical color models with application to skin detection*, IJCV, volumen 46, pginas 81-96, 2002.

[6] J. Friedman, T. Hastie and Robert Tibshirani, *Additive Logistic Regression: a Statistical View of Boosting*, Annals of Statistics, volumen 28, pginas 2000-2030, 1998.

[7] A. Pentland, *Socially aware computation and communication*, Computer, volumen 38, pginas 33-40, 2005.