Automatically Analyzing Facial-Feature Movements to Identify Human Errors

Maria E. Jabon, Sun Joo Ahn, and Jeremy N. Bailenson, *Stanford University*

Using facial feature points automatically extracted from short video segments, researchers couple computer vision with machine learning to predict performance over an entire task and at any given instant within the task.

Every day countless human errors occur around the globe. Although many of these errors are harmless, disastrous errors—such as Bhopal, Chernobyl, and Three Mile Island—demonstrate that developing ways to improve human performance is not only desirable but crucial. Considerable research exists in human-error identification (HEI), a field devoted to developing systems to predict human errors.[1] However, these systems typically predict only instantaneous errors, not overall human performance. Furthermore, they often rely on predefined hierarchies of errors and manual minute-by-minute analyses of users by trained analysts, making them costly and time-consuming to implement.[1] (See the "Related Work in Facial Recognition" sidebar for more details on previous and ongoing research work.)

Using facial feature points automatically extracted from short video segments of participants' faces during laboratory experiments, our work applies a bottom-up approach to predict human performance. Our method maximizes data usage and allows us to predict both instantaneous errors (individual errors occurring at any time during the task) and task-level performance (speed, accuracy, and productivity over the entire task). This enhancement takes our method of human error prediction beyond the capabilities of existing HEI techniques.

Current Approach

To create our performance models, we first collected videos and performance logs of participants performing a laboratory task. We synchronized the videos with the performance logs and segmented the videos into meaningful chunks based on cues in the logs. An example of a meaningful chunk might be the time interval preceding one error instance.

We then extracted key facial feature points from the videos, such as the mouth and eye positions. We calculated time and frequency domain statistics over the facial feature 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

1541-1672/11/$26.00 © 2011 IEEE
Published by the IEEE Computer Society

Related Work in Facial Recognition

With its ability to create more than 10,000 expressions, the face has greater variability than any other channel of nonverbal expression. Thus, automated facial-feature tracking lets researchers tap into a rich resource of behavioral cues. In the past, researchers have shown great interest in using micro-momentary facial expressions to predict human emotion and mental states.[1–4] In a recent study, Rana El Kaliouby and Peter Robinson developed a general computational model to recognize six classes of complex emotions and implemented this model as a real-time system.[5] Their approach used dynamic Bayesian networks to recognize emotions. In a related study that used computer vision to detect emotional states, Jeremy N. Bailenson and his colleagues demonstrated that automated models could be developed to detect and categorize the felt emotion of individuals.[2] By training machine-learning algorithms that link certain facial movements to subjective perception of emotions, they were able to create real-time models that classified three emotions (sad, amused, and neutral) based on facial features and physiological responses. Rosalind Picard and her colleagues in the Affective Computing Research Group at Massachusetts Institute of Technology also demonstrated across a number of systems that tracking various facial aspects can give insight into the mental state of the person whose face is being tracked.[6,7]

Picard and colleagues then 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 user's nonverbal and affective cues and responds accurately to these cues will be most effective in human-computer interaction.[6] Some of the current work from the Affective Computing Research Group includes using multimodal sensory inputs for the following purposes:

- Predicting frustration in a learning environment;[8]
- Detecting and responding to a learner's affective and cognitive state;[9] and
- Assisting individuals diagnosed with autism spectrum disorders in social interaction.[10]

Following suit, researchers have now developed systems to model deception. Thomas O. Meservy and his colleagues extracted macro features such as head and hand position and angle from video cameras taken during an experiment where a mock theft took place in the lab.[11] Afterward, in an interview with a trained researcher, participants were either truthful or deceptive regarding the theft. Using machine-learning algorithms, the team was able to create models that obtained up to 71 percent correct classification of truthful or deceptive participants based on just the features extracted from the video recording of the subject.

Our work extends previous work in many ways. First, we predict a unique and multifaceted human behavior (human performance). Furthermore, we use a bottom-up approach that lets us link specific facial features directly to human performance and evaluate those features over varying time intervals. We can identify the most valuable pre-error intervals and the most informative interval lengths. We extend current HEI techniques in that we make predictions at two different temporal layers: instantaneous and task level in an intuitive and relatively painless method with highly accurate results. These advancements offer a cost-effective solution in terms of labor, time, and finances in human-performance prediction. We feel that such merits will yield significant benefits to both researchers and industry personnel who yearn to find answers within a face.

