

Predicting Performance on a Repetitive Task through Automatic Analysis of Facial Feature Movements

Maria E. Jabon, Sun Joo Ahn, Jeremy N. Bailenson

Every day countless human errors occur around the globe. While many of these errors are harmless, just a few disastrous errors, such as Bhopal, Chernobyl, or Three Mile Island, convince us that developing ways to improve human performance is not only desirable but crucial.

Considerable research exists in the area of Human Error Identification (HEI), a field devoted to the development of systems to predict human errors¹. However, these systems typically predict only instantaneous errors, not the many other possible factors involved in human performance such as speed or productivity. Furthermore, they often rely upon pre-defined hierarchies of errors and minute-by-minute analyses of users by trained analysts, making them costly and time-consuming to implement¹.

The current work proposes a novel bottom-up approach to human performance prediction using facial features automatically extracted from short video segments of participant faces during laboratory experiments. This bottom-up approach maximizes data usage and gives us the power to create performance models on three different layers: *instantaneous errors* (individual errors occurring at any time during the task), *phase level* performance (speed and accuracy on a single iteration of the task) and *task level* performance (speed, accuracy, and productivity over the entire task). This enhancement takes our model beyond the capabilities of current HEI techniques, which focus solely on predicting instantaneous errors, and expands its practicality in implementation.

Using computer vision also eliminates substantial costs associated with standard HEI techniques. Instead of devoting the time and resources to train and employ analysts, we use

computers to process our data. This allows for models that, once trained and verified, can run in real-time without human intervention. Furthermore, by using computers we eliminate inconsistencies in human analyses; computer vision can iteratively apply a model across many users, providing reliable and individualized statistics. Lastly, computer vision is notably unobtrusive; it is able to execute facial tracking and analysis with just a small web camera mounted in a work area while an individual remains within the context of a controlled environment. These benefits would allow for widespread adoption, adding to the practicality of our models.

Background and Motivation [pullout]

In the past, researchers have shown great interest in using micro-momentary facial expressions to predict human emotion and mental states¹⁻⁴. In a recent study, El Kaliouby and colleagues⁵ developed a general computational model to recognize six classes of complex emotions for facial affect inference and implemented this model as a real-time system. Their approach used dynamic Bayesian networks for recognizing emotions. In a related study that used computer vision to detect emotional states, Bailenson and colleagues² demonstrated that automated models could be developed to detect and categorize the felt-emotion of individuals. By training machine learning algorithms that link certain facial movements to subjective perception of emotions, Bailenson and colleagues were able to create real-time models that classified three emotions (i.e., sad, amused, and neutral) based on facial features and physiological responses. Picard and colleagues^{6,7} have also demonstrated across a number of systems that tracking various aspects of the face can give insight into the mental state of the person whose face is being tracked.

In a similar fashion, systems have been developed to model more complex human states, such as deception. Meservy et al.⁸ extracted macro features such as head and hand position and angle from video cameras taken during an experiment where a mock theft took place. After the theft, participants were either truthful or deceptive in an interview with a trained researcher regarding the mock theft in the lab. Using machine learning algorithms, the team was able to create models that obtained up to 71% correct classification of truthful or deceptive participants based on just the features extracted from the video recording of the subject.

Picard and colleagues in the Affective Computing Research Group have advanced the field of behavior prediction to include affective state prediction. Particularly interested in the development of Affective Learning Companions (also referred to as the Affective Intelligent Tutoring System), Picard emphasizes that technology that recognizes the non-verbal and affective cues of the user and responds accurately to these cues will be most effective in human-computer interaction⁶. Some of the current work from the Affective Computing Research Group includes using multi-modal sensory inputs for the following purposes: 1) to predict frustration in a learning environment, 2) to detect as well as respond to the learner's affective and cognitive state, and 3) to develop portable aids which track, capture, and interpret facial-head movements of other people to assist individuals diagnosed with autism spectrum disorders in social interaction⁹⁻¹¹.

Our work extends previous work by introducing several new features. First and foremost, our model goes beyond mere categorization of emotion and instead links facial movements directly to both error *and* performance using a bottom-up approach linking facial features directly to outcome. Moreover, we approach the problem with three different temporal layers: instantaneous, phase level, and task level. The face, with its ability to create over 10,000

expressions, has greater variability than any other channel of nonverbal cues. We propose that automated facial feature tracking allows researchers to tap into a rich resource of behavioral cues.

The Current Approach

Our approach to modeling human performance by means of computer vision and machine learning follows the general pattern used in many machine learning classification problems. In this pattern, raw data is collected and segmented into discrete chunks. General metrics, known as features, are extracted from these chunks. The features are analyzed to find the most predictive subset, and finally, models are trained using the best features.

To create our performance models, we first collected videos and performance logs of participants performing a repetitive laboratory task. We synchronized the videos with the task performance logs and segmented the videos into meaningful chunks based upon cues in the performance logs. An example of a meaningful chunk might be the time interval preceding one error instance or one phase of the task. We then extracted key facial points from the videos, such as the mouth and eye positions. We calculated both time and frequency domain statistics over the facial points in each segment, and ranked these features according to their chi-square value. Finally, using the highest ranked features, we trained machine learning classifiers to predict participant performance on the entire task (i.e. task level performance), in each phase of the task (i.e. phase level performance), and at any given instant within the task (i.e. instantaneous errors). Figure 1 depicts the entire process.

