Using Automated Facial Expression Analysis for Emotion and Behavior Prediction

Sun Joo Ahn
Jeremy Bailenson
Jesse Fox
Maria Jabon

Abstract
We present an overview of facial expressions and emotions and measurement methodologies used in prior studies. Computer models that classify emotions and predict behaviors based on facial expressions can be an effective alternative to human coders. Most predictions can occur based on a small sample of behavior. Empirical studies using the proposed methodology are introduced to demonstrate the use of automated facial expression analysis as a new methodological tool for the investigation of facial expressions.
Introduction

The expression of emotion is achieved through intricate combinations of both verbal and nonverbal information produced from various sources of the body and the brain. Nonverbal information encompasses any message that is not expressed in words (Depaulo & Friedman, 1998) including gestures, postures, and vocal tones. Though people often do not invest much thought to the nonverbal aspect of communication due to its automaticity, nonverbal cues are the primary vehicles for expressing emotions and managing emotional experiences (Burgoon & Hoobler, 2002). Among the communicative channels that express emotions, the face is often considered to be the richest source of nonverbal information (e.g. Collier, 1985; Ekman, 2003).

In the past, much of the research on facial expression has been based on Ekman’s proposed method of measurement by dividing expressions into categories of predefined emotion labels (Ekman & Friesen, 1978). Not only is this process labor-intensive but it is also restrictive in the sense that the interpretation of facial expressions must be based on this predefined number of categories, regardless of the intensity, duration, and variety of feature movements. Compared to conventional methods, computers now enable researchers to process incredible volumes of data in a short amount of time in real-time (Colmenarez, Xiong, & Huang, 2004) but even automatic facial feature detection is largely dependent on conventional measurements upon analysis.

In this chapter, we propose a new model of approaching facial expression detection and analysis by incorporating automated technologies that can make and improve prediction models based on the raw feature movement data as input. Using a computer is advantageous because it is able to process incredible amounts of raw data to select facial features that are most relevant to improving the prediction model. Thus, more data can be utilized without having to fit it into predefined categories. Moreover, our suggested methodology is notably unobtrusive when compared to other behavioral measures such as physiological measures, which require sensors to be stuck to the body, or pupil tracking, which requires the head to stay stationary. Instead, computer vision uses just a small camera to execute facial tracking.

We first begin our discussion with a brief survey of facial expressions as a source of emotion signaling and conventional measurement methodologies used in prior studies of facial expression. We will then introduce the use of computers to classify emotions and predict behaviors of individuals based on raw data of facial feature movements. This prediction occurs via ‘thin-slicing,’ or a prediction based on a short sample of behavior (Ambady & Rosenthal, 1992). Finally, several empirical studies using the proposed methodology are introduced to demonstrate that automated facial expression analysis may be a valid new methodological tool.

Facial Expression of Emotion
People have long believed that the face is a window to the soul but the scientific community has yet to reach a consensus regarding the relationship between facial expressions and emotions (for review, see Bente, Krämer, & Eschenburg, 2008; Manstead & Fischer, 2002; Russell & Fernandez-Dols, 1997). The assumptions underlying our proposed methodology are that the automatically gauged facial expressions will be true reflections of internal intent, which will allow us to predict future behaviors. Thus, the debates most relevant to our proposal would be those regarding the automaticity (vs. conscious control) and universality (vs. specificity) of facial expressions across individuals. Automaticity would vouch for the veracity of the internal intents reflected through facial expressions, and universality would imply the generalizability our findings.

The bulk of the debate surrounding the emotion-expression link is based on dual-processing theories that give way to models with differing arguments for the interplay between automatic and controlled processing following stimulus (e.g., Gilbert, 1991; Trope, 1986). These models have contesting views on the degree of influence exercised by conscious control over the emotional response, determining the degree of expressed emotion. At the same time, the models agree that the initial emotional response following stimulus is automatic (Barrett, Ochsner, & Gross, 2007), implying that the production of nonverbal behaviors such as facial expressions may be “mindless” (Burgoon et al., 2000). Furthermore, as conscious controlling of emotion expression tends to be driven by social context (Jakobs, Manstead, & Fischer, 1996) or particular goals (Zaalberg, Manstead, & Fischer, 2004), it may be inferred that in situations devoid of particular social interactions or end-goals, the emotional responses following stimuli are generated automatically.

In relation, Ekman and Friesen (1969a) coined the term display rules, which are socially learned and often culturally distinct rules about the management of expression. Emotional responses are initially automatic and not controlled by conscious will (LeDoux, 1996), but it appears that social rules can regulate the appropriate manifestation of emotion expressions. Ekman (1972) provided empirical evidence of these display rules by demonstrating in a series of studies that upon viewing films of surgery and accidents, Japanese people tend to mask negative expressions with a smile when in the presence of another person. When left alone, however, the Japanese participants displayed the same facial expressions demonstrated by American participants in a separate group. Thus, in situations where individuals need not incorporate such conscious control of emotion expression, they will allow the initial automatic responses to govern, yielding veridical reflections of the internal state of their minds.