References

1. Z. Zeng et al., "A Survey of Affect Recognition Methods: Audio, Visual, and Spontaneous Expressions," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 31, no. 1, 2009, pp. 39–58.
2. J.N. Bailenson et al., "Real-Time Classification of Evoked Emotions Using Facial Feature Tracking and Physiological Responses," *Int'l J. Human Machine Studies*, vol. 66, no. 5, 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, pp. 144–169.
4. Y.S. Shin, "Recognizing Facial Expressions with PCA and ICA onto Dimension of the Emotion," *Structural, Syntactic, and Statistical Pattern Recognition*, Springer, 2006, pp. 916–922.
5. R. El Kaliouby and P. Robinson, "Mind Reading Machines: Automated Inference of Cognitive Mental States from Video," *Proc. IEEE Int'l Conf. Systems, Man and Cybernetics*, vol. 1, IEEE Press, 2004, pp. 682–688.
6. R. Picard, *Affective Computing*, 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'l J. Human-Computer Studies*, vol. 65, no. 4, 2007, pp. 361–375.
8. A. Kapoor, W. Burleson, and R. Picard, "Automatic Prediction of Frustration," *Int'l J. Human-Computer Studies*, vol. 65, no. 8, 2007, pp. 724–736.
9. S. D'Mello et al., "AutoTutor Detects and Responds to Learners Affective and Cognitive States," *Proc. Workshop Emotional and Cognitive Issues at Int'l Conf. Intelligent Tutoring Systems*, 2008, pp. 31–43; http://affect.media.mit.edu/pdfs/08.dmello-etal-autotutor.pdf.
10. M. Madsen et al., "Technology for Just-In-Time In-Situ Learning of Facial Affect for Persons Diagnosed with an Autism Spectrum Disorder," *Proc. 10th ACM Conf. Computers and Accessibility* (ASSETS), ACM Press, 2008, pp. 19–26.
11. T.O. Meservy et al., "Deception Detection through Automatic, Unobtrusive Analysis of Nonverbal Behavior," *IEEE Intelligent Systems*, vol. 20, no. 5, 2005, pp. 36–43.

entire task (task-level performance) and at any given instant within the task (instantaneous errors).

Figure 1 illustrates these steps.

Task Setup

Our experimental task consisted of fitting screws into holes for half an hour (see Figure 2). To perform the task, participants had to pick up a screw (item one in Figure 2a) from one of the virtual boxes using a Sensable Phantom Omni haptic pen (item four

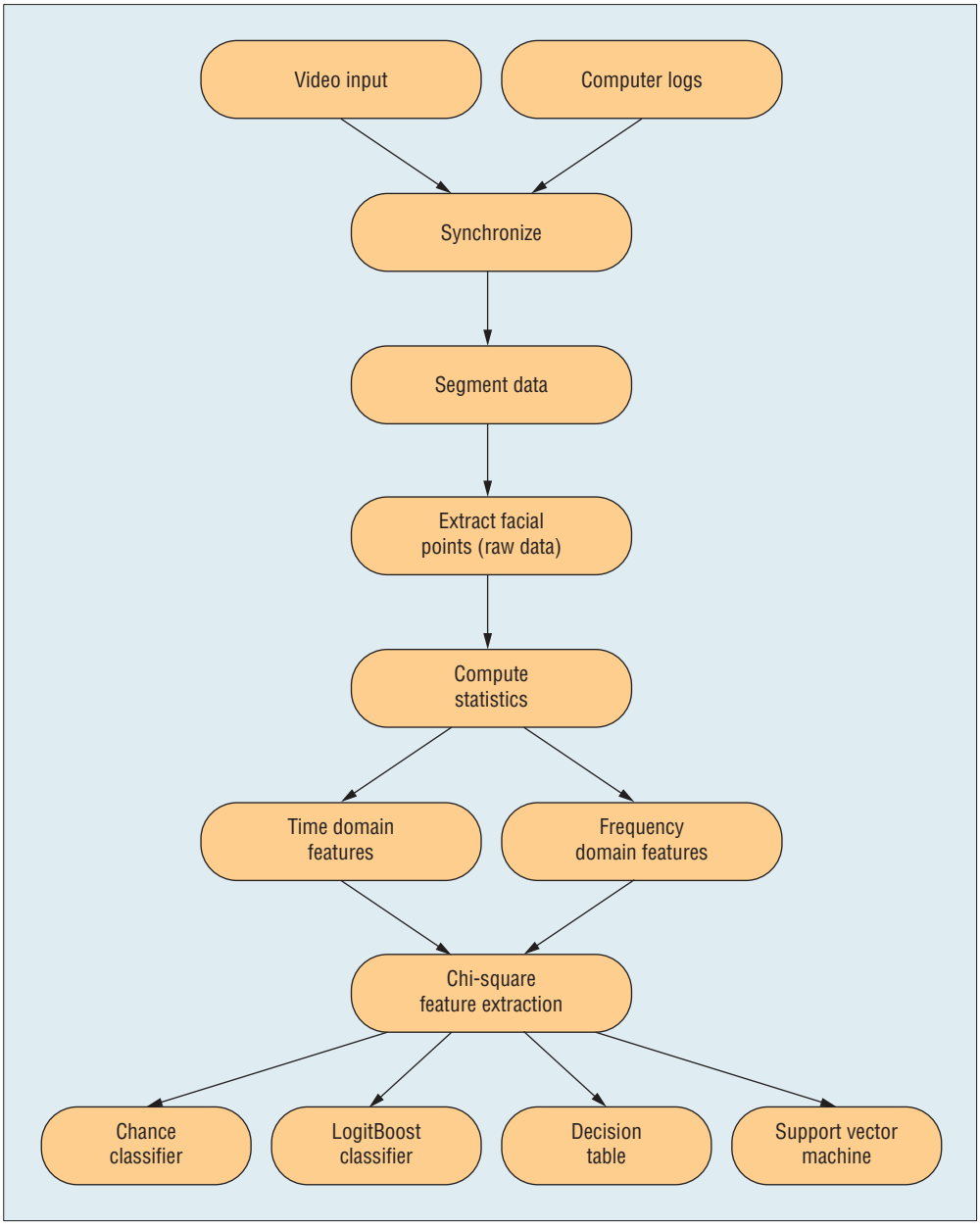

Figure 1. Human error identification approach. To create our performance models, we followed a series of steps to extract training data from a set of videos and performance logs.