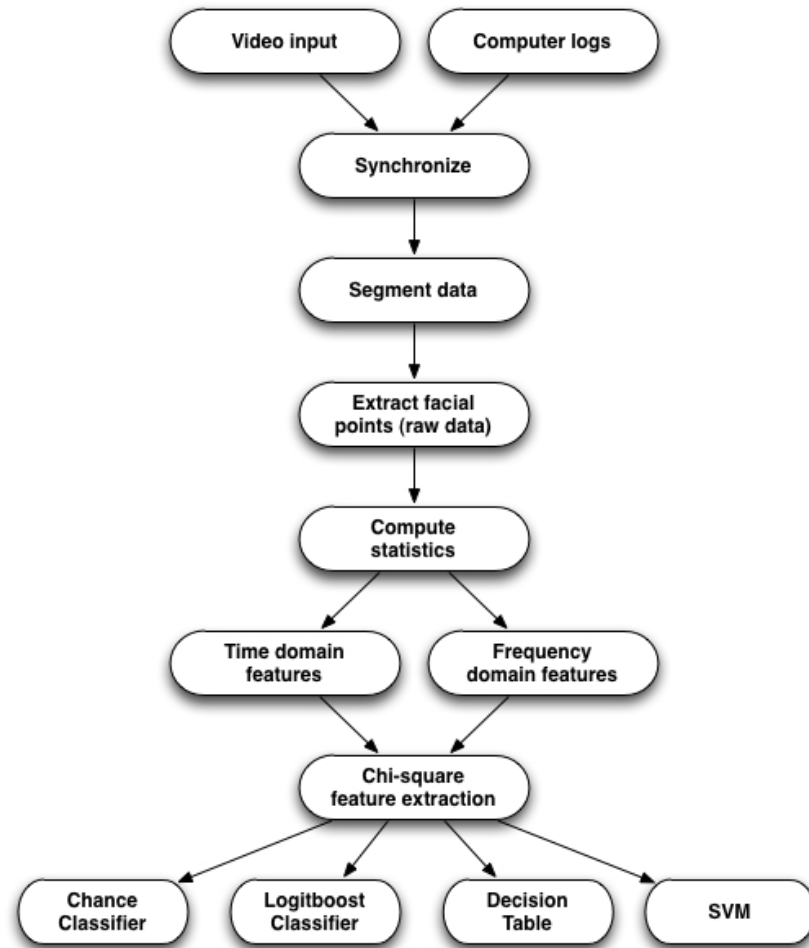


Figure 1: Steps in predicting human performance based upon facial videos

Task Setup

The task used in our experiment simulated an assembly line and was an iteration of the simple, monotonous operation of fitting screws into designated holes for half an hour. The session was administered at a computer station with a flat screen monitor adjusted to a resolution of 640x480 pixels. On the left-hand side of the screen participants were presented with three boxes (Figure 2A), each containing a screw with a different part number. In the center of the screen there was a large wooden board (Figure 2B) with seven holes labeled with a randomly selected set of the different part numbers.

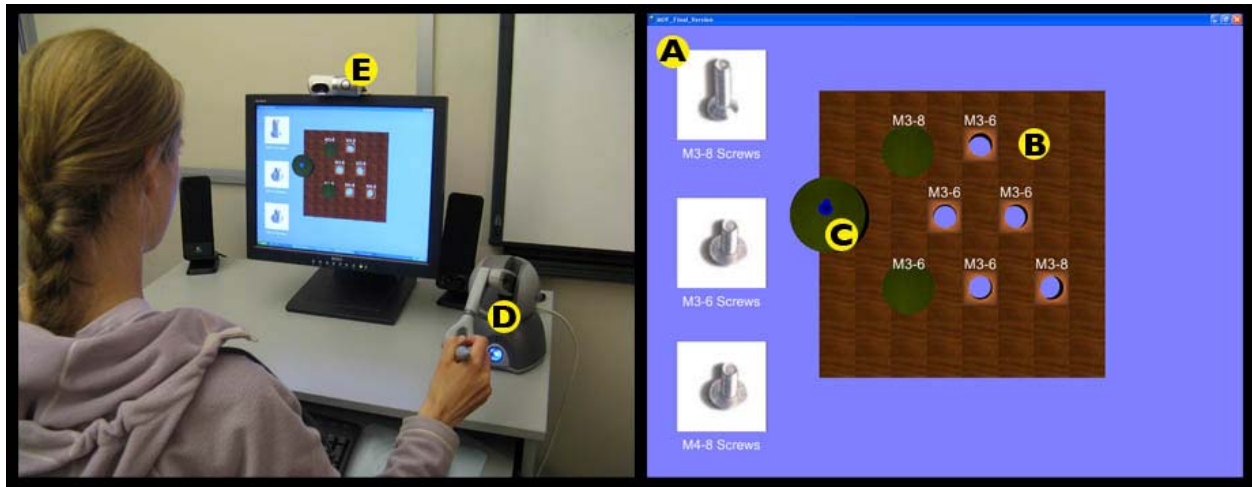


Figure 2. Experimental setup
 A) virtual screw box B) virtual board C) virtual screw D) haptic pen E) webcam

To perform the task, participants had to pick up a screw (Figure 2C) from one of the virtual boxes using a Sensable Phantom Omni haptic pen (Figure 2D) and insert it into a hole with the correct part label. The haptic pen, a device with six degrees of freedom (x, y, z, pitch, yaw, and roll), allowed the user to “feel” the hardness of the box and the depth of the screw. This way, they felt as if they were really pushing a screw into the board. Success or failure to screw in the parts was indicated by a beep; when a participant successfully screwed in a screw by holding the screw inside the hole for one second, he/she heard a beep. The lack of a beep signaled that the participant had not screwed in the part completely. The wooden boards were programmed to refresh to a new board with a new set of seven empty holes after a pre-programmed amount of time regardless of the participants’ progress. Each presentation of a board was considered to be one ‘phase’ of the experiment. The first board was timed to refresh after 45 seconds but every time the participant successfully filled two consecutive boards without any errors (indicating that the level of difficulty was too low), the given phase time was curtailed by three seconds. This ensured that the task would not be too easy or too difficult, allowing the participant to work at a pace that was feasible yet challenging for him/her.

A high resolution Logitech QuickCam UltraVision web camera (Figure 2E) affixed to the top of the monitor captured the participant's face at a rate of 15 frames per second for the duration of the experiment. Videos were compressed to .avi format using standard video recording software (Video Capturix 2007). This procedure allowed us to capture and save every facial movement of our participants for future analysis. Performance logs from the task were also recorded on the local desktop machine. Performance measures included time stamps for each error (defined as placing a screw in an incorrect hole, dropping a screw, or not holding the screw in place until the beep was heard), correctly placed screw, and board refresh. In addition, the amount of time the participant spent holding screws was also measured. In this way we could measure the instantaneous performance (i.e. errors) of the participant as well as the performance in a given phase (i.e. how quickly and accurately the participant filled out one board) and the performance over the entire experiment (i.e. overall error rate, speed of completion) and synchronize these measures with the facial videos.

A total of 57 students were recruited for our experiment, with data from eight of these participants discarded due to technical problems during data collection. Thus, the final results are based on logs and videos from 21 female and 28 male participants.

Feature Computation

As a next step we extracted all the facial feature points from the facial videos using the commercially-available OKAO vision library. This library, developed by OMRON Corporation, automatically detects and tracks 37 points on the face, head movements such as pitch, yaw, and roll, and two meta-features: eye and mouth openness level. The points are tracked relative to the captured frame. However, for the purpose of our calculations we standardized all points to be

relative to the center of the face. Figure 3 presents a screenshot depicting the OKAO face tracking points.

