2006-1700: MEASURING USER SATISFACTION BY DETECTING AND MEASURING EMOTIONS

John Fernandez, Texas A&M University-Corpus Christi

Dr. Fernandez is Assistant Professor of Computer Science in the Department of Computing and Mathematical Sciences. Having served 20 years in the U.S. Air Force and 10 years in private industry, Dr. Fernandez brings real-world experiences into the classroom for his students. His research interests are in HCI, information assurance, and software engineering.

Phillip Wilson, Texas A&M University-Corpus Christi

Mr. Phillip Wilson is a graduate assistant at Texas A&M University – Corpus Christi. He has a BS in Computer Science and Mathematics and is currently pursuing a Master's in CS. He has been a USAA intern for three years and has accepted an offer to work as an IT analyst/programmer for USAA upon graduation. His interests are in biometrics and information assurance.

Measuring User Satisfaction by Detecting and Measuring Emotions

Introduction

The measurement of user satisfaction was not much of a concern for the software community until the field of human computer interaction (HCI) became a recognized contributor to the discipline of computer science. HCI is concerned with the design, evaluation, and implementation of interactive computing systems for human use and with the study of related major phenomena[9]. The preponderance of research in HCI has been focused on the components of design and implementation of interactive computing systems. The main goal of HCI is to build interactive systems that are easy to learn, effective to use, and enjoyable from the user's perspective[9]. These characteristics are summed up in one word – usability. Usability can only be understood from the user's mind-set. Glass (as cited in Pressman[14]) contends that even the quality of a system is not as important as the user being satisfied, because if the user isn't satisfied, nothing else really matters. Therefore, usability is of utmost importance in measuring a software product's positive impact on the user.

Since the focus is on satisfying the needs and desires of the user, the development of interactive systems should follow a user-centered approach that is sensitive and responsive to those needs. This foundational principle becomes a challenge for the typical software developer because it calls for skills and abilities that are not normally part of the software engineer's portfolio[5]. HCI is an interdisciplinary methodology that involves skills from disciplines such as psychology/cognitive science, engineering, informatics, computer science/software engineering, ergonomics, human factors, and social sciences[13]. Therefore, HCI is closely aligned with user-centered development. Web-based systems are good examples of HCI-type systems because of the availability of the medium upon which they operate.

This paper addresses the measurement of user satisfaction by detecting emotions. The initial discussion deals with experiments conducted to measure emotions after the user had completed a test scenario. The paper then discusses the ongoing work to create an environment that will provide a real-time assessment of emotions. Before presenting the experiments and ongoing work, some preliminary discussions on the subject of user satisfaction are necessary.

User Satisfaction

Krug[7] points out that the real expert in usability is the user. Despite the best efforts of HCI experts and Web designers, the ones who ultimately determine if a Web site is useful (and therefore successful) are the users. Users, as a collective group, bring such a wide diversity of skill levels and backgrounds, that it is impossible for Web designers to anticipate and address every technical or cultural nuance. Therefore, it is worthwhile to give the users a voice in the design of Web sites through low fidelity prototyping and other techniques for user involvement.

Since developers are pursuing the user's approval, they are interested in the user's bottom-line satisfaction with the Web site. This phenomenon begins to enter into the world of users' emotional reactions to a Web site. According to Marcus[8], the HCI community's interest in emotions is heating up. Many questions arise about which emotions measure user satisfaction and Marcus includes a good discussion of the history of emotions research and states that this area of study is growing in scope for HCI professionals.

One can also approach the study of user satisfaction from a more traditional product evaluation perspective. Priesmeyer et al.[15] present an excellent case for measuring emotions to understand people's purchase intent of products and services.