in Figure 2b) and insert it into a hole with the correct label. The pen, a device with six degrees of freedom (x, y, z, pitch, yaw, and roll), let the user "feel" the hardness and the depth of the box. A beep indicated the user's success or failure to screw in the parts. The wooden boards refreshed after a preprogrammed amount of time regardless of the participant's progress. One board refresh was termed one phase of the experiment. The first board refreshed after 45 seconds. If the participant successfully filled two consecutive boards without any errors (indicating that the level of difficulty was too low), the phase time was curtailed by three seconds.

A high-resolution Logitech QuickCam UltraVision webcam (item 5 in Figure 2b) affixed to the top of the monitor captured the participants' faces at a rate of 15 frames per second while video recording software (Video Capturix 2007) compressed the data to AVI format. We also recorded performance logs, including measures such as time stamps for each error (that is, placing a screw in an incorrect hole, dropping a screw,

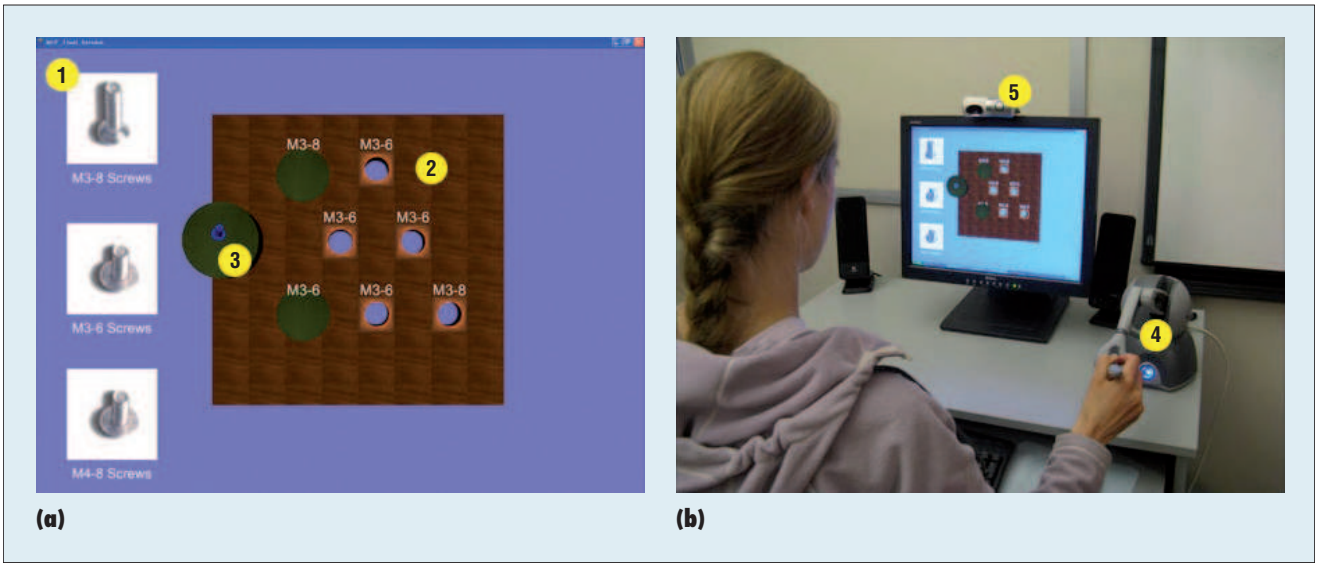

Figure 2. Experimental setup. The task was administered at a computer station with a flat-screen monitor adjusted to a 640 × 480 pixel resolution. (a) We presented the participants with three boxes, each containing a screw with a different part number. (b) The center of their screen contained a large wooden board with seven holes labeled with a randomly selected set of the different part numbers. The screen also contained a virtual screw box (1), virtual board (2), and virtual screw (3). The participants needed to put the screws in the holes using a haptic pen (4). We recorded each session using the mounted Web camera (5).

or failing to hold the screw in place until the beep), each correctly placed screw, each board refresh, and overall time spent holding screws. In this way, we measured each participant's instantaneous performance (errors) and performance (overall error rate and speed of completion) over the entire experiment.

We collected data from 57 students (25 female and 32 male) but discarded data from eight participants due to technical problems with collection.

Feature Computation

We extracted facial-feature points from the videos using the OKAO vision library (see Figure 3). To map the facial data with task performance, we then synchronized our facial data with the performance logs. Finally, we programmatically segmented the data according to our two prediction intervals: instantaneous and task. Instantaneous intervals corresponded to data from short time intervals directly preceding error instances, and task level intervals corresponded to data for entire phases of the task. We discarded any intervals with average face-tracking confidence (that is, how confident OKAO was in its measurement) lower than 60 percent.

