Multimodal Assessment of Conversational Engagement in Persons with Parkinson's Disease

Ajjen Joshi
Department of Computer Science
Boston University
Fall 2014 Directed Study Report
ajjendj@bu.edu

December 10, 2014

Abstract

Our capacity to engage in meaningful conversations depends on a multitude of communication cues: verbal delivery of articulate and intelligible speech, tone and modulation of voice, exhibition of a range of facial expressions, and display of body gestures, among others.Parkinson's disease diminishes verbal and non-verbal communication facilities in affected persons. Occupational therapists measure the expressive behavior of persons with Parkinson's through observation and interaction. In this paper, we propose a computational model of measuring expressiveness and engagement. We record persons engaged in conversation and compute descriptors from multiple modalities (face, body, voice). We train a regression model from these multimodal features to scores obtained from domain experts and seek to answer the following questions. First, can a machine learning model be trained to accurately score expressive behavior? Second, what multimodal features are important in order to assess the level of conversational engagement? Third, what is the minimum length of observation required by the model to make an accurate prediction?

1 Introduction

The ability to effectively express ourselves is an important part of daily living as it affects every aspect of our social lives, such as forming and maintaining relationships and creating impressions. Because the quality of communication is dependent on a variety of speech, voice, emotional and gestural signals, the task of assessing the quality of conversational engagement requires the holistic study of these separate signals. When evaluating persons with communicational impairments, experts are trained to observe cues from these multimodal channels. Recent advances in computer vision and audio signal processing has enabled the accurate extraction of many multi-modal features, which we hypothesize can be used to train a computational model capable of automatically analyzing expressive behavior and conversational engagement. We focus on the problem of assessing quality of conversational engagement in persons with Parkinson's disease.

Parkinson's disease is a progressive neurodegenerative disease affecting over 1,000,000 people in the United States [15]. Parkinson's impairs patients' motor abilities with symptoms including bradykinesia and muscle rigidity. Rigidity of the facial musculature can lead to "facial masking", a loss in the ability to express facial emotions, whereas rigidity of oral and respiratory musculature hampers the capacity to effectively deliver speech [15]. Because Parkinson's affects both verbal and non-verbal channels of communication, the ability to accurately measure communication abilities becomes paramount for both patients and caregivers. A computational tool capable of doing so would help therapists diagnose and evaluate patients, as well as identify specific features responsible for the assessment score in order to provide individualized therapy.

Human behavior and movement analysis has been an important research domain in computer vision and pattern recognition. Most works deal with specific problems such as face detection [27], detection of facial alignment and landmarks [2], emotion recognition [13], gesture segmentation and recognition [19], emotion recognition from speech signals [29], action quality assessment [23] etc. Most databases that researchers work with are limited to a single modality. However, given the dependence of communication patterns on multimodal signals, efforts have been made to create and study problems that leverage signals from a variety of channels. Motor

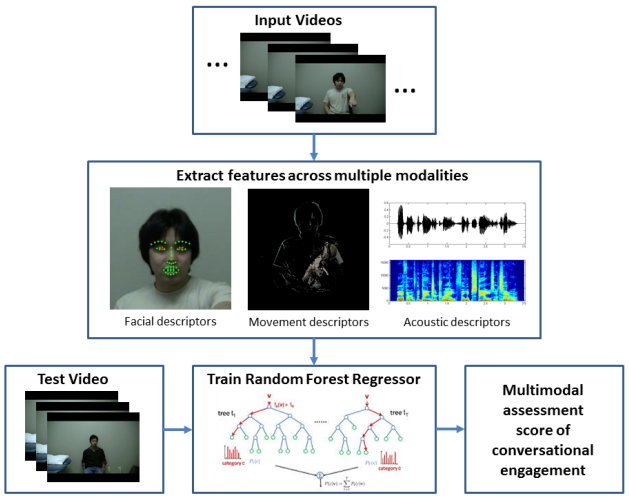

Figure 1: Overview of multimodally computing assessment score of conversational engagement.

symptoms resulting from Parkinson's disease have been studied a great deal in medicinal and rehabilitational literature, but research insights are based more on observational than on computational models [15].