The issue of universality in facial expressions across culturally diverse individuals has also gained some persuasive evidence. Recently, Matsumoto and Willingham (2009) tested the universality of emotion expression via facial expressions by using blind participants. Their results indicated that people who were born blind, people who became blind later on in their lives, and normally sighted people all shared common expressions of happy smiles after winning an athletic event. Further evidence is provided by neuroscience such as Panksepp’s (1998) neurobiological model (see also Panksepp et al., 2000). Similar to Ekman’s work (1972) on universality of the expression for a set of ‘basic’ or ‘core’ emotions (e.g., anger, fear, joy, sadness, playfulness) across cultures, Panksepp also argues
for basic emotion systems which are hardwired at birth by using positron emission
tomography (PET) and functional magnetic resonance imaging (fMRI) technologies.
Although all emotion cannot be categorized into discreet, predetermined categories, these
researchers maintain that the emotional range is created from subtle variations of the basic
set of emotions (e.g., Ekman, 1992).

Thus, according to the theories of automatic and universal expressions of emotion,
detecting and assessing facial expressions will yield insights to internal states of mind and
these results may be generalized across the population. Therefore, the discussion of various
methodologies to scientifically measure and analyze facial movements is important in our
attempt to capture the expressions of the face. Wagner (1997) divides these methods into
two groups: measurement studies, which answer questions regarding facial behavior by the
objective measurement of facial movements, and judgment studies, which answer research
questions by gauging responses from observers. This chapter focuses on the former, as it
applies to the prediction of emotion and behavior. One of the two most widely used
methods in measurement studies of the face is electromyography (EMG), which measures
the contraction of the muscles underlying facial skin (Cacioppo, Tassinary, & Fridlund,
1990). However, the EMG tends to be intrusive due to the use of electrodes, the equipment
is costly, and the output data can be unreliable (Wagner, 1997).

The other form of measurement uses objective coding schemes based on visible
units of facial behavior. This is usually done by post-processing video recorded during an
experiment, using slow-motion replay for a frame-by-frame analysis of facial movements
by human coders who systematically follow descriptive rules of judgment. Systematic
measurements, such as the Facial Action Coding System (FACS) created by Ekman and
Friesen (1978), the Maximally Discriminative Facial Movement Coding System (MAX:
Izard, 1979), and emotional FACS (EMFACS: Ekman & Friesen, 1982) have been
developed to objectively study the relationship between facial expressions and emotions.
However, use of these methodologies requires extensive prior training of the coders and a
frame-by-frame analysis of videotaped participants, which is labor intensive and raises
speculations regarding inter-observer reliability (Sayette et al., 2001). Furthermore, this
process often leaves room for measurement variance due to human error, as there is no way
to determine the precise onset, progression, and termination of each expression.

A factor that makes the measurement of facial expression difficult is that
manifestation of emotion is fleeting and transitory, typically lasting from half a second to
four seconds (Ekman & Friesen, 1969a, 1969b; Ekman, 1984; Izard, 1997). Despite
conscious efforts to conceal true emotions following display rules, these involuntary
microexpressions tend to slip through and ‘leak’ veridical internal states during extended
observation (Ekman, 1997). For instance, in deception, a person may be able to consciously
maintain an expressionless face (i.e., “poker face”) but may unconsciously shake his or her
leg at the same time, leaking signs of nervousness.

In order to catch these veridical leakages of emotion on the face, extensive
observation may be required, in which case using a computer to replace the human coder
would yield optimal outcome. Frank (1997) comments that automatic facial tracking “has
the potential to save countless hours in scoring the dynamics of facial actions, thus it may make viable the discovery and exploratory work that is needed on dynamic markers of emotion in the face” (p. 240). As such, computers that are able to collect, classify, and interpret various forms of nonverbal cues such as facial expressions, voice, or gestures are garnering attention from scientists as a more effective means of investigating emotion expressions compared to the traditional human-coder.

For instance, Picard (1997) also emphasizes that technology which recognizes nonverbal and affective cues of the user and responds accurately to them will be most effective in aiding individuals in their learning processes. Her recent work includes using sensory inputs from multiple sources (e.g., facial expressions, head movement, and posture) to predict frustration in learning environments (Kapoor, Burleson, & Picard, 2007); monitoring dialogue, posture, and facial features to detect and respond to a learner’s emotions and cognitive states (D’Mello et al., 2008); and developing portable aids which track, capture, and interpret facial and head movements of other people to assist individuals diagnosed with autism spectrum disorders in social interaction (Madsen et al., 2008).