We then calculated means, medians, minimums, maximums, standard deviations, ranges, and wavelet transformations on each of the raw facial-feature points in each interval (see Figure 4). We calculated these statistics because facial signals are dynamic, and their micro-momentary movements can leak information about the person's internal state.[2]

Feature Selection

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 using the freely distributed machine-learning software package Waikato Environment for Knowledge Analysis (WEKA).[3] We also performed a best-first search to determine the optimal cutoff for features to keep in our analyses. Similar methods have been successful in other classification problems, such as Arabic text classification[4] and gene-based cancer identification.[5]

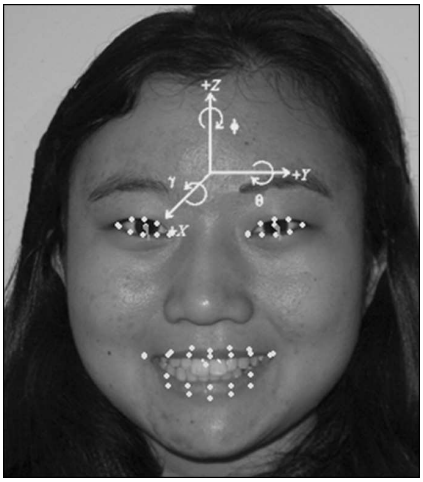

Figure 3. OKAO computer vision algorithm tracking points on a participant's face. We tracked 37 points on the face, along with head movements such as pitch, yaw, and roll and eye and mouth openness level. We tracked the points relative to the captured frame. However, for our calculations, we standardized all the points to be relative to the center of the face.

Figure 5 shows the performance curve of each analysis using different

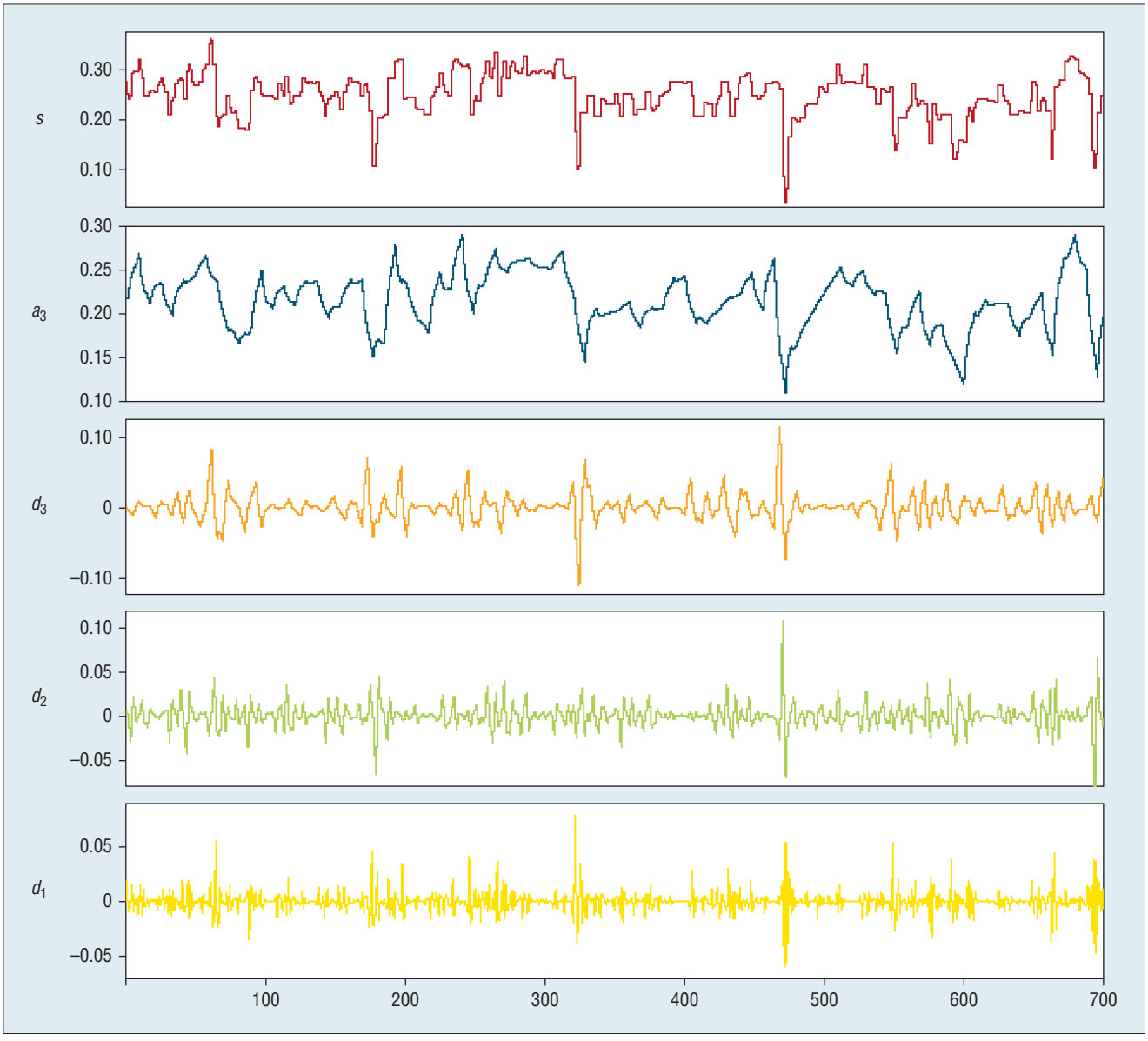

Figure 4. Wavelet decomposition of right eye openness level. The *s* is the original signal, a_3 is the order one decomposed signal, and d_1 through d_3 represent the three levels of decomposition: $s = a_3 + d_1 + d_2 + d_3$.