Given recordings of persons with Parkinson's engaged in conversation, we compute feature descriptors from both video and audio. We train a random forest regression model from these multimodal features to scores obtained from domain experts. The aim of this work is to answer the following questions. First, can a machine learning model accurately predict the score corresponding to the level of expressive behavior? Second, what modality and combination of multimodal features is the most important in order to assess the level of conversational engagement? Third, can the model make accurate predictions by observing "thin slices" i.e. short segments of the video which are perhaps inadequate for a human observer to make a judgment?

2 Related Work

The problem of understanding human behavior by computationally analysing one or more channels of information, such as emotions, facial gestures and body movement, has been extensively studied in research. Here, we list and briefly explain some important work in the literature relevant to ours.

Vinciarelli et al. [26] provided a comprehensive survey of the domain of social signal processing, which refers to the measurement and analysis of social signals of human communication. Behavioral signals have been extracted from physical appearance [1], gestures [20], face and eye behavior [9] and vocal features [28]. The analysis of such signals involves recording of a scene where communicational signals can be observed, detecting people engaged in communication, extracting features related to appearance, gesture, facial emotion, speech and voice etc. and performing classification into social target labels [26]. Another research question that has been posed is to determine the amount of time for which a human needs to be observed in order to derive meaningful behavioral information. For example, Curhan et al. [6] used a "thin slice" of conversation in the context of employment negotiation to make accurate predictions of the eventual outcome.

The study of detecting, recognizing and analyzing the meaning of gestures has also been a popular subject of inquiry in computer vision and pattern recognition research. A comprehensive survey of gesture classification techniques have been explained by Mitra et al. [19]. Understanding and interpreting hand gestures is especially important for applications in sign language recognition [25]. Gunes et al. [12] attempted to predict human affective behavior by analyzing multimodal features from the body and face. Given the advances made in the tasks of detecting and classifying actions, Pirsiavash et al. [23] tackled the problem of assessing the quality of the actions performed.

The expression of facial emotion is an essential component of effective communication. The classification of facial emotions is another long-studied problem in computer vision. Bartlett et al. [3] performed a comparison of machine learning methods applied to the problem of emotion classification based on visual data of human faces. Metallinou et al. [18] compared the ability of facial features and voice features to predict certain emotional behavior. De Silva et al. [7] studied the relationship between posture and affective expression.

Zeng et al. [30] provided a survey of affect recognition techniques, enumerating work who use features from visual channels, audio channels as well as work that combines features from multiple modalities. Park et al. [22] argued about the benefits of combining computational descriptors from multiple communication modalities in tasks related to behavior analysis. Mehrabian [17] famously stated that perception of an individual is determined mostly by facial and bodily cues, followed by tone of voice and verbal content. Pantic et al. [21], in their survey of machine understanding of human behavior, argued for caution in combining multiple modalities, stating that information from too many channels can have the disadvantage of confusing human judges.

The study of behavioral engagement in persons with Parkinson's disease is important because of the degradation of effective communication capacities in patients.

Lyons et al. [15] discussed a method to describe expressive behavior in patients with Parkinson's disease which rely on observation of video tapes by trained personnel. Hemmesch et al. [14] reported how the onset of symptoms such as facial masking affects what impression patients make on others.

3 Datasets

There has been a significant amount of research on the computational analysis of various aspects of human behavior. Because many interesting research questions can be asked within the general domain of the computational study of human behavior, large multimodal datasets for researchers to study behavioral engagement and compare state-of-the-art methods are missing from the literature. However, due to the availability of easily available video data in services such as Youtube, the development of efficient feature extraction technologies from video, audio and text, as well as crowdsourcing platforms, such as the Amazon Mechanical Turk, which enable researchers to quickly annotate large amounts of data, the number of datasets where these topics can be explored have been increasing in recent times.

The YouTube Personality dataset [4] consists of videos collected from 404 video bloggers in YouTube, where the vloggers face a webcam and speak about a number of different topics. The dataset comes with a number of behavioral features extracted from both video and audio, as well as speech transcriptions, along with personality impression scores. Park et al. [22] have introduced the Persuasive Opinion Multimedia (POM) dataset, which contains 1,000 movie review videos, with each video containing a front-facing speaker reviewing a particular product.