Advantages of Automatic Facial Expression Analysis

Computer systems have opened up new horizons for emotion detection by recognizing and detecting nonverbal cues via automated devices. Although these systems are still far from achieving the capacity of human perception, they are able to classify and assess user emotions through predetermined mathematical models with limited human intervention (for a review, see Konijn & Van Vugt, 2008). Among the various modes of nonverbal communication, we focus on facial expressions which are captured by small cameras and later analyzed with computer software. Before introducing our proposed method in detail, the following is a brief overview of the advantages of using technology to replace more traditional forms of measurement discussed earlier.

First of all, automatic facial expression analysis performs with a higher accuracy than human coders who have larger margins of error and may overlook important information. As Osgood (1953) noted, “From the total splurge of sounds made by an actively vocal infant, only the small sample that happens to strike the observer is recorded at all” (p. 684). In contrast, the computer is able to detect micro-expressions, even those that last for only a few seconds at a time that a human coder may miss. In addition, the computer does not fatigue and shows relatively little inconsistency in performance. As Webb et al. (2000) point out, “People are low-fidelity observational instruments... recording and interpretation may be erratic over time, as the observer learns and responds to the research phenomena he [or she] observes” (pp. 143-144).

Recent studies applied automated feature extraction and classification to extract macro features such as the head and hand position and angle from video cameras (but not changes in facial features) taken during an experiment where a mock theft took place (Meservy et al., 2005). Computer models obtained up to 71% correct classification of innocent or guilty participants based on the macro features extracted from the video camera. Furthermore, in an overview of the deception detection research, Meservy et al. (2008)
noted that the accuracy of humans coding behavioral indicators only falls around 50%, but by using computers to automatically extract and identify relevant behavioral cues, deception can be detected at a significantly accurate level without the invasiveness of other methods (e.g., physiological measures such as polygraph machines, or lie detectors) and the cost of extensively trained human interviewers.

Yet another advantage of using automated facial detection technology coupled with computational models is that once the system is able to secure the parameters for a model, prediction of behavior (vs. simple detection and classification) can be made using only a small sample. This is a computational rendering of what social psychologists call ‘thin-slicing,’ a way people sample a short excerpt from social behavior to draw inferences about states, traits, and other personally relevant characteristics (Ambady & Rosenthal, 1992). Prior research has demonstrated that initial moments of behavior have high correlations to the decisions made and actions taken in the future (Carrere & Gottman, 1999; Gottman & Levenson, 1992; Gottman & Notarius, 2000). For instance, based on an observation of a three minute video clip of a conflict between a married couple, Carrere and Gottman (1999) were able to predict the outcome of that marriage after six years. Similarly, we are able to mimic this human cognitive process using computers and develop a model that can predict behavior based on a short sample of facial feature measurements.

Using this approach, Curhan and Pentland (2007) demonstrated the use of computer coding to show that thin-slices of behavior in the first few minutes can predict the outcomes of a negotiation task. The methodology used in their study is comparable to ours in that they used data collected via computers and thin-slicing to predict behavior. In their study, computers were used in place of human coders to detect vocal behaviors (i.e., time spent speaking, influence over conversation partner, variation in pitch and volume, and behavior mirroring) during a negotiation task. Their results imply that the speech features extracted during the first five minutes of negotiation are highly predictive of future outcomes. The researchers also note that using computers to code speech features offers advantages such as high test-retest reliability and real-time feedback.

As a cost-effective and relatively accurate method to detect, track, and create models for behavior classification and prediction, automatic facial expression analysis has the potential to be applied to multiple disciplines. Capturing behavioral data from participants such as facial expressions or head movements may be a more accurate representation of how and what they feel, and a better alternative to self-report questionnaires that interrupt participants’ affective-cognitive processes and are subject to bias (Reynolds & Picard, 2005; Picard & Daily, 2005).

In the studies detailed in the following pages, we present some empirical applications of automatic facial detection coupled with learning algorithms that predict future behavior based solely on facial expressions. We begin with an introduction of the common methodology used in all of our empirical work and describe how we are able to derive behavior predictions from facial features.

Overview of Proposed Methodology
Facial expression recognition is a method of classifying facial motion and facial feature deformation into abstract classes based on visual information alone (Fasel & Luettin, 2003). It is a bottom-up approach in which the correlation between facial movement and a particular output (e.g., behavior, attitude, or emotion) becomes the formula for establishing not just a classification model but a prediction model as well. As all of the raw data measuring the precise movement of facial features are processed to create these models, they reflect intricate combinations of feature movements and movement patterns which would be lost on the human coder using conventional measurement schemes.

Computer vision and machine learning form the cornerstones of our approach to modeling and predicting human behavior. Machine learning is a technique in which computers are programmed to modify and improve their performance based on novel data input (Nilsson, 2005). In this way the machines can mimic the process of learning, autonomously adapting to external change in the environment (i.e., new input factors) and reducing the need for constant redesign.