numbers of features. Table 1 lists the top 10 features for each analysis, and Figure 6 shows the meaning of each feature. On average, 60 percent of the top features were wavelet coefficients of face signals, which shows the power of the wavelet analysis.

Performance Prediction

From our experimentations with decision tables, support vector machines (SVMs), LogitBoost classifiers, and Bayesian nets, we found decision tables and LogitBoost classifiers to be the most powerful performance predictors. In many ways these algorithms mirror the cognitive process used by humans, whereby the sequence of behaviors (or features) that lead up to an error are learned by observing many examples. In machine learning, the examples are a set of vectors containing all the features and a label. From this training set, the classifier tunes its algorithms to predict unlabeled vectors in a *test set*.

Specifically, the decision table will, given a training set and another set of unlabeled instances I, search for the set $£$ of instances that match the features of I. It will then predict I to be of the majority class in $£$. If $£ = \varnothing$, then the majority class of the training set is predicted.[6] If the features are continuous they are divided into two discrete classes based on where they

fall in relation to the median value of the feature.

LogitBoost classifiers work by sequentially applying a classification algorithm to reweighted versions of a data set to make a series of classifiers.[7] A majority vote of these classifiers then determines the final class prediction of unlabeled instances.[7] 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.

Results

To gauge the performance of our classifiers, we calculated three main measures for each classifier: overall accuracy, precision, and recall. *Overall accuracy* is 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. *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.

We then looked at the overall accuracy, precision, and recall of our classifiers and compared them against the corresponding values from a chance classifier. We define a *chance classifier* as a classifier that guesses the class of an instance using the proportional split of the data.

Task-Level Performance Prediction

Our first goal was to predict participants' overall task performance by using only the first few minutes of facial data in their videos—that is, create models capable of prescreening individuals for aptitude at the task. We defined overall performance as a normalized sum of how many phases each participant completed, fastest phase completion time, total number of phases completed at maximum speed, mean error rate, amount of time spent holding screws, and mean number of filled holes per box. Participants with scores in the top quartile were labeled high performers, and those in the bottom quartile were labeled low performers. We used data from two independent data sets for testing and training. This assured the generalizability of our results across individuals.

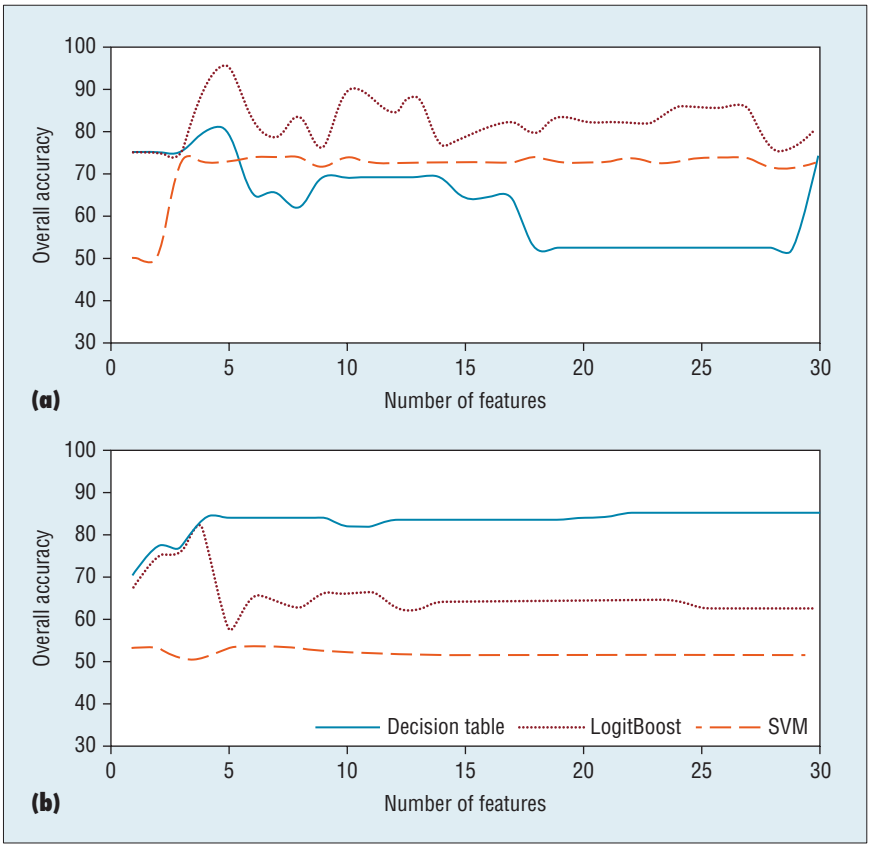

Figure 5. Accuracy of performance predictions. We used different numbers of features to determine the overall accuracy for (a) task performance and (b) instantaneous errors.