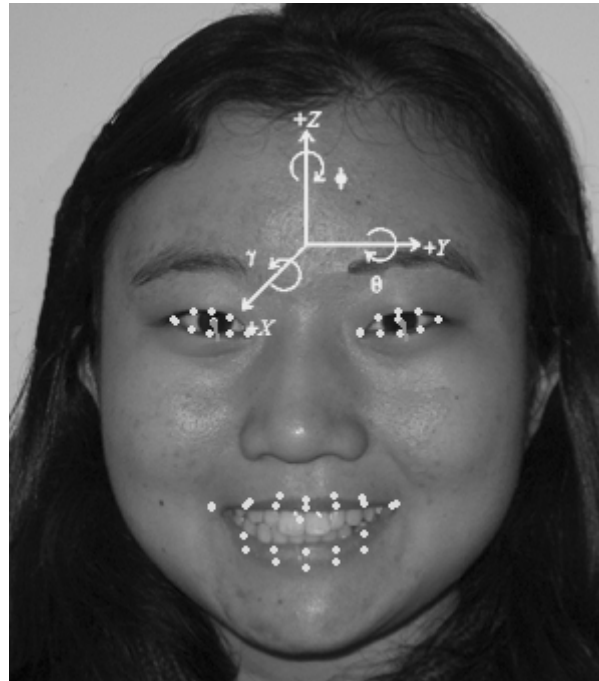


Figure 3. OKAO computer vision algorithm tracking points on participant's face

After facial feature extraction, we synchronized our video footage and computer logs to ensure correspondence between facial feature input and performance measures. This was done by manually recording the time in the video when a bell chimed (time zero in our performance logs) and adding that timestamp to all the timestamps in our performance logs. In this way we assured synchronization to match within 10 ms. We then programmatically segmented the data according to our three temporal levels of prediction: instantaneous, phase, and task. Task length intervals corresponded to all of the data for one participant, phase length intervals corresponded to all of the data for one board completion, and instantaneous intervals consisted of one second running intervals collected throughout the experiment. We discarded any intervals in which the

average face-tracking confidence (i.e., the measure of how confident the face tracking software was in its measurement) was lower than 60%.

From the raw facial data in each temporal level we then computed many features. These included both time and frequency domain statistics calculated over each facial point. We calculated these additional statistics because facial signals are dynamic, and their momentary movements can leak information about the internal state of the person making the expression². Particularly in pre-error situations, we expected many changes in signal dynamics. Thus, we calculated means, medians, minimums, maximums, standard deviations, and ranges for each point. In addition, we performed a Wavelet transform on each of the points. We used the MATLAB Wavelet toolbox to perform the discrete Wavelet transform, in particular the Daubechies Wavelet family with orders one, two, and four. For each order, we performed a level three decomposition of the input signal and collected statistics over the detail coefficients of each level including averages, ranges, histograms, and variances. In previous research, such wavelet transforms were applied successfully for identifying drowsiness with input biometric data such as the EEG³. In Figure 4, we present an example of an original signal and the wavelet decomposition for the “right eye openness” over one phase of the experiment. Note the y coefficients at different scales.

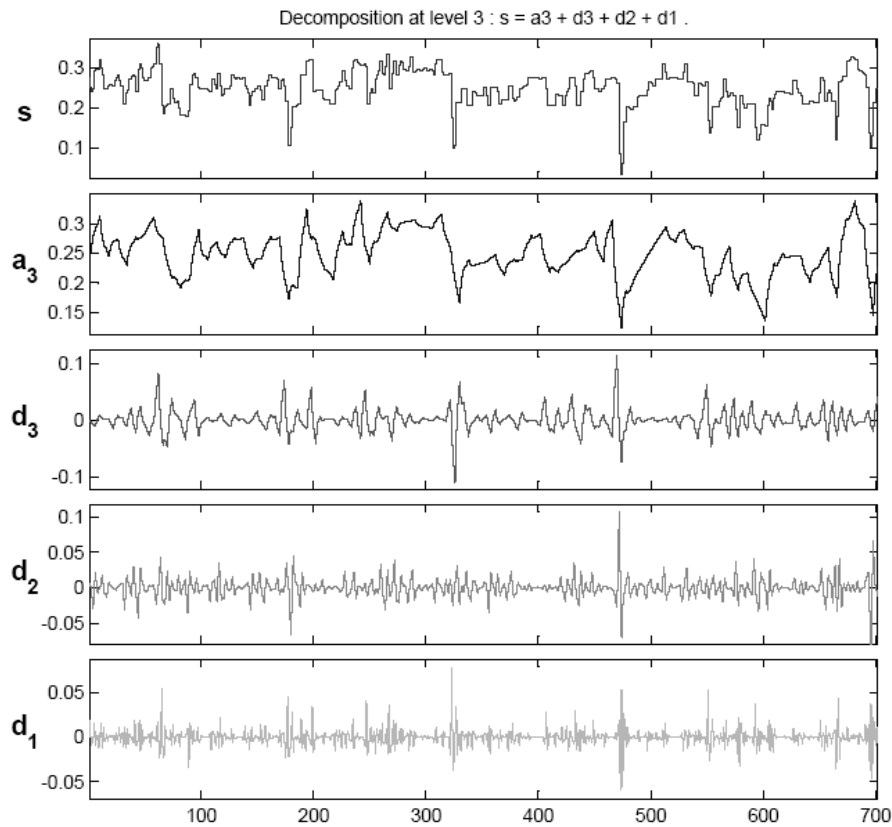


Figure 4: Wavelet decomposition of “right eye openness” level. s represents the original signal, a_3 the order one decomposed signal, and d_1 - d_3 represent the three levels of decomposition.