Biel et al. [4] explore the possibility of classifying user personality into one of big five personality traits [16], while Park et al. [22] predict persuasiveness scores based on multimodal descriptions of their data. We are interested in assessing conversational engagement in people with impairments that prevent them from displaying normal emotional and gesticular cues. Therefore, there is the need to collect audio and visual data from persons suffering from such impairments, as well as persons capable of engaging in normal conversations, in order to build a robust computational system capable of accurately assessing conversational behavior.

4 System Overview

An overview of our conversational engagement computational model is shown in Figure 1. Here, we explain in detail the elements of our framework:

4.1 Input

The input to our framework consists of RGB videos along with corresponding audio, which have been recorded using a regular Canon consumer videocamera. Each input video contains a slice of conversation, where the user is facing the camera and answering questions posed by the interviewer. For the purpose of consistency, the videos have been recorded in a studio setting, where the distance between the camera and the user is constant and the background is plain.

4.2 Multimodal Descriptor Extraction

It is clear that conversational engagement can be observed and assessed using a combination of facial, emotional, vocal and gestural signals. Therefore, we extract feature descriptors for the face, body movement, and voice from each frame in the input video.

4.2.1 Facial Descriptors

Our system computes two optical-flow based facial features per frame. The first feature at frame t is:

$$OFdense(t) = \frac{\sum_{v \in V_d(t)} \sum_{u \in U_d(t)} \sqrt{u^2 + v^2}}{h(t)w(t)} \quad (1)$$

$OFdense(t)$ measures the average magnitude of the optical flow in all pixels inside the bounding box of the detected face normalized by the area of the bounding box. Here, $U_d(t)$ and $V_d(t)$ are respectively the sets of x and y components of the flow vector in all pixels in the bounding box of the face at frame t, $h(t)$ and $w(t)$ represent the height and width of the bounding box of the face. We use OpenCV's implementation of face detection [27] and dense optical flow [10] algorithms to compute this feature.

The second facial feature at frame t is:

$$OFsparse(t) = \frac{\sum_{v \in V_s(t)} \sum_{u \in U_s(t)} \sqrt{u^2 + v^2}}{|V_s(t)|} \quad (2)$$

$OFsparse(t)$ measures the average magnitude of the optical flow in a sparse set of pixels inside the bounding box of the detected face normalized by total number of sparse features. Here, $U_s(t)$ and $V_s(t)$ are respectively the sets of x and y components of the flow vector in the sparse set of pixels in the bounding box of the face at frame t, $|V_s(t)|$ represents the number of salient pixels whose optical flow is being measured. The sparse set of features to track are determined from the texture properties of the pixels inside the bounding box of the face [24].

Another important component of the facial descriptor is the vector containing the pitch, roll and yaw that describes the pose of the face, along with the change in those values every frame. Finally, our system computes normalized positional and velocity features for fifty different facial landmarks. The landmarks describe the position and shape of the eyebrows, eyes, axis of the nose, and mouth. Since the size of the face is not constant across users, we first normalized the positional coordinates of the facial landmark using the height of the facial bounding box as a reference.The normalized position vector of landmark j at time t is:

$$\mathbf{W}_l(t) = \frac{\mathbf{W}_l^r(t) - \mathbf{W}_{nose}^r(t)}{h(t)}, \tag{3}$$

where $\mathbf{W}_l^r(t)$ is the raw position vector for landmark l at time t, $\mathbf{W}_{nose}^r(t)$ is the raw position vector for the landmark at the tip of the nose at time t, and h is the height of the bounding box of the face. Our system uses the normalized positional coordinates W_x, W_y of all landmarks and determines the values for their velocities W'_x, W'_y. We used the face tracker by Asthana et al. [2] to localize the facial landmarks.

4.2.2 Movement Descriptors

The movement descriptor attempts to describe qualities that correspond to body movement and consists of two features computed every frame. The first feature $MHI(t)$ is the total number of pixels in a motion history image silhouette normalized by the area of the face. The motion history image is a robust method to represent motion in a sequence of images, where the intensity value of the pixel depends upon the recency of motion in the corresponding location.

$$MHI(t) = \frac{|S_t|}{h(t)w(t)} \tag{4}$$

Here, S_t is the set of all pixels of the motion history image silhouette. The second feature is the global motion orientation of the motion history image.