Many of the problems that can be approached with machine learning are linearly separable, which means that there is a clear distinction between one group and another in a given space (e.g., cats on the left side, dogs on the right side). In this case, a model can be created to perform classifications in which parameters are established and modified until an optimal standard for categorizing different objects into a certain space is set. That is, through repeated trial and error, the model improves the method of calculating and estimating how to correctly classify individual instances of data until it reaches its optimum performance level. This is considered to be the training process through which these machine algorithms learn.

Although the computations behind this process are extremely complex, in greatly simplified terms, machine learning used in linearly separable classifications can be likened to a linear regression. There is an input layer, which is similar to independent variables, and an output layer, which is analogous to dependent variables. By detecting correlations between the input and output layers, the computer ‘learns’ how to derive the output with the given input. Once the computer goes through numerous sequences of input and output information, it is able to secure the parameters for a model. Then, as in regression, this model can be used to predict a novel output with novel input data and thus successfully complete classification tasks.

However, not all problems in the real world are linearly distinguishable. The strength of certain machine learning methods, such as the Support Vector Machine (SVM), is that the machine classifier is effectively able to handle nonlinear problems (Elizondo, 2006). Although this requires complex computations that map nonlinearly separable data into numerous independent linear dimensions until the data becomes linearly separable, using computers to comb through large amounts of data and to detect linear patterns from what seems to be a complex, nonlinear spread of data offers powerful alternatives to conventional statistical methods in finding causal relationships in social science research.
To create our models we used facial feature data extracted from recorded video clips as the input layer, and participant behavior gauged through behavioral measures or surveys as the output layer. Once the computer secured the parameters of its computational model, we were able to predict the behavior of a new person based solely on his or her new face feature data. Because we are looking at 22 or 37 points on the face rather than at a single feature to derive correlations between facial features and behaviors, our input space is often complex and non-linear. Using machine learning algorithms provides us with the analytic power to search for patterns on multiple dimensions.

Although previous work has used machine learning to detect and classify the movement of facial features using machine learning, our approach extends these efforts by focusing on the predictive function of these models. That is, although prior models stop at either detecting facial movements or classifying them into pre-designated categories, our model goes beyond to predict the future behavior within a given task (e.g., a virtual car accident or an error in performance). This opens up the possibility of such models becoming a common methodology in social scientific and behavioral research.

Figure 1 summarizes the typical phases of data analysis used in our experiments.

Figure 1. Data analysis procedure

A. Facial Feature Extraction

The first step in the construction of a typical dataset includes post-processing the videos of facial movements recorded during experimental tasks to extract key facial features and head movements. In our studies, a computer vision library such as the Neven Vision library or the OKAO vision library was used. The Neven Vision library automatically detects, without any preset markers worn on the face, the x and y coordinates of 22 points along with eyes and mouth openness levels and head movements such as yaw, pitch, and roll (the measurement of object rotation for X, Y, and Z axes, respectively) for each frame of the video. The OKAO vision library, developed by OMRON Corporation, automatically detects and tracks 37 points on the face, head movements such as pitch, yaw, and roll, as well as eye and mouth openness levels. In Figure 2, we present a screenshot of OKAO face tracking software.
B. Data Synchronization and Time Series Statistics Calculation

In the next phase of analysis, video recordings are synchronized with data collected from experimental tasks such as surveys or simple motor tasks. This is done to map the extracted facial geometry information to behavioral output data. In the experiments described in this chapter, three to five second intervals of facial expressions were taken one to two seconds before each instance of the behavior to be predicted and used as the input data. After data synchronization we also computed a series of time-domain statistics on the coordinates in each interval to use as inputs in our classifiers. Statistics calculated included averages, velocity, maximum value, minimum value, standard deviation, and range. All of these values were computed for each of the NEVEN or OKAO outputs. For some important facial characteristics, such as the eye and mouth openness levels, we created five-bin histograms from 0% to 100% to capture the distribution of eye and mouth state over the interval. We also calculated the percentage of the time the eyes were closed (PERCLOS) during each interval.

C. Chi-Square Feature Extraction

In our experiments, we discovered that classifiers using a reduced feature set tended to perform better because training the model with an excessive number of features led to over-fitting. One reason for over-fitting was because there was much overlap among our features. For instance, the left eye x-coordinate usually correlated well with the right eye x-coordinate. Including this redundant information resulted in reduced accuracy in our models. Thus, in order to reduce our datasets to only the most relevant features for each analysis, we performed a chi-square analysis for each of our datasets. A chi-square analysis evaluates the worth of an attribute by computing the value of the chi-squared statistic with respect to the class. By performing this analysis and ranking the obtained scores we were
able to determine the most informative features for a given dataset and select only those above a certain threshold to use in our final classifications. This not only alleviated the problem of over-fitting, but also reduced computation costs and identified which features were most important indicators of the behavior we were aiming to predict, speeding up the process of training.