After statistical calculation, the complete set of raw and summary features consisted of 4242 features for each input instance. In order to speed up the training of our algorithms, prevent over-fitting, and identify which features were most useful in predicting performance, we performed a chi-square feature selection on each dataset. A chi-square selection evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class⁴. This method is recommended when the numerical force of the training set is not large in relation to the number of features⁴. We performed the chi-square analyses using the freely distributed machine learning software package Waikato Environment for Knowledge Analysis (WEKA)⁵. To determine the optimal cutoff for features to keep in our analyses we performed a

best first search starting with the top ranked chi-square feature and adding features according to chi-value. Similar selection methods have been shown successful in other classification problems with numerous features, such as Arabic text classification⁶ and gene-based cancer identification⁷. We found the optimal cutoff to vary depending on the type of classifier and the dataset. However, we found in general increasing above 10 features added little to no improvement in results. Thus, we set a maximum of 10 on the number of features we used for each classification. This reduced the complexity of our models and ensured that feature computation and predictions could be made in less than 10 ms. Table 1 presents the top 10 chi-square features for each analysis. Figure 5 depicts the meaning of these features and Figure 6 shows the performance curve of the task, phase, and instantaneous performance predictions with varying numbers of features. Note the power of the wavelet analysis; on average, 73% of the top features were wavelet coefficients of face signals.

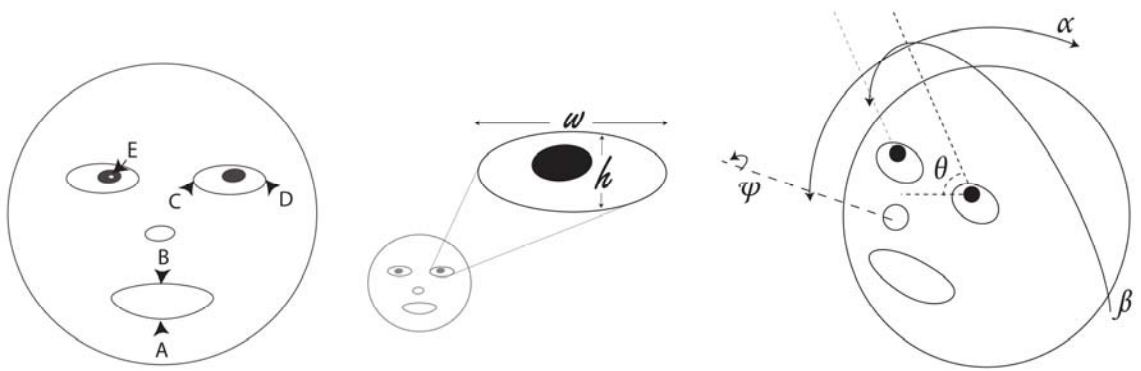


Figure 5: Significant facial feature definitions

	Statistic	Feature	Definition (as depicted in Fig. 5)	Chi Value
Task	Ave	Ave Y	Vertical position of face	108.5
	Max	Ave Y	Vertical position of face	78.90
	Wav	Gaze Tilt	θ (radians)	70.50
	Min	Ave Y	Vertical position of face	64.53
	Wav	Gaze Tilt	θ (radians)	63.56
	Ave	Lower Lip Center Y	A (y coordinate)	61.80
	Wav	Right Eye Open Level	h (cm)	55.31
	Wav	Right Eye Ratio	w/h	55.25
	Wav	Gaze Tilt	θ (radians)	54.26
	Wav	Right Eye Ratio	w/h	52.50
	Phase	Wav	Right Eye Open Level	h (cm)
Wav		Gaze Tilt	θ (radians)	204.0
Wav		Left Inner Eye Corner Y	C (y coordinate)	200.9
Wav		Right Eye Ratio	w/h	198.5
Wav		Right Pupil Y	E (y coordinate)	198.2
Wav		Left Lower Lip Y	A (y coordinate)	197.0
Wav		Left Outer Eye Corner Y	D (y coordinate)	195.4
Wav		Pitch	α (radians)	195.4
Wav		Right Inner Eye Corner X	C (x coordinate)	193.2
Wav		Left Eye Open Level	h (cm)	193.1
Instantaneous		Vel	Roll	ψ (radians)
	Vel	Yaw	β (radians)	449.7
	Vel	Roll	ψ (radians)	424.6
	Wav	Left Outer Eye Corner Y	D (y coordinate)	379.1
	-	Eye Per Close Rate	% time both eyes $h < .15$	370.9
	Wav	Ave X	Horizontal position of face	370.6
	Wav	Left Lower Lip Y	A (y coordinate)	367.1
	Wav	Left Upper Lip Y	B (y coordinate)	356.0
	Wav	Left Pupil Y	E (y coordinate)	354.3
	Wav	Left Upper Lip X	B (x coordinate)	353.2