4.2.3 Audio Descriptors

Qualities in speech and voice are important indicators of conversational engagement. Our system uses the publicly available Covarep [8] to extract features which are commonly used in speech analysis. The features we extract include pitch, which is the fundamental frequency of the speech segment, along with formants and Mel frequency cepstral coefficients (MFCC). Formants define the spectral peaks produced by the acoustic resonance of the human voice. We use formants with the five lowest frequencies. The Mel frequency cepstrum describes the power spectrum of a sound, which can be represented by a series of coefficients (MFCc). MFCCs are used in speech and speaker recognition systems [11]. We use the first 24 MFCCs as features.

4.2.4 Data Representation

After rescaling the features so that all features are weighed equally during training, we removed the effects of noisy measurments by smoothing all features using a moving average filter spanning 5 frames. Because our input videos are temporal sequences of varying length, there arises the need to represent every data instance with a feature vector of the same length. We achieved this by dividing the gesture into 20 equal-length temporal segments, and representing each temporal segment with a vector of the median elements of all features. The representative vectors of each temporal segment were then concatenated into a single feature vector.

4.3 Regression Model Training

We defined the training set as $\mathcal{D} = \{(\mathbf{X}_1, Y_1), ..., (\mathbf{X}_n, Y_n)\}$. Here, $(\mathbf{X}_1, ..., \mathbf{X}_n)$ corresponds to the uniform-length feature vector representing each training video, and $(Y_1, ..., Y_n)$ represents their corresponding engagement scores.

A random forest regression model consists of several regression trees $\{t(\mathbf{x}, \phi_k), k = 1, \dots\}$ [5]. Here $\mathbf{x}$ is an input vector and ϕ_k is a random vector used to generate a bootstrap sample of objects from the training set $\mathcal{D}$. The prediction of the regression model is determined by taking the average over all trees in the forest. The root mean-squared generalization error (RMSE) for the model is given by:

$$RMSE = \sqrt{\frac{\sum_{i=1}^{N}(Y_i - Ypred_i)^2}{N}} \tag{5}$$

Here, Y_i and $Ypred_i$ correspond to the ground truth scores and predicted scores and N is the number of test samples. Thus, the RMSE computes the average error in the predictions made by the model.

Another method used to evaluate the accuracy of a regression model is the R^2 score, which is based on the ratio of the error made by the model to the error made by a baseline predictor that always predicts the mean score of the training data.

$$R^2 = (1 - \frac{\sum_{i=1}^{N}(Y_i - Ypred_i)^2}{\sum_{i=1}^{N}(Y_i - Ymean)^2}) \times 100 \tag{6}$$

Let d be the dimensionality of the feature vector of the inputs. At each internal node of the tree, m features are