D. Dataset creation

To prepare our data for classification, we split each dataset into two sets - an independent test and training set - by selecting independent test instances from only a certain group of participants and training instances from a different, non-overlapping set of participants. This was done so that we could secure the parameters of our model from one group of participants and be sure that none of those participants appeared in the dataset we tested our models on, which would result in falsely high reported accuracies.

E. Evaluation of Classifier Performance

The technique of machine learning classifiers discussed above was used to analyze the dataset created. To select the classifier which would best perform this task, we tested out some state of the art machine learning classifiers, including Support Vector Machines, Logitboost with a decision stump weak classifiers, Multilayer Perceptron Neural Networks, and Bayes Net Classifiers (Freund & Schapire, 1996; Friedman, Hastie, & Tibshirani, 2000). We built all classifiers using the freely distributed machine learning software package Waikato Environment for Knowledge Analysis software package.

To gauge the performance of our classifiers we used two main measures: precision and overall accuracy. Overall accuracy is defined as the total number of correctly classified instances. Precision is defined as the number of correctly classified instances in a given class divided by the total number of instances for that class. One can think of this as the probability that the system will correctly identify instances of a given class. Note that overall accuracy is a gross measure of performance; precision is a measure of the accuracy within a class individually. Thus in looking at both overall accuracy and precision we are able to obtain a comprehensive measure of the performance of our classifiers.

When viewing results, however, one must also consider the chance overall accuracy and the chance precision for each class. 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 is defined as the percentage of instances such a naïve classifier would correctly classify and the chance precision for each class is the accuracy it would achieve in each class individually.
In this section, we introduce some of our empirical studies that used computer vision coupled with machine learning algorithms to detect and classify facial features as well as predict future behavior based on these facial features. Compared to facial expression studies that use human coders, all of our studies were able to process more data faster with less labor, achieve high test-retest accuracy across all participants, and consequently detect patterns in facial movements that predict future behaviors that may have been lost to the human eye. Along with research benefits for social scientists, implications for real life applications will also be discussed for each.

A. Real-time Classification of Evoked Emotions

This study demonstrated the application of automatic facial recognition and machine learning to analyze extensive amounts of facial data and to detect patterns that correlate facial feature movement to discrete emotions. The processes and results of this study will help readers understand how automated facial tracking and prediction based on computer vision and machine learning algorithm works.

The input data consisted of videotapes of 41 participants watching films to elicit emotions of either amusement or sadness, along with measures of their cardiovascular activity, somatic activity, and electrodermal responding. It should be noted that the recorded expressions were spontaneously elicited expressions, unlike the photographs of deliberately posed faces often used in prior facial expression research (e.g., Ekman & Friesen, 1971). Therefore, the probability that we accessed veridical emotions is higher than in studies that used deliberately posed faces (see Nass & Brave, 2005, for further discussion of this distinction).

The output criteria was the intensity of the two emotions rated by trained coders looking at a second-by-second assessment of sadness or amusement of the recorded facial expressions using coding software informed by Ekman and Friesen’s FACS. We correlated these input layers to the output layer (i.e., emotion) by building algorithms that used extracted points from the participants’ faces as well as their physiological responses to predict the level of amusement or sadness for every person at every second.

The prediction of the intensity of emotion is a much more difficult task than the simple binary detection of emotions because the former is a linear classification problem, but our model yielded encouraging results. In predicting the emotional intensity of amusement, the model performed with a correlation coefficient of 0.53 ($MSE = 1.44$) based on just facial features. For predicting the intensity of sadness, the linear regression function performed with a correlation coefficient of 0.23 ($MSE = .79$). This suggests that machine learning algorithms may be successfully used to predict emotional intensity of amusement and sadness based solely on facial expression.

We also created binary classifiers to classify detected discrete emotions (amusement, sadness, and neutral). These models were even more successful than the emotional intensity models, demonstrating 65% precision in classifying amusement, 35% accuracy in classifying sadness, and an average of 85% in classifying neutral expressions.
using only facial feature data. Furthermore, the precision rates increased when physiological data were included with facial expressions as input layers.

The results imply that computer vision coupled with machine learning algorithms may be used to detect, classify, and predict emotion and emotion intensity with relatively high accuracy. Along with important academic implications, the possibilities of industry applications are infinite. Examples of interested parties may include advertisers wanting to know the emotions of consumers, participants on dating sites interested in the emotion felt by their partners, and television producers interested in the emotions of their viewers.

B. Predicting Unsafe Driving Behavior

Driving is an integral activity in the lives of many and emotions may well play a large role in driver safety (Lerner, Singer, & Huey, 2008; Ranney, 2008). There have been active research efforts to develop vehicle collision avoidance systems (Bertozzi, Broggi, & Lasagni, 2005; Kuribata, Takahashi, & Ide, 2005; Suzuki, Fujii, & Tanimoto, 2005; Williamson & Chamberlain, 2005). These studies demonstrated that systems inside vehicles can use information collected about the driver, such as emotional states or distraction levels, to predict unsafe driving behavior. We extend the previous work by attempting to use facial expressions in not only detecting mental and emotional states but also predicting driving accidents two full seconds before they occur.