Table 1. Top 12 chi-square features for each prediction

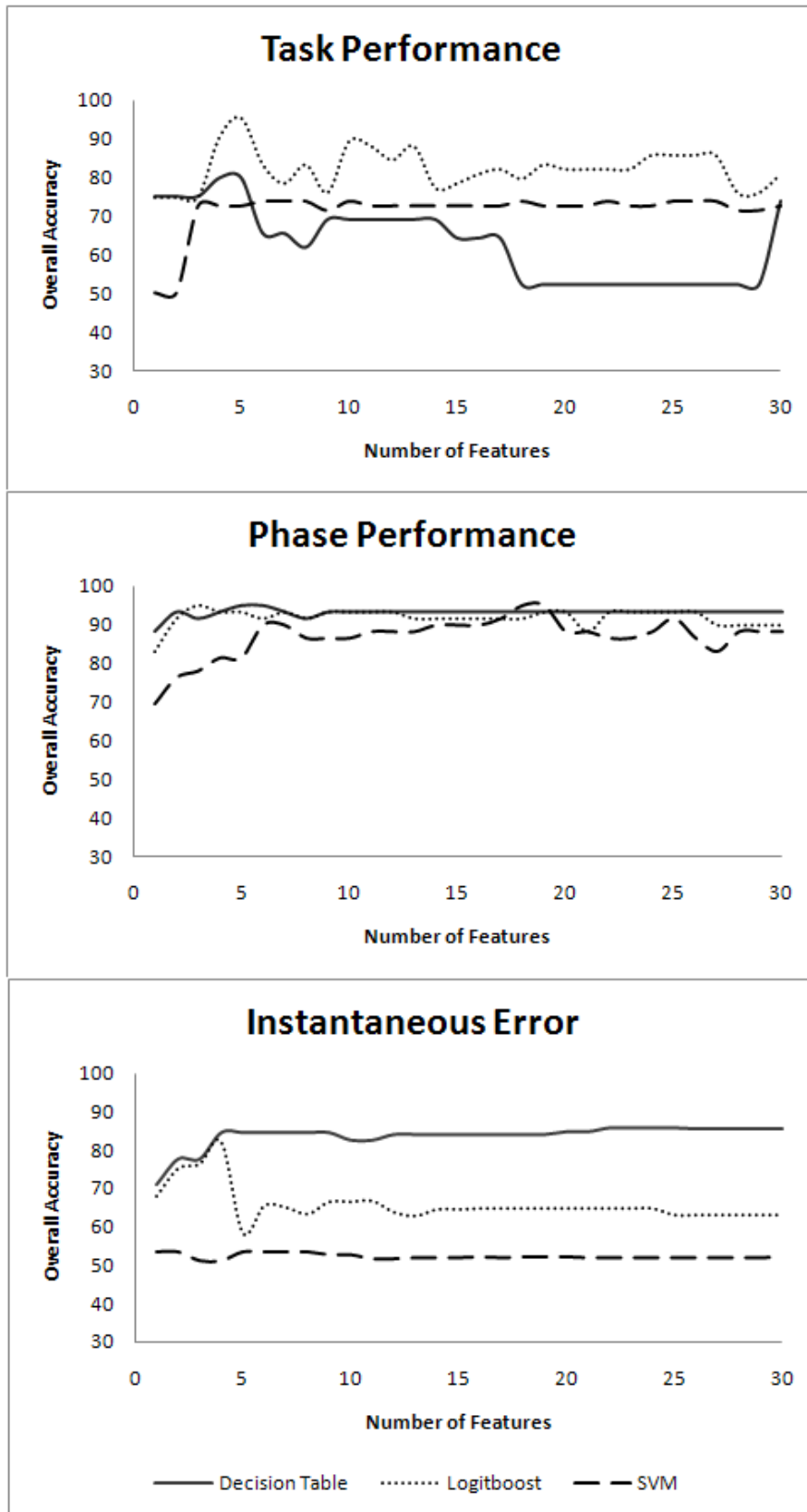


Figure 6: Accuracy of performance predictions using different numbers of features.

Performance Prediction

We experimented with numerous state of the art classifiers to model human performance. Classifiers explored included Decision Tables, Support Vector Machines, LogitBoost classifiers, and Bayesian nets. We found Decision Tables and LogitBoost classifiers the most powerful in predicting human performance. Given a training set of instances and an un-labeled instance I , the Decision Table classifier will search for the set \mathcal{L} of instances that match the features of I . It will then predict I to be of the majority class in \mathcal{L} . If $\mathcal{L} = \emptyset$, then the majority class of the training set is predicted⁸. If the features are continuous they are discretized by the median value. In a way, this mirrors the cognitive process used by humans in many existing HEI techniques, where humans will learn what sequence of behaviors (i.e. features) lead up to an error by observing many examples (i.e. a training set). LogitBoost classifiers work by sequentially applying a classification algorithm to re-weighted versions of a dataset in order to make a series of classifiers⁹. A majority vote of these classifiers then determines the final class prediction of unknown instances⁹. For our LogitBoost classifier we chose a simple decision stump classifier as our base algorithm and built a LogitBoost classifier by performing 40 boosting iterations on the decision stump.

In order to gauge the performance of our classifiers we calculated three main measures for each classifier: *overall accuracy*, *precision*, and *recall*. *Overall accuracy* is simply the total number of correctly classified instances. *Precision* is the number of instances correctly predicted to be in a class divided by the total number of instances for that class. Precision can be thought of as the percentage of predictions that are correct for a given class. *Recall* is the total number of instances correctly predicted to be in a class divided by the total number of instances predicted to be in that class. Recall can be thought of as the percentage of the instances labeled to be in a

class that actually were in that class. Note that overall accuracy is a gross measure of performance, while precision and recall are measures of the accuracy within an individual class.

We also calculated the overall accuracy, precision, and recall of a *chance classifier* for each classification. We define a chance classifier as a classifier that would naively guess the class of an instance using the proportional split of the data; if the dataset consisted of 25 error instances and 75 non-error instances the chance classifier would classify an instance as error 25% of the time and as non-error 75% of the time. The chance overall accuracy, precision, and recall are the same performance measures as defined above but calculated on a chance classifier. In order to gauge the significance of our results we looked at the overall accuracy, precision, and recall of our classifiers relative to the chance values.

Results

Task Level Performance Prediction

Our first goal was to predict task performance of the participants using only the first few minutes of facial data in their videos. In order to determine the overall performance of our participants we calculated many statistics for each participant such as how many phases the subject completed, fastest phase completion time, number of phases completed at maximum speed, mean error rate, amount of time spent holding screws (a rough gauge of productivity), and mean number of filled holes per box. We normalized all values to be between 0 and 1 and added the values to create an overall performance score for each participant. We took the participants with overall performance scores in the top quartile to be high performers and those in the bottom quartile to be low performers and combined their data into one large dataset. We then split this dataset into two independent subsets - a test and train set - by selecting test instances from only a

certain random subset of participants (ordered by participant ID) and train instances from a different, non-overlapping subset of participants. This assured the generalizability of our results across individuals. Subsequently, we extracted the segments of data corresponding to each participant's initial phases to use in our classification algorithms. We trained a Decision table classifier and a LogitBoost classifier to predict the high performing and low performing participants. The results of these classifications using the differing numbers of phases as input are presented in Figure 7. We found the results varied with differing numbers of phases, but the general trend was in increase in accuracy with more phases used. We did note significant drops in performance with certain numbers of phases, such as 12 phases in the Decision table analysis, which could be caused if certain phases were less predictive than others. If this were the case including them would result in reduced accuracy.

While the overall accuracy peaked for both classifiers when 20 phases were used as input, the overall accuracy obtained with 10 phases was only 2-10% lower than overall accuracy obtained with 20 phases. This demonstrates the power of our approach; using only very small amounts of facial data we can gauge participant performance over the entire task. If early prediction were essential in an application, just 10 phases (five to seven minutes) of data would be sufficient data to classify participants as high or low performers. Detailed results from the most successful analyses are presented in Table 2.