Measuring Emotions To Detect User Satisfaction

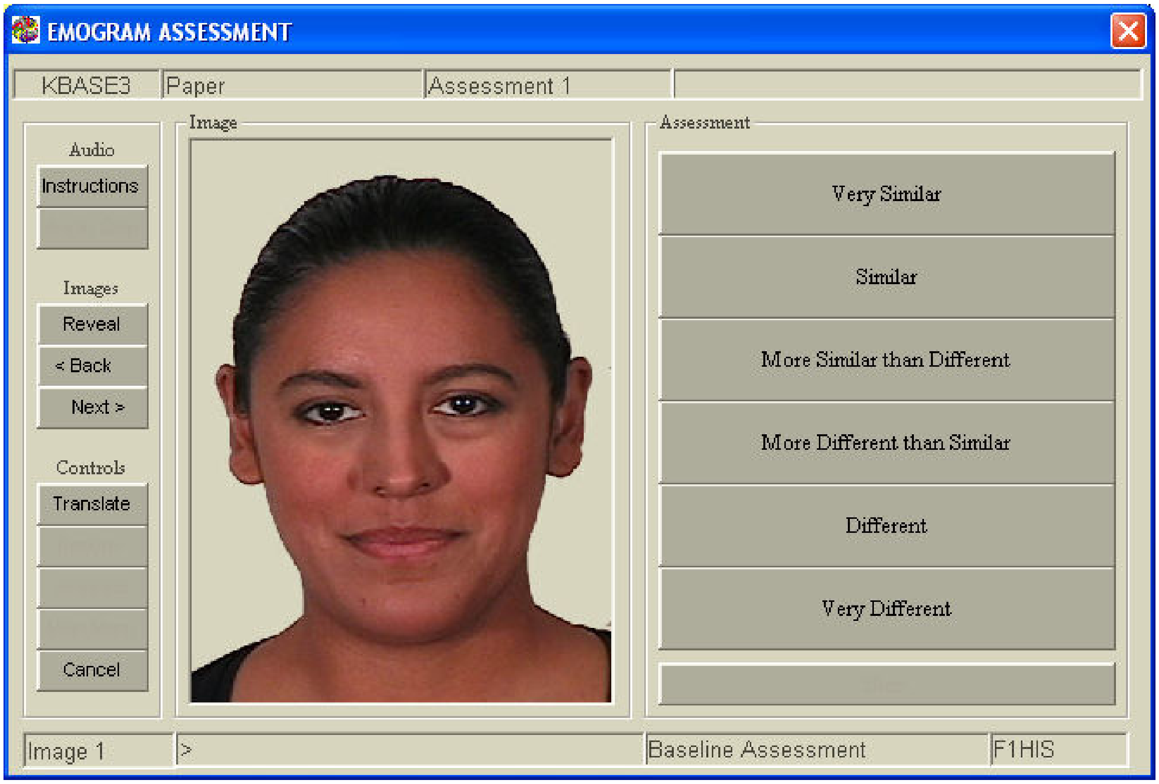
Priesmeyer et al.[15] developed a computer-based system that measures and interprets eleven human emotions. Although Marcus[8] presents a variety of studies that utilize a different number of emotions, there is no universal agreement as to the correct set or number of emotions that should be considered. Priesmeyer et al.[15] used the computer program, called the Emogram to determine the emotional quality of products as evaluated by a user.

Although Darwin made the first major contribution to the study of emotions, Paul Ekman (as cited in Priesmeyer et al.[15]) advanced the study of emotions by focusing on facial expressions. Ekman identified specific muscle groups that relate to specific emotions and developed a coding system (Facial Action Coding System) that allows one to identify an emotion from the combination of facial muscles used to express it.

Based on substantial research and cited studies, Priesmeyer et al.[15] decided to include the following list of basic emotions in the Emogram system: happiness, interest, surprise, contempt, disgust, shame, fear, anger, distress, sadness, and anxiety. Richard Lazarus (as cited in Priesmeyer et al.[15]) confirmed many of the basic emotions in this list through his own research and provided much discussion about the interaction of emotions that directly contributed to the interpretation and analysis of emotional dynamics in the Emogram system. An important step in the development of Emogram was the creation of precise photographs depicting varying degrees of the eleven basic emotions.

The Emogram system provides measures of each of the emotions by combining responses to low, medium, and extreme expressions of each. It also computes an overall Emotional Quality (E-Quality) score that reflects the overall emotional state of the individual at the time of the assessment. This E-Quality score is computed as the difference between the average of the pleasant emotions (happiness, interest and surprise) and the average of the unpleasant emotions (contempt, disgust, shame, fear, anger, distress, sadness and anxiety). The difference is then recalibrated to range from +100 to -100. The result is a measure that reflects a more satisfying overall emotional state as the measure approaches +100 and an increasingly uncomfortable state as the score approaches -100. In addition, the Emogram results for each emotion can range from a low value of one (1) to a high score of six (6). Some of the emotions are combined into six other measures computed from the emotional scores: openness, internal accountability, external accountability, congruence, incongruence, and relevance. Each of these additional measures can also range from a high of six to a low of one.

To measure the emotional qualities of a product using the Emogram, individuals take a baseline Emogram prior to being exposed to a product. The individual is then exposed to the product and then takes the Emogram a second time. Each Emogram assessment takes approximately six to eight minutes. The Emogram program then computes the strength of each emotion based on the reaction of the consumer and analyzes the changes in the emotions from the baseline to the second assessment. During the assessments the consumer is not asked to think about how he feels, but rather to respond to each photograph at an emotional level to more accurately reflect his or her emotional state at the time of the assessment. Specifically, the question posed to the individual is, “To what extent do you feel the way the person in the photo feels?” The Emogram assessment screen and the available responses to this question are shown in Figure 1 below[15]. This protocol effectively bypasses many of the problems associated with attempts to measure emotions with words. It should be noted that administrators of the Emogram must be trained by certified Emogram trainers and be awarded a license to use the instrument.