For our study, participants were asked to drive through a 40 minute course in a driving simulator. The driving simulator was set up to run on a single PC and project the simulated image of the roadway onto a white wall in the lab. The course simulated driving in a suburban environment with conditions varying from light to intense traffic.

![Figure 3. Participant being monitored in STISIM driving simulator](image)

Based just on facial features, our system was able to predict minor and major automobile accidents with an average accuracy of 78.8%, which is 35% better than chance. We also found that including supplementary information from the car, such as brake and accelerator maneuvers, improved the accuracy by 10 percentage points or more.
The results indicate that it would be relatively easy to create and apply a driver safety system using a small camera and a small computer to detect and analyze driver emotions and behaviors through facial expressions in real-time. Prior studies have shown that drivers who are more likely to become angry (i.e., those with trait anger) tend to engage in more aggressive behavior on the road, which can result in negative outcomes such as crashes (Deffenbacher et al., 2003). Our model of accident prediction in this study and the model to predict emotional state and intensity discussed earlier could be coupled and applied to predict such emotional states and accidents before they occur.

C. Monitoring Operator Fatigue

The possibility that automatic facial expression recognition can be used for modeling behavior yields enormous potentials for application. Another area that we focused on is performance prediction, in particular, the learning process of people tackling new and unfamiliar motor tasks. Many activities in everyday life require learning and performing some sort of motor skill, as do many jobs at workplaces. The ability to gauge the process of learning and predict performance would lead to solutions that could optimize the productivity of both the individual and the organization.

This study proposed that by using computer vision coupled with machine learning algorithms, we can predict errors before they occur, correctly identify whether a participant is learning or has already mastered a given task based on facial expressions, and also predict overall participant performance quality based only on facial features. Moreover, we used thin-slicing to base our prediction of overall performance quality on facial data from the first seven minutes of participation.

The experimental task in this study was a simulation of a factory line in which participants fit screws into designated holes using a haptic device, an input-equipment which allows users to feel, touch, and manipulate objects in virtual space). On the left-hand side of the monitor, participants saw three boxes, each containing a screw with a different part number. In the center of the screen, there was a wooden box with seven holes labeled with part numbers. The wooden boards were programmed to refresh to a new board after a pre-programmed amount of time. Every time the participant successfully filled out two consecutive boards without any errors (indicating that the level of difficulty was too low), the given time was curtailed by three seconds. This ensured that the participants were working at a rate of difficulty that continuously adjusted to their level of learning.
Using the collected facial expression data from the web-camera recordings and the performance data, datasets were constructed to build our computational models.

For this study, we built five separate models. The first used facial data collected two seconds before an error occurred to predict impending error. The second classified whether or not a participant was in a ‘learning’ phase or ‘mastery’ phase of the experiment. *Learning* was defined as sessions in which the given time continuously decreased. Decreasing completion time indicated that participants were adjusting to the task and that they were gaining speed and accuracy. *Mastery* was defined as sessions in which the given time had stopped decreasing and had leveled off. This indicated that a participant reached his or her maximum performance capacity and was unable to complete the task any faster. The final three classifiers predicted the overall quality of participant performance. In making the participant classifiers we used only the data from the first seven minutes of facial features. In this way we simulated thin-slicing.

Results indicated that the overall accuracy rate for error prediction two seconds before the event was 88.7%, compared to the chance predictor performance of 85.4%. Overall accuracy of classifying learning and mastery phases were 78.8% and 57.0% respectively, which were over 15% better than the chance predictors in both categories. Finally, the models built on data from the first 10 phases of the experimental task were able to predict the overall performance quality of an individual. The overall accuracy of these predictions ranged from 70-98%, which is 20-40% above those of chance predictors.

Thus, we have evidence that computer vision coupled with machine learning algorithms is able to mimic thin-slicing. This may also be indicative of the predictive power of patterns within facial movements. That is, the patterns of facial expressions detected within the first few minutes of the task are actually indicative of overall performance outcomes at the end of the task. An example of a real-life implementation of
this finding may be a system with which the facial expressions of workers are monitored to assess fatigue levels and establish the most effective work-rest schedules.

D. Online Shopping

This study explored the possibility of predicting whether or not a consumer on an online shopping site will purchase a certain product by reading his or her facial expressions moments prior to the purchase. Online behavior involves countless dispositional and environmental variables. By predicting consumer behavior based on facial expression alone, our system would not only shed insight into online consumer decision-making process but also yield important implications for industry in developing advertisements or sales aids that react to the facial expressions of consumers. This could also lead to the establishment of an interactive shopping experience.