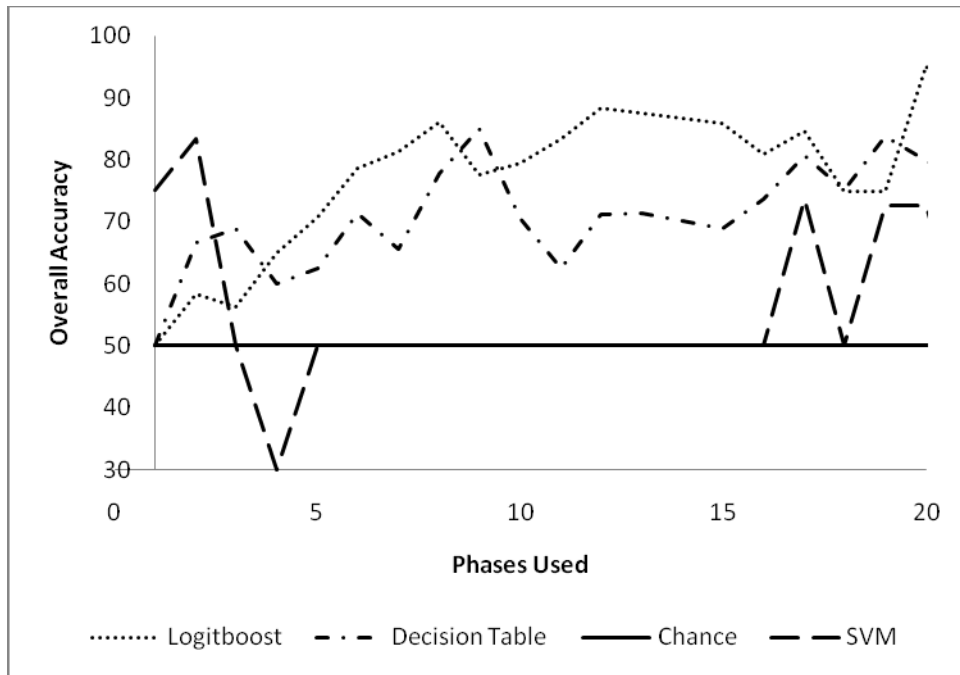


Figure 7. Face inputs predict task performance using different amounts of data.

Classifier	Overall accuracy	Class	Precision	Recall
Chance Classifier	50.0%	High	50.0%	50.0%
		Low	50.0%	50.0%
LogitBoost Classifier	95.2%	High	91.3%	100%
		Low	100%	90.5%
Decision Table	79.8%	High	71.2%	100%
		Low	100%	59.5%
SVM	72.6%	High	65.1%	97.6%
		Low	95.2%	47.6%

Table 2. Face inputs predict task performance using 20 phases of data.

We noted that the use of boosting significantly improved results; the LogitBoost performed almost 15% better than the simple Decision Table, classifying participant performance with 95.2% accuracy. This accuracy is over 44% higher than chance level for our dataset. Recall was also notably strong for the high performers in both algorithms, indicating a low false alarm rate for high performers. The SVM performed notably worse than the Decision Table and LogitBoost classifier, suggesting the SMO algorithm may not be well suited for performance prediction.

Phase Level Performance Prediction

In our next analysis, we attempted to predict participant performance on any given phase of the experiment. To do this we first calculated the lowest phase time achieved by each participant. Since phase time was curtailed until the participant was no longer able to completely finish a box in the allotted time, the lowest phase time indicated the participant's maximum

speed. Then, for each participant, we extracted all the phases in which he or she was working at his calculated maximum speed, made two or fewer errors, and filled at least half the board. We then attempted to detect these high performance phases from any phase that did not meet the high performance criteria.

To do this, we again split the data into independent test and train sets based upon participant ID to avoid participant overlap. We then trained and tested a Decision Table classifier and a LogitBoost classifier on the data from the training dataset and tested it on the data from the test set. The results of these classifications are presented in Table 3.

Classifier	Overall accuracy	Class	Precision	Recall
Chance Classifier	51.5%	High	58.6%	58.6%
		Low	41.4%	41.4%
LogitBoost Classifier	94.9%	High	90.6%	100%
		Low	100%	90.0%
Decision Table	94.9%	High	90.6%	100%
		Low	100%	90.0%
SVM	89.8%	High	84.8%	96.6%
		Low	96.2%	83.3%

Table 3. Face predicts phase performance.

Again, the face proved to be a powerful predictor of performance; both the LogitBoost and Decision Table classifiers performed with overall accuracies greater than 90%, which is over 40% higher than chance levels. Both classifiers had a high performance phase recall of 100% and a low performance phase precision of 100%. This indicates a low false alarm rate for both high performance phases and low performance phases.

Instantaneous Performance Prediction

In our final analysis we attempted to predict instantaneous errors (defined as dropping a screw, screwing a screw into the wrong hole, or not holding a screw in the hole long enough) using only the facial data collected from our videos. To do this we compiled a dataset consisting of all the *pre-error intervals* (defined as the window of facial data of length I beginning D seconds before the error, where I ranged from one to five seconds and D ranged from one to three seconds). We then added to the dataset an approximately equal number of randomly selected intervals that did not contain errors.

Again, we split our dataset into two independent subsets - a test and train set - by selecting test instances from only a certain subset of participants and train instances from a different, non-overlapping subset of participants and trained a LogitBoost classifier and a Decision Table classifier on the training data and tested on the test data. The results of these classifications are presented in Figure 8. In the top panel, D is held constant at one second while I is varied between one and five seconds. In the bottom panel I is held constant at two seconds while D is varied between one and three seconds. Detailed results for the most successful classification using an I of two seconds and a D of one second are presented in Table 4.