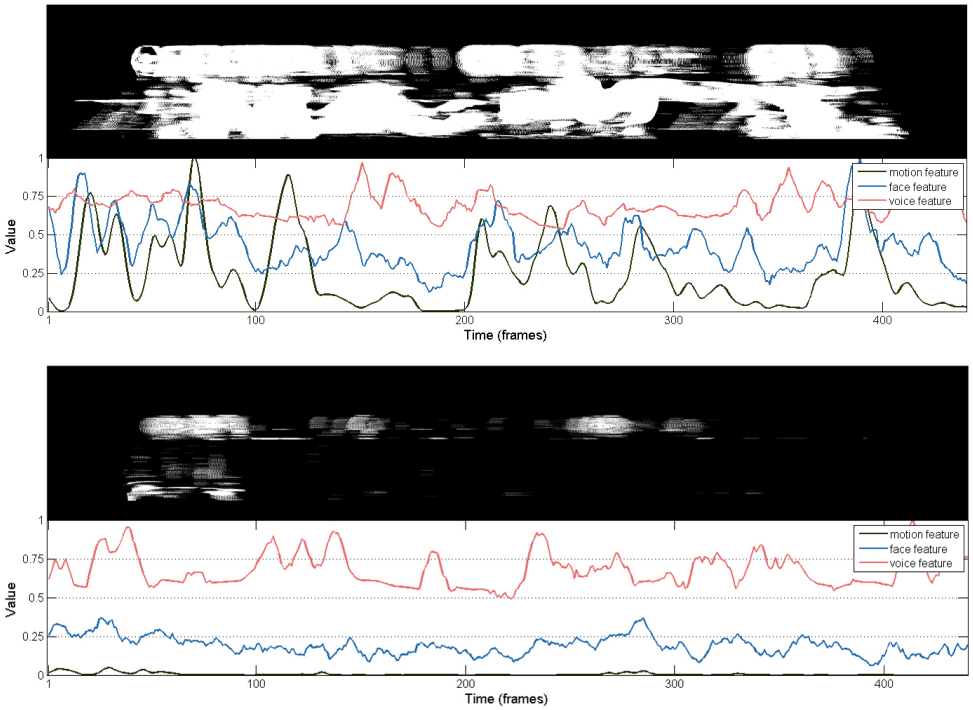

Figure 2: Comparison of visualizations of multimodal features for user with a) normal facial and gestural behavior, and b) decreased facial and gestural behavior.

selected randomly from the available d, such that $m < d$. We chose $m = \frac{d}{3}$ to train our model.

4.4 Testing

The task during training is to use our trained regression model to predict the engagement scores of our test videos. From each test video, the same features which were used to train the model were extracted and each video was represented by a uniform length feature vector as per the training procedure. The regression score for the test video was computed based on this feature vector.

5 Experiments

Here we describe the experiments performed to visualize the features computed by our system, and evaluate our system.

5.1 Visualization

Because we have yet to record data involving persons affected by Parkinson's disease, we visualized features extracted from a toy dataset. The toy dataset consists of only two video samples, recorded as a preliminary test in the same laboratory studio setting where we plan to videotape patients. One of the videos consists of a subject engaged in normal conversation with the interviewer, whereas the other video consists of a subject who was instructed to display muscular rigidity symptoms associated with Parkinson's i.e. a decreased level of face and body gesticulation.

In Figure 2, we compare the visualizatios of a small subset of the multimodal features computed by our system. The top half of the visualizations consists of a composite representation of body movement obtained by aggregating motion history images over time. The bottom half of the visualization consists of a plot of three features from different modalities over time. The motion feature MHI is the total number of pixels in a motion history image silhouette normalized by the area of the face. The facial feature $OFdense$ measures the average magnitude of the optical flow in all pixels inside the bounding box of the detected face normalized by the area of the bounding box. The vocal feature plotted in the graph measures the pitch of the voice sampled at corresponding temporal locations of the video.

It is quite clear from Figure 2 that the top visualization belongs to the subject with normal facial and gestural behavior and the bottom visualization belongs

Table 1: Comparison of personality impression scores predicted by our Random Forest regressor and Biel et al.'s SVM regressor

Models	E		C		O		A		ES	
	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2	$RMSE$	R^2
RF regressor	0.82	26	0.75	1	0.68	6	0.87	-3	.79	-5
SVM regressor [4]	0.80	36	0.74	10	0.67	10	0.87	-2	.80	-4

to the subject with restrained facial and gestural behavior. The composite motion image for the first subject consists of a lot of white segments corresponding to the amount of movement. In comparison, the motion image for the second subject contains a lot less white pixels. Similarly, the features representing facial and body movement have consistently higher values for the subject with normal facial and gestural behavior. For this pair of videos, the pitch of the voice is the only feature where the computed values do not clearly differentiate a subject with normal facial and body gesticulation from a subject with restrained facial and body gesticulation. From this preliminary analysis, we believe that conversational engagement of subjects can be predicted well if the appropriate set of features is chosen.

5.2 Regression

Since we do not yet possess a dataset suitable to train a sophisticated regression model, we used the dataset provided by Biel et al. [4] to perform some primary experiments related to finding a mapping from the feature representation of multimodal data to some behavioral score.