Until recently, clickstream has been the most widely used measure of online consumer behavior. Clickstream is a tracking technique that denotes the path a visitor takes through websites (Bucklin et al., 2002). Two aspects of browsing behavior are tracked: the visitor’s decision to continue browsing or exit the site, and the length of time spent in each page. However, Fox and Spencer (2006) point out that although clickstream data can point out the ‘when’ and ‘what’ of web visits, it fails to answer questions about the ‘how’ and ‘why’ of consumer site use. In this light, supplementing the facial data with click-through data will provide greater accuracy in predicting online consumer behavior.

The first part of the experimental task was to answer two rating questions regarding the 24 products selected from the pretest. The first question measured how much a participant liked the product and the second measured how likely the participant was to purchase it. The second part of the task showed a complete list of the 24 products rated by the participant in thumbnail pictures and required the participant to choose the final set of five products that he/she would like to purchase. This was to relate the purchase intention measured in the first task to actual purchase behavior measured by this second task.

The dataset construction first required the extraction of facial features from the web camera videos taken during the study and synchronizing that data with the user responses to the survey. Statistics were then computed on the facial points.

Overall, classifiers demonstrated better accuracies when predicting buy behavior than when predicting like behavior. This implies that consumers incorporate different information processing tactics when they evaluate products versus when they make purchasing decisions. In relation, previous work by Sood and Forehand (2005) demonstrated that that memory tracing for choice tasks is higher than for judgment tasks. Choice tasks can be paralleled to the buy ratings in our study and judgment tasks to our like ratings. The authors agree that choice tasks encourage consumers to rely on heuristics which implies that the buy rating may be based more on heuristics than the like ratings. Because heuristic decision-making takes into account peripheral cues such as emotions, and an increase in emotional stimuli leads to an increase in facial expressions, participants
are likely to show richer facial expressions for buy ratings than for like ratings, thus increasing the probability of better predictions.

Products that invoked amusement in people also showed markedly higher prediction rates than for non-humorous items. This could be accounted for by the fact that humor triggers an orienting response where individuals display a certain level of arousal and increased attention to process the novel stimulus. This could lead to a moment of leakage in facial expressions where the individual temporarily loses cognitive control of the face and true expressions leak to the surface, thus allowing the face to display genuine thoughts or feelings.

Another result is that high involvement products had higher predictive power than the medium or the low involvement products. Intuitively, high involvement products encourage consumers to become personally engaged in the decision-making process (Greenwald & Leavitt, 1984; Zaichkowsky, 1985) and such personal connections are bound to bring in heuristic cues heavily loaded with emotions and personal memories. High involvement products may then yield richer facial expressions, which would explain the higher predictive power of these models. In particular, we obtained better results with buy ratings (which also incorporate heuristic choices) than like ratings in the high involvement product category as well.

Although this study did not focus on detecting the emotions of consumers as they shopped online, it is highly likely that some level of affect was involved in producing facial expressions, especially for high involvement and humorous products. One of the strengths of our approach is the system’s ability to predict behavior without having to define emotions or to decide their relationships with facial expressions. The model is able to objectify facial movement regardless of the specificities of the felt emotion and establish a real-time system that can predict the purchasing behavior of consumers.

Discussion and Implications

The face is an important nonverbal channel to signal emotions and its expressions hold important implications in communication and social interactions. Although more than 10,000 expressions can be made using just the facial muscles (Ekman, 2003), scholars are still actively debating about a handful of basic emotions expressed through spontaneous facial movements (e.g., Fridlund, 1994; Izard, 1990; Matsumoto, 1987; Russell, 1997).

The difficulty of the topic of facial expression and emotions lies in the fact that, as noted by scholars such as Ruch (1997) and Zajonc (1998), emotions tend to be a slippery, hypothetical concept that no one can define and observe directly, but can only be inferred from several indicators such as behavior, physiological changes, or self-reports. Using a systematical measurement tool such as FACS helps researchers by breaking down facial expressions into distinctively measurable units, providing a tangible approach to investigating the causes and effects of facial expressions. However, FACS is labor-intensive, susceptible to problems of inter-coder reliability, and limits the usage of data.
Computer vision coupled with machine learning offers a cost effective and reliable alternative to human coding of emotion. Computers are able to offer the absolute measurement of facial feature movement with precision that would be difficult to be matched by human coders and provides accurate data on the duration and intensity of the movement. With the incredible rate of technical development, the machine’s ability to detect and analyze facial expressions is likely to increase astronomically in the future. Just as computer programs now execute most statistical calculations, it may be not long enough before we rely heavily on the computer power to plough through the massive amounts of data to figure out the relationship between facial expression, emotion, and behavior.

Our empirical studies demonstrate that coupling computer vision and machine learning algorithms can take facial expression research beyond the limited capacities of prior methodologies and help researchers obtain a firmer grip on this slippery topic. Based only on the raw data of facial feature movements, our models were successfully able to classify different emotions; predict whether a consumer will purchase a product or not; predict an automobile accident before it happens; classify whether a person is learning or has mastered a task; predict an error for a given task before it occurs; and predict whether a person will excel at a task, learn quickly, or fatigue quickly. All of these predictions are made via unobtrusive real-time collection of facial features. The models are even able to mimic thin-slicing and make predictions based on a short sample of data.