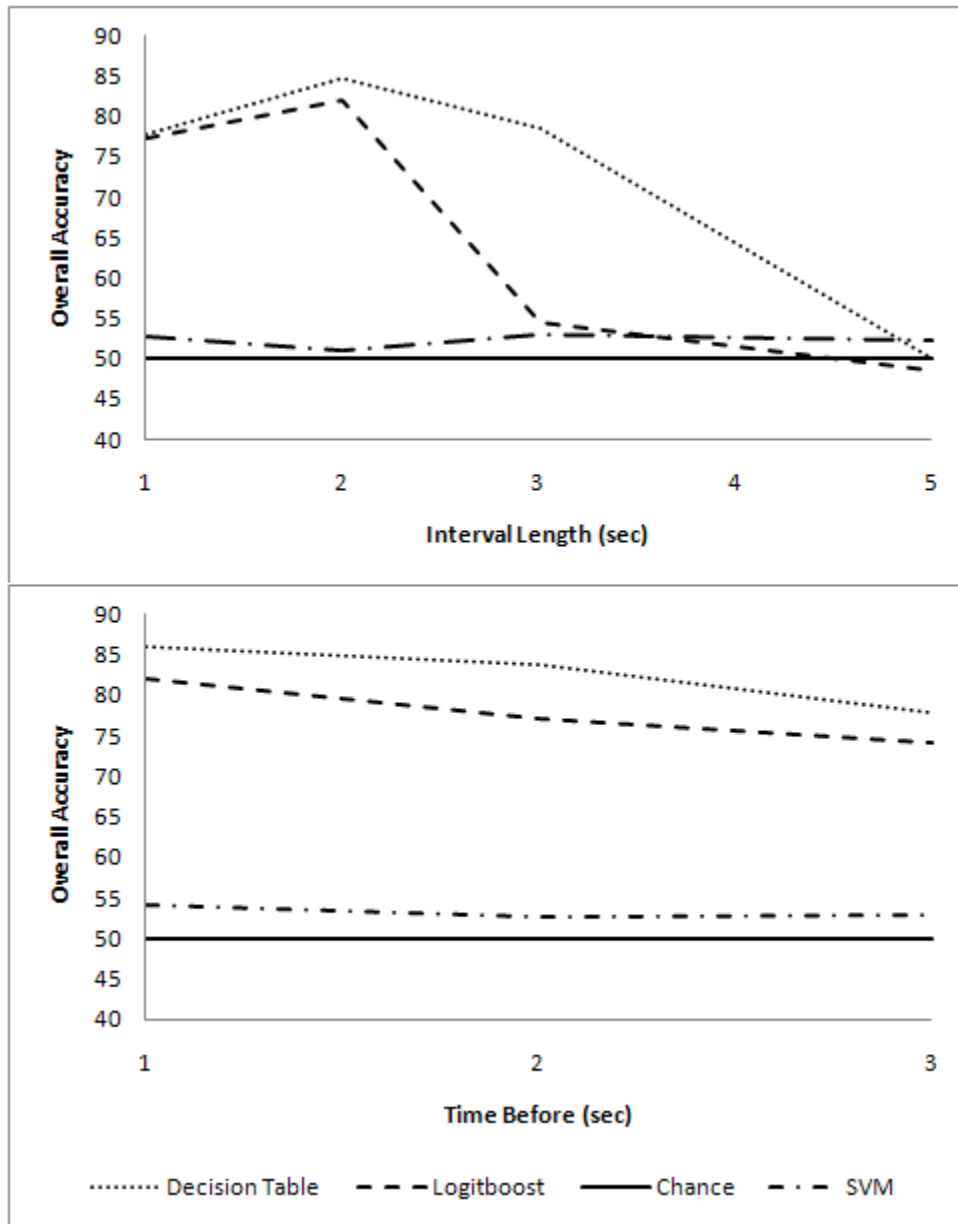


Figure 9. (Top) Face inputs predict errors one second before they occurred using different interval lengths. (Bottom) Face inputs predict errors at different times before using two seconds of data.

Classifier	Overall accuracy	Class	Precision	Recall
Chance Classifier	50.1%	Error	47.2%	47.2%
		Correct	52.8%	52.8%
LogitBoost Classifier	82.0%	Error	84.1%	75.7%
		Correct	80.4%	87.5%
Decision Table	84.7%	Error	83.2%	84.1%
		Correct	86.0%	85.2%
SVM	53.3%	Error	0%	0%
		Correct	53.3%	100%

Table 4. Face inputs predict errors two seconds before they occur using one second of data.

As Figure 7 shows, performance peaked using two seconds of data. This suggests that the micro-momentary expressions indicative of errors may be very short in duration (less than three seconds). We also found that accuracy increased as we reduced the time interval before prediction, indicating the expressions occur at one second or less before an error. Results using one second of data from one second before error instances yielded encouraging results. The Decision table performed with an overall accuracy of 84.7 %, a full 34.6 % over chance levels.

While perhaps not yet at a level of accuracy suitable for all applications, our models could benefit many by reducing errors and maintaining an optimal level of production with performance monitoring. Implications for system applications are discussed in the following section.

System Use

Our results, most of which are in the 90th percentile, indicate that our approach of using facial movements to predict errors and model human performance has significant potential for actual application. Human error plays a role in many industrial incidents; the nuclear power industry attributes up to 70-90% of failures to human error, and other industries report similar rates (e.g. airline 90%, medicine 98%)^{10,11}. In these safety-critical applications, warnings could be issued to eliminate costly, or even deadly, incidents of error.

By presenting both a micro and a macro view of human errors made on a task, our model enables managers to gain a better understanding of worker performances on the task level and the phase level in addition to individual error instances. This is an important contribution as simple aggregates of individual errors do not necessarily demonstrate the person's performance during the duration of the entire task.

Our task level performance prediction could be implemented to pick out individuals more suited to a particular task. Companies could use such a model to pre-screen employees or to match individuals to jobs they have a higher aptitude for, saving time and resources. For instance, if individuals were able to find out their future performance on a task from only a short video clip taken of their face, it would be of great assistance in the search for a perfect job-to-person fit. Similarly, instructors could gauge the abilities of students with just a short task, and then tailor coursework to the specific abilities of individual students.

Next, our phase level performance prediction could be used to gauge worker state in any given segment of a task and then take appropriate actions (i.e., a break when workers showed signs of fatigue). This would allow for preemptive (versus reactive) actions to optimize production capacity and prevent unfortunate accidents. Furthermore, as the detection, tracking, and training/learning process of our models are completely automated, this can free up valuable

human resources and reduce overhead costs. Our results implicate that we are able to outperform human analysts with only a fraction of the effort required in the traditional HEI systems. Thus, our model can either completely substitute human data analysts or supplement their work as an additional layer of analysis to build an error detection system that is close to infallible as possible and which also goes on to predict future performance.

In addition, due to the ease of adoption (only a small webcam and processor are needed to run our models), individuals and corporations alike could reap the benefits of performance prediction at little cost, allowing more widespread use, and thus greater potential to avoid error.

System limitations

Despite these encouraging results and important implications for real-life applications, the current study has several limitations. The most obvious would be that although the models may be generalized across individuals, they cannot be generalized across tasks; the models only predict the performance quality of an individual in our particular task, not in any general task. Furthermore, our experiment was conducted in a lab setting, not in a real-life work environment. One can imagine in certain environments the face would not be visible, in which case our models would be unable to predict an error. In future work we could explore the possibility of using the same framework to predict errors and model performance on other, more general tasks and in more naturalistic settings. In addition, the models could be benchmarked against more traditional HEI methods.