Figure 7 shows the face inputs we used to predict task-level performance using varying amounts of data. The results changed with differing numbers of phases, but in general accuracy increased with more phases. We noted significant drops in performance with certain numbers of phases (for example, 12 phases), which could be caused if certain phases were less predictive than others.

Although the overall accuracy peaked for both classifiers when 20 phases were used as input, the overall accuracy obtained with 10 phases was only 2 to 10 percent lower than the overall accuracy obtained with 20 phases. This demonstrates the power of our approach; using only 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 to classify participants as high or low performers.

Table 2 shows the face inputs we used to predict task-level performance using 20 data phases. Boosting significantly improved results; the LogitBoost performed almost 15 percent better than the simple decision table, classifying participant performance with 95.2 percent accuracy.

Table 1. Top 10 chi-square features for each prediction.

Level	Statistic	Feature	Definition	Chi value
Task	Average	Average *Y*	Vertical position of face	108.5
	Maximum	Average *Y*	Vertical position of face	78.90
	Wavelet	Gaze tilt	θ (radians)	70.50
	Minimum	Average *Y*	Vertical position of face	64.53
	Wavelet	Gaze tilt	θ (radians)	63.56
	Average	Lower lip center *Y*	*A* (*y* coordinate)	61.80
	Wavelet	Right eye open level	*h* (cm)	55.31
	Wavelet	Right eye ratio	w/h	55.25
	Wavelet	Gaze tilt	θ (radians)	54.26
	Wavelet	Right eye ratio	w/h	52.50
Instantaneous	Velocity	Roll	ψ (radians)	790.8
	Velocity	Yaw	β (radians)	449.7
	Velocity	Roll	ψ (radians)	424.6
	Wavelet	Left outer eye corner *Y*	*D* (*y* coordinate)	379.1
	–	Eye per close rate	Percent of time both eyes $h < .15$	370.9
	Wavelet	Average *X*	Horizontal position of face	370.6
	Wavelet	Left lower lip *Y*	*A* (*y* coordinate)	367.1
	Wavelet	Left upper lip *Y*	*B* (*y* coordinate)	356.0
	Wavelet	Left pupil *Y*	*E* (*y* coordinate)	354.3
	Wavelet	Left upper lip *X*	*B* (*x* coordinate)	353.2

This accuracy is more than 44 percent higher than the chance level. Recall was also notably strong for the high performers in both algorithms, indicating a low false-alarm rate. The support vector machine performed worse than the decision table and LogitBoost classifier, suggesting the support vector algorithm might not be well suited for performance prediction.

Instantaneous Error Prediction

In our second analysis, we predicted instantaneous errors. To do this, we compiled a data set of all the pre-error intervals, which we 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 this set an approximately equal number of randomly selected non-error intervals. We used two independent sets of participants for test and training sets and trained a LogitBoost classifier and a decision table classifier on each set.

Figure 8a shows the facial features predicting errors one second before they occurred with varying *I*, and Figure 8b shows the facial features predicting errors at varying *D* using two seconds of data. Performance peaked using two seconds of data one second before the error occurred. This suggests that the micro-momentary expressions indicative of errors might be short in duration (less than three seconds) and that the expressions occur at one second or less before an error. Table 3 shows results for the most successful classification.

System Use

Our results, most of which are in the 90th percentile, indicate that our method 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 to 90 percent of failures to human error, and other industries report similar rates—90 percent for airlines and 98 percent in medicine.[8,9] In these safety-critical applications, warnings could be issued to eliminate costly, or even deadly, incidents.

Moreover, given that our models provide both micro- and macro-views of human-task errors, our methods could also give managers a better understanding of worker performance at both the task and individual error levels.

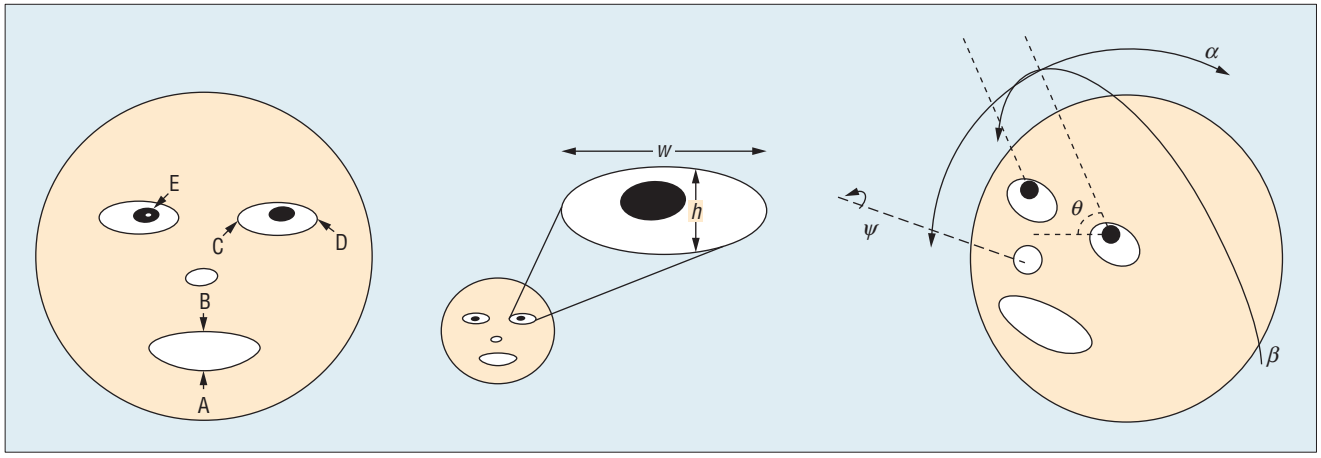

Figure 6. Significant facial features. Table 1 defines each feature.