In addition to some of the implications discussed with the empirical studies, our models have application potentials for other studies of mass media that involve emotion and media effects. One area that would benefit greatly from adopting automated facial detection technology would be research on children and media. Most studies use self-reports to measure responses from children after they view television programs (Cantor, Mares, & Oliver, 1993; Cantor & Nathanson, 1996), but children have underdeveloped cognitive abilities compared to adults (Flavell, Miller, & Miller, 2002) and may have difficulty fully comprehending some of the questions used in scales developed for adults. Measurement of children’s emotions and responses via spontaneous facial expressions would be much more accurate than these self-reports.

This technology would also facilitate research of other media effects, such as ones that investigate the motivation and selection of media content. For instance, there have been numerous studies that incorporate Zillmann’s concept of mood management (Zillmann, 1988a), which explains media selection and consumption as a means for users to sustain positive moods or to terminate negative moods (Zillmann, 1988b; Knobloch & Zillmann, 2002). Using our system, researchers would be able to monitor emotional fluctuations in real time as participants make their selection of media content and verify whether or not the choices are indeed contributing toward managing an optimal state of happiness based on their facial expressions. Moreover, predictions of future media selection could be made based on the theory of mood management supplemented by analyses of facial expressions, further progressing current findings.

Another possible application in media research would be as a tool that supplements another methodology. One such example would be physiological measures that are used to
gauge real-time responses to media exposure to assess such variables as attention, arousal, and orienting responses. In addition to measures of facial electromyography (EMG), skin conductance, and heart rate, which are often the standards in physiological measurements (Bradley, Cuthbert, & Lang, 1996; Bradley et al., 2001), automatic facial expression tracking could either successfully replace any one of these measures or complement them. Because facial EMG involves obtrusive electrodes being placed directly on the face, and skin conductance and heart rate can reliably measure arousal but not valence (Winton, Putnam, & Krauss, 1984), the use of our proposed system could offer simple solutions to overcome these restrictions.

Beyond gauging intrapersonal responses, this technology could also serve to assist research endeavors of interpersonal communications. Interpersonal communication is vital to relationship management as it is a skill essential to the initiation, negotiation, maintenance, and termination of relationships between individuals (Burleson, Metts, & Kirch, 2000). For instance, prior research has demonstrated that communication success is one of the most important indicators of marital satisfaction (Steggell & Harper, 1991). Since interpersonal competence, or the quality and skills of interaction, differ from individual to individual (Spitzberg & Cupach, 1984), our system could improve the veracity and fidelity of interpersonal interactions by providing information from facial expressions that supplement the verbal message. More importantly, the system could aid in the resolution of interpersonal conflict, particularly when the complaint is withheld rather than confronted. Such unexpressed conflicts are harmful as they can lead to delayed and excessively emotional confrontations that the confronted person may think are unjustified (Baumeister, Stillwell, & Wotman, 1990). Our proposed methodology could help detect dissatisfactions even when they are not expressed verbally and prevent relationship failures.

Finally, we discussed earlier how facial expressions could yield meaningful cues for deception detection and consumer behavior, which are highly relevant to influence and persuasion. Studying interpersonal influences are important as data suggest that influence attempts are more often successful than not (Cody, Canary, & Smith, 1994). Our method of facial expression analysis could facilitate the efforts in compliance-gaining research. Many researchers have lamented about the shortcomings of relying on paper-and-pencil measures in lieu of behavioral measures, and basing their observations only on a single episode of message delivery rather than extended interactions (Berger, 1985; Miller, 1987). The nonverbal response data from tracking facial expressions will provide more concrete cues to behavioral changes induced by interpersonal influences compared to the use of self-reports alone. Furthermore, our system will endow researchers with the power to closely observe and analyze participant behavior for an extended period of time.

The face, as rich as its expressions are, is in no way a singular source of emotional expression that gives us insight into the human mind. Human emotion and its expressions are intricate combinations of mental, physical, internal, and external processes. Nevertheless, we focus on the face because it is the most evident signal of emotion and there is much to benefit from studying the face and its gestures. In the same way, the computer, lacking instinct or 'gut feelings,' may never perform entirely like the human mind. In fact, using computer vision and machine learning to predict human behavior the
system is far from flawless. But we focus on the system because such shortcomings pale in comparison to the obvious advantages offered by using automatic facial tracking and machine learning algorithms to predict behavior. If the face has a story to tell, our system of real-time behavior prediction based on facial expressions will be an effective methodological tool to help detect and interpret its message.
References


aggression, risky behavior, and crash-related outcomes in three groups of drivers. *Behaviour Research and Therapy, 41*, 333–349.