The dataset consists of videos collected from 404 video bloggers in YouTube, where the vloggers face a webcam and speak about a number of different topics. The authors do not provide the original video data; instead, a feature representation of each of the videos is made available to the research community. Each video is represented by a set of 25 attributes: 21 collected from speaking activity and prosody cues and 4 collected from movement cues. Each of the video samples are given personality impression scores on a 7-point scale in five criteria: Extraversion (E), Agreeablesness (A), Conscientiousness (C), Emotional Stability (ES), and Openness to Experience (O). The scores were collected via the crowdsourcing platform of Amazon Mechanical Turk.

We evaluated our regression model independently for each personality impression criterion in the task of predicting scores associated with each sample. For each regression task, we trained a random forest regression model using a 10-fold cross-validation approach. That is, the dataset was divided randomly into 10 non-overlapping folds. At each iteration, the model was trained on samples from 9 folds and tested on the samples from the remaining fold. The evaluation scores were then averaged over all iterations.

Table 1 illustrates the performance accuracy of our regression model in comparison with the SVM regressor used by Biel et al. [4]. The SVM regressor used by the authors is trained using an RBF kernel on an expanded feature set that includes attributes representing look, pose and other multimodal cues. These attributes were not provided with the dataset. The performance of our baseline model compares favorably with that of the original authors, as can be seen from the table. Both $RMSE$ and R^2 scores are similar across both models for all personality traits. Both regressors have a positive R^2 score for three of the five traits (Extraversion, Conscientiousness, and Openness to Experience) indicating better performance on these criteria, with the best scores being achieved for the Extraversion personality impression score.

6 Conclusion

In this paper, we have presented motivations along with a framework for studying an important problem: the computational assessment of conversational and behavioral engagement in people suffering from certain motor impairments resulting from conditions such as Parkinson's disease. We have described methods to extract multimodal features from videos of subjects. We believe these features extracted from gestural cues of the face and body, along with speech and prosodic attributes contain information that can be mapped to behavioral engagement scores, which are assessed by experts through observation and interaction. We hypothesized the effectiveness of these features by visualizing them for two subjects, one of whom was instructed to deliberately restrain the urge to gesticulate and emote as is done during normal conversation.

Finally, we tested our baseline random forest regression model on a separate dataset and achieved comparable performance accuracy to that produced by the original authors using only a subset of the features they used.

The next step in this work is to gain access to or collect data involving subjects suffering from varying levels of motor impairments, and test our hypotheses.

References

[1] Parham Aarabi, Dominic Hughes, Keyvan Mohajer, and Majid Emami. The automatic measurement of facial beauty. In *Systems, Man, and Cybernetics, 2001 IEEE International Conference on*, volume 4, pages 2644–2647. IEEE, 2001.

[2] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Incremental face alignment in the wild. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1859–1866. IEEE, 2014.

[3] Marian Stewart Bartlett, Gwen Littlewort, Mark Frank, Claudia Lainscsek, Ian Fasel, and Javier Movellan. Recognizing facial expression: machine learning and application to spontaneous behavior. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 2, pages 568–573. IEEE, 2005.

[4] Joan-Isaac Biel and Daniel Gatica-Perez. The youtube lens: Crowdsourced personality impressions and audiovisual analysis of vlogs. *Multimedia, IEEE Transactions on*, 15(1):41–55, 2013.

[5] Leo Breiman. Random forests. *Machine learning*, 45(1): 5–32, 2001.

[6] Jared R Curhan and Alex Pentland. Thin slices of negotiation: predicting outcomes from conversational dynamics within the first 5 minutes. *Journal of Applied Psychology*, 92(3):802, 2007.

[7] P Ravindra De Silva and Nadia Bianchi-Berthouze. Modeling human affective postures: an information theoretic characterization of posture features. *Computer Animation and Virtual Worlds*, 15(3-4):269–276, 2004.

[8] Gilles Degottex, John Kane, Thomas Drugman, Tuomo Raitio, and Stefan Scherer. Covarep-a collaborative voice analysis repository for speech technologies. In *IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP'14), Florence, Italy*, 2014.

[9] Rana El Kaliouby and Peter Robinson. Real-time inference of complex mental states from facial expressions and head gestures. In *Real-time vision for human-computer interaction*, pages 181–200. Springer, 2005.

[10] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, pages 363–370. Springer, 2003.