This is an important contribution because simple aggregates of individual errors do not necessarily demonstrate a person's overall performance during a task. Companies could use these models to prescreen employees or match individuals to jobs better suited to their skills , saving time and resources.

Furthermore, our results indicate that we are able to outperform human analysts with only a fraction of the effort required in the traditional HEI systems. Thus, our models can either completely substitute human data analysts or supplement their work, saving valuable human resources.

Finally, because of the ease of adoption (our models require only a small webcam and processor), individuals and corporations alike could reap the benefits of performance prediction at little cost, allowing more widespread use and thus offering greater error avoidance potential.

System Limitations

Despite these encouraging results, the current study has several limitations. Although the models can be generalized across individuals, they cannot be generalized across tasks; our models were based on arbitrary definitions of error and performance specifically tailored to our laboratory experiment. Furthermore, we conducted our experiment in a laboratory setting with the face clearly visible. In work environments where the face is not visible our models would be unable to predict performance.

In addition, the facial-recognition software and machine-learning technology we used is not specifically tailored to the task of performance

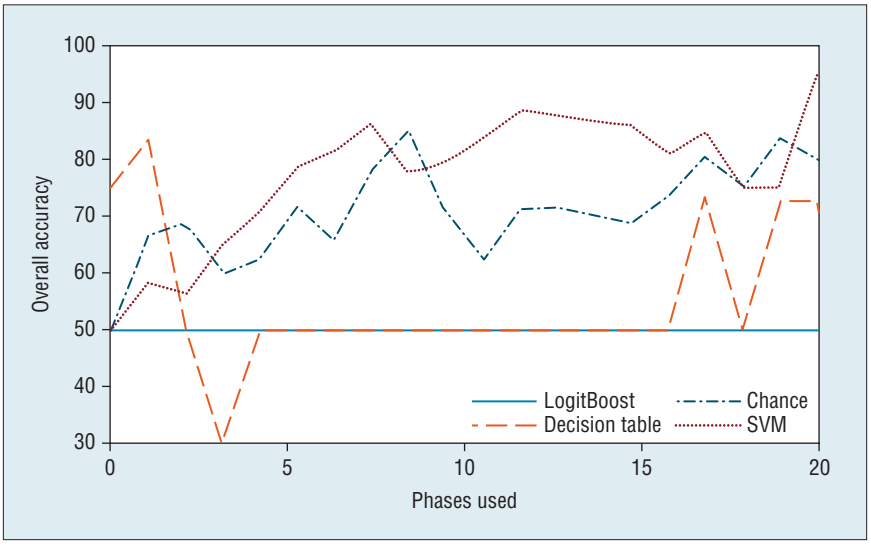

Figure 7. Task-level performance results. The results indicate that we can gauge participant performance over the entire task using only small amounts of facial data.

Table 2. Prediction results for task performance using 20 data phases.

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
Support vector machine	72.6	High	65.1	97.6
		Low	95.2	47.6