KBASE3 | Paper | Assessment 1

Image 1 | > | Baseline Assessment | F1HIS

Figure 1: Sample Emogram Evaluation Screen

Experiments Relating Emotions to Web Site Usability

Fernandez et al.[6] describe the details of several experiments with Web site usability. The results of the experiments were very interesting because it was quite clear that the E-Quality score of each student evaluator measured the degree of user satisfaction with the test scenario. Figure 2 shows some of the results of the preliminary experiment using Emogram for Web usability assessment. The E-Quality score of the baseline clearly shows that the student evaluators were feeling much better before completing the test scenario of the study.

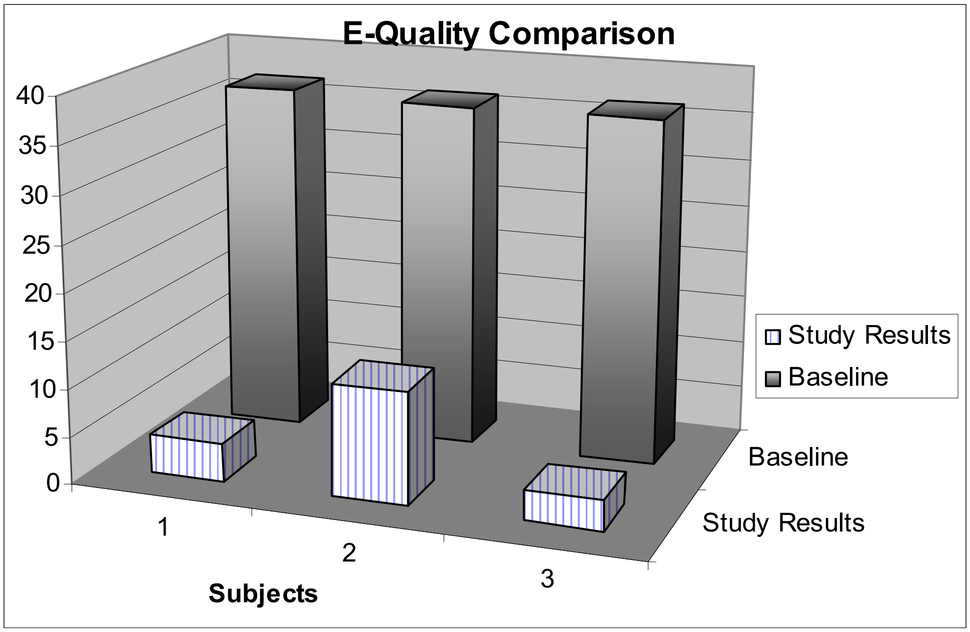


Figure 2: Comparison of Emotional Quality Scores

Multiple Web Sites Assessed with Emotions Measurement

Graduate and undergraduate students in HCI course were assigned Web site development projects for community organizations. Fernandez et al.[6] contain details on the work done with undergraduate students. Each development team prepared a test scenario based on the main tasks that the primary Web site users were supposed to be able to accomplish. The professor (J. Fernandez) provided a questionnaire (usability survey) to be used by each evaluator after the completion of the test scenario. This generalized questionnaire was standardized for use by all evaluators so that it could be compared with the Emogram assessments. Once the student completed the evaluation and the usability questionnaire, he or she was immediately administered the Emogram while targeting the student's experience with the Web site evaluated.

During the work with the graduate student projects and associated evaluations, Fernandez followed the additional step within the protocol to incorporate the administration of the Emogram prior to the student evaluation of the assigned Web site. The student evaluators followed the protocol of pre-evaluation or baseline Emogram, Webs site evaluation, post-evaluation usability survey and post-evaluation Emogram. The protocol provided a broader suite of data for analysis.

Two primary Web sites were used in the experiment. There were a total number of 21 student evaluators of these two Web sites: 11 females, 10 males, 13 minority and 8 anglo. Of the 21 student evaluators, 13 were computer science majors and 8 were other majors, ranging from math to business.

The scores on the usability survey were determined with the following mapping: strongly agree 20, agree 10, neutral 0, disagree -10, and strongly disagree -20. A perfect score on the usability survey was 100 reflecting a highly positive experience with a Web site. The results of the previous work by Fernandez et al.[6] permitted Fernandez to develop a four-part hypothesis using the usability survey. The hypothesis stated that there was no relationship between the survey results and the Emogram metrics: (1) survey results and