[11] Todor Ganchev, Nikos Fakotakis, and George Kokkinakis. Comparative evaluation of various mfcc implementations on the speaker verification task. In *Proceedings of the SPECOM*, volume 1, pages 191–194, 2005.

[12] Hatice Gunes and Massimo Piccardi. A bimodal face and body gesture database for automatic analysis of human nonverbal affective behavior. In *Pattern Recognition, 2006. ICPR 2006. 18th International Conference on*, volume 1, pages 1148–1153. IEEE, 2006.

[13] Hatice Gunes and Massimo Piccardi. Bi-modal emotion recognition from expressive face and body gestures. *Journal of Network and Computer Applications*, 30(4): 1334–1345, 2007.

[14] Amanda R Hemmesch, Linda Tickle-Degnen, and Leslie A Zebrowitz. The influence of facial masking and sex on older adults' impressions of individuals with parkinson's disease. *Psychology and aging*, 24(3):542, 2009.

[15] Kathleen Doyle Lyons and Linda Tickle-Degnen. Reliability and validity of a videotape method to describe expressive behavior in persons with parkinson's disease. *American Journal of Occupational Therapy*, 59(1):41–49, 2005.

[16] Robert R McCrae and Oliver P John. An introduction to the five-factor model and its applications. *Journal of personality*, 60(2):175–215, 1992.

[17] Albert Mehrabian. Silent messages. 1971.

[18] Angeliki Metallinou, Sungbok Lee, and Shrikanth Narayanan. Audio-visual emotion recognition using gaussian mixture models for face and voice. In *Multimedia, 2008. ISM 2008. Tenth IEEE International Symposium on*, pages 250–257. IEEE, 2008.

[19] Sushmita Mitra and Tinku Acharya. Gesture recognition: A survey. *Systems, Man, and Cybernetics, Part C: Applications and Reviews, IEEE Transactions on*, 37 (3):311–324, 2007.

[20] Sylvie CW Ong and Surendra Ranganath. Automatic sign language analysis: A survey and the future beyond lexical meaning. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 27(6):873–891, 2005.

[21] Maja Pantic, Alex Pentland, Anton Nijholt, and Thomas S Huang. Human computing and machine understanding of human behavior: a survey. In *Artifical Intelligence for Human Computing*, pages 47–71. Springer, 2007.

[22] Sunghyun Park, Han Suk Shim, Moitreya Chatterjee, Kenji Sagae, and Louis-Phillipe Morency. Computational analysis of persuasiveness in social multimedia: A novel dataset and multimodal prediction approach. *International Conference on Multimodal Interaction, Proceedings of*, pages 50–57, 2014.

[23] Hamed Pirsiavash, Carl Vondrick, and Antonio Torralba. Assessing the quality of actions. In *Computer Vision–ECCV 2014*, pages 556–571. Springer, 2014.

[24] Jianbo Shi and Carlo Tomasi. Good features to track. In *Computer Vision and Pattern Recognition, 1994. Proceedings CVPR'94., 1994 IEEE Computer Society Conference on*, pages 593–600. IEEE, 1994.

[25] Thad Starner, Joshua Weaver, and Alex Pentland. Real-time american sign language recognition using desk and wearable computer based video. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 20(12): 1371–1375, 1998.

[26] Alessandro Vinciarelli, Maja Pantic, and Hervé Bourlard. Social signal processing: Survey of an emerging domain. *Image and Vision Computing*, 27(12):1743–1759, 2009.

[27] Paul Viola and Michael J Jones. Robust real-time face detection. *International journal of computer vision*, 57 (2):137–154, 2004.

[28] Bogdan Vlasenko, Björn Schuller, Andreas Wendemuth, and Gerhard Rigoll. Frame vs. turn-level: emotion recognition from speech considering static and dynamic processing. In *Affective Computing and Intelligent Interaction*, pages 139–147. Springer, 2007.

[29] Sherif M Yacoub, Steven J Simske, Xiaofan Lin, and John Burns. Recognition of emotions in interactive voice response systems. In *INTERSPEECH*, 2003.

[30] Zhihong Zeng, Maja Pantic, Glenn I Roisman, and Thomas S Huang. A survey of affect recognition methods: Audio, visual, and spontaneous expressions. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(1):39–58, 2009.