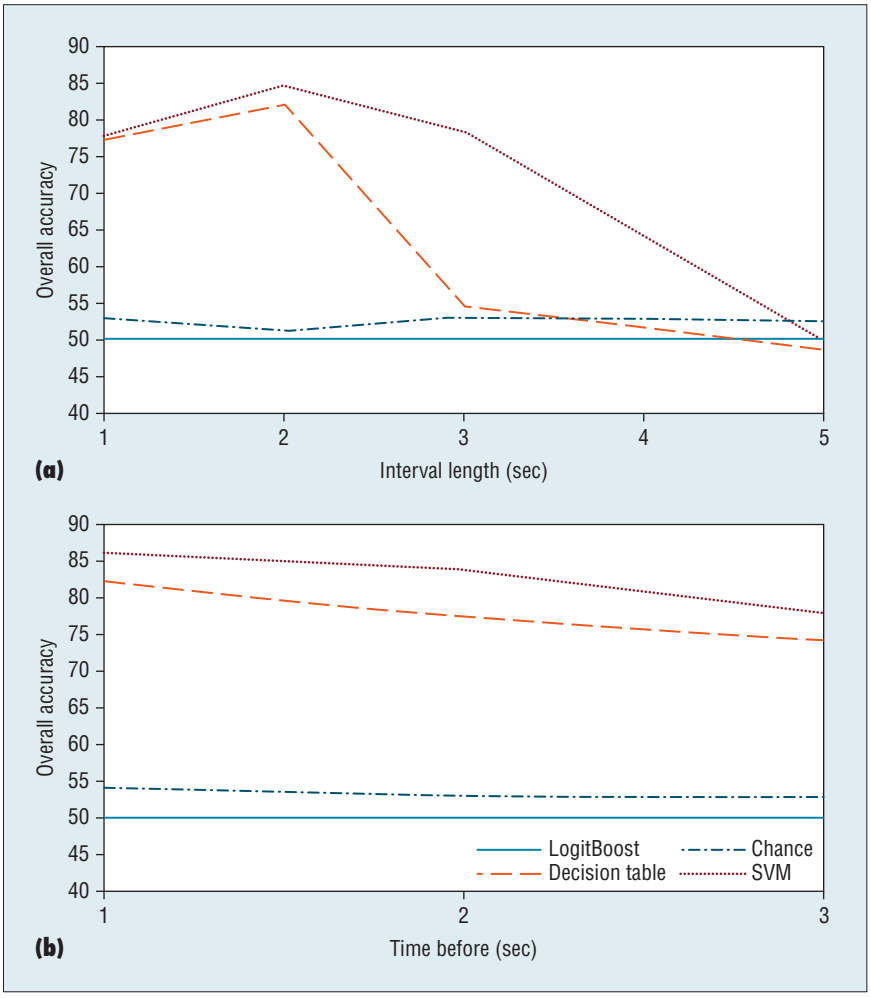

Figure 8. Instantaneous error prediction. We measured the facial features predicting errors (a) one second before they occurred with varying amounts of data (b) at varying times before the error occurred using two seconds of data.

Table 3. Face input predicting errors one second before they occur 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
Support vector machine	53.3	Error	0	0
		Correct	53.3	100

prediction, and the performance of our models is closely related to the quality of our software. OKAO vision tracks only 37 points on the face and might not yield accurate results for people wearing glasses. A custom-made library that automatically detects and tracks more points on the face and works around obstructions on the face such as glasses might yield better results. Similarly, alternative algorithms might also yield better accuracies.

Even with the technical challenges and limitations discussed here, we believe that our video-based performance prediction system demonstrates potential for many applications. Although it is difficult to expect any system, even human, 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, our models could serve as effective, cost-efficient, and unobtrusive monitors that assist both individuals and corporations in achieving maximum safety and output. ■

Acknowledgments

We thank Ritsuko Nishide, Shuichiro Tsukiji, Hiroshi Nakajima, and Kimihiko Iwamura for early guidance on the project, and OMRON Silicon Valley for partial funding of the project. We also thank Suejung Shin and Steven Duplinsky for their help in making the images, and Joris Janssen, Kathryn Segovia, Helen Harris, Michelle Del Rosario, and Solomon Messing for helpful comments on earlier drafts of this article.

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," *Human-Centered Computing: Cognitive, Social, and Ergonomic Aspects*, Lawrence Erlbaum Associates, vol. 3, 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*), Oxford Univ. Press, 1997.
3. H.I. Witten and E. Fank, *Data Mining: Practical Machine Learning Tools and Techniques*, 2nd ed., Morgan Kaufmann, 2005.
4. P. Kosla, *A Feature Selection Approach in Problems with a Great Number of Features*, Springer, 2008, pp. 394–401.
5. X. Jin et al., "Machine Learning Techniques and Chi-Square Feature Selection for Cancer Classification Using SAGE Gene Expression Profile," *Proc. Data Mining for Biomedical Applications*, LNBI 3916, Springer, 2006, pp. 106–115.
6. R. Kohavi, "The Power of Decision Tables," *Proc. 8th European Conf. Machine Learning*, Springer-Verlag, 1995, pp. 174–189.
7. J.H. Friedman, T. Hastie, and R. Tibshirani, "Additive Logistic Regression: A Statistical View of Boosting," *Annals of Statistics*, vol. 28, no. 2, 2000, pp. 337–407.
8. J.W. Senders and N. Moray, *Human Error: Cause, Prediction, and Reduction*, Lawrence Erlbaum Associates, 1991.
9. A. Isaac, "Human Error in European Air Traffic Management: The HERA project," *Reliability Eng. and System Safety*, vol. 75, no. 2, 2002, pp. 257–272.

THE AUTHORS

Maria E. Jabon is a software engineer at LinkedIn. While attending Stanford, she worked as lead systems engineer in the Virtual Human Interaction Lab and research assistant in the Nolan Lab. Her interests include machine learning and Web interfaces for the visualization of high-dimensional data sets. Jabon has an MS in electrical engineering from Stanford University. Contact her at mjabon1@gmail.com.

Sun Joo Ahn is a doctoral candidate at the Department of Communication at Stanford University. Her research interests include emotion and behavior prediction based on automated facial-feature tracking and consumer psychology within virtual environments. Ahn has an MA in communication from Stanford University. Contact her at sjahn@stanford.edu.

Jeremy N. Bailenson is the founding director of the Virtual Human Interaction Lab and an associate professor in the Department of Communication at Stanford University. His main interest is the phenomenon of digital human representation, especially in the context of immersive virtual reality. Bailenson has a PhD in cognitive psychology from Northwestern University. Contact him at bailenso@stanford.edu.