Another limitation is that the models are based on arbitrary definitions of error and performance quality, and may yield completely different results with different definitions of these elements. For instance, if we were to discard our comprehensive definition of error and

choose finer grained classifications of error, our models may not be as successful. Similarly, if we used different measures to classify our participants as high performers our models may yield different results. Similarly, our choice of phases greatly impacted the task level prediction, and different combinations of input layers may offer different results. In sum, automated facial recognition is a systematic and objective methodological tool, but it must be understood that the established models are heavily dependent on our definitions of measured items and chosen input data which may be arbitrary or subjective.

Finally, the facial recognition and machine learning technology we use is not new, only our applications of it are. Thus the performance of our models is closely related to the quality of our software. In situations where the face is obscured and tracking is lost, performance cannot be predicted. Also, the OKAO vision library tracks only 37 points on the face and may not yield accurate results for people wearing glasses. One could envision a library that automatically detects and tracks more points on the face and works around obstructions on the face such as glasses. Similarly, we only used existing classifier models; unique classifiers could be designed and more specifically tailored to the task of human performance prediction. In future work alternative algorithms could be explored.

Conclusion

Even with the technical challenges and limitations discussed above, we believe that our video-based performance prediction system demonstrates potential for many applications. Although it is difficult to expect any system, including humans, to make perfectly accurate judgments on processes as complex as human behavior, our results are encouraging and we anticipate improvements with future research. By incorporating these models into the workplace

and learning environment, they can serve as effective, cost-efficient, and unobtrusive monitors that assist both individuals and corporations in reaping maximum safety and output.

Acknowledgements

We would like to thank Ritsuko Nishide, Shuichiro Tsukiji, Hiroshi Nakajima, and Kimihiko Iwamura from OMRON Corporation for partial funding of the project. We would also like to thank Suejung Shin and Steven Duplinsky for their help in making images, and Joris Janssen, Kathryn Segovia, Helen Harris, and Solomon Messing for helpful comments on earlier drafts of this paper.

References

1. P. Salmon, *et. al.*, "Predicting Design Induced Pilot Error: A comparison of SHERPA, Human Error HAZOP, HEIST and HET, a newly developed aviation specific HEI method," *Proc. of the Tenth International Conference on Human-Computer Interaction*, 2003, pp. 567-571.
2. P. Ekman and E. L. Rosenberg, *What the face reveals: Basic and applied studies of spontaneous expression using the Facial Action Coding System (FACS)*, New York: Oxford University Press. 1997.
3. M. A. Schier, "Changes in EEG alpha power during simulated driving: a demonstration," *International Journal of Psychophysiology*, vol. 37, no. 2, pp. 155-162, 2000.
4. C. Manning, P. Raghavan, and H. Schütze, *Introduction to Information Retrieval*, Cambridge University Press. 2008.
5. H.I.Witten and E. Fank, *Data mining: Practical machine learning tools and techniques*, second edition. San Francisco: Morgan Kaufmann. 2005.
6. P. Kosla, *A Feature Selection Approach in Problems with a Great Number of Features*, Heidelberg: Springer Berlin. 2008. pp. 394-401.
7. Xin Jin, Anbang Xu, Rongfang Bie' and Ping Guo, *Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profile*, Springer Berlin / Heidelberg. 2006. pp. 106-115.
8. R. Kohavi, "The Power of Decision Tables," *8th European Conference on Machine Learning*, pp. 174-189, 1995.

9. J.H. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: a Statistical View of Boosting," Stanford University, 1998.
10. J. W. Senders and N. Moray, *Human error: Cause, prediction, and reduction*, Series in applied psychology. Hillsdale, N.J.: L. Erlbaum Associates, 1991.
11. A. Isaac, "Human error in European air traffic management: the HERA project," *Reliability Engineering and System Safety*, vol. 75, no. 2, pp. 257-272, Feb 2002.

Pullout References

1. Zhihong Zeng, Maja Pantic, G. I. Roisman, T. S. Huang, "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, pp. 39-58, January, 2009.
2. J.N. Bailenson, E.D. Pontikakis, I.B. Mauss, J.J. Gross, M.E. Jabon, C.A. Hutcherson, C. Nass, and O. John, "Real-time classification of evoked emotions using facial feature tracking and physiological responses," *International Journal of Human Machine Studies*, vol. 66. 2008. pp. 303-317.
3. R. Picard and J. Klein, "Computers that Recognize and Respond to User Emotion: Theoretical and Practical Implications," *Interacting with Computers*, vol. 14. no. 2. 2002.
4. Y.S. Shin, "Recognizing Facial Expressions with PCA and ICA onto Dimension of the Emotion," *Structural, Syntactic, and Statistical Pattern Recognition*, Heidelberg: Springer Berlin. 2006. pp. 916-922.
5. R. El Kaliouby and P. Robinson, "Mind reading machines: automated inference of cognitive mental states from video," *Systems, Man and Cybernetics, IEEE International Conference on*, vol. 1. 2004. pp. 682-688.
6. R. Picard, *Affective computing*. Cambridge: MIT Press. 1997.
7. R. W. Picard and K. K. Liu, "Relative participative count and assessment of interruptive technologies applied to mobile monitoring of stress." *Int. J. Hum.-Comput. Stud.* vol. 65. no. 4. pp. 361-375. 2007.
8. T.O. Meservy, M. L. Jensen, J. Kruse, J.K. Burgoon, and J.F. Nunamaker Jr., "Deception detection through automatic, unobtrusive analysis of nonverbal behavior," *IEEE Intelligent Systems. Sept/October.* vol. 20. no. 5. pp. 36-43. 2005.

9. A. Kapoor, W. Bursleson, and R. Picard, "Automatic prediction of frustration," *International Journal of Human-Computer Studies*, vol 65. pp. 724-736. 2007.
10. S. D'Mello, T. Jackson, S. Craig, B. Morgan, P. Chipman, H. White, N. Person, B. Kort, R. el Kaliouby, R. Picard, and A. Graesser, "AutoTutor detects and responds to learners affective and cognitive states," *Workshop on Emotional and Cognitive Issues at the International Conference of Intelligent Tutoring Systems*, Montreal, Canada. June 23-27, 2008.
11. M. Madsen, R. el Kaliouby, M. Goodwin, and R. W. Picard, "Technology for just-in-time in-situ learning of facial affect for persons diagnosed with an autism spectrum disorder," *Proceedings of the 10th ACM Conference on Computers and Accessibility (ASSETS)*, Halifax, Canada. October 13-15, 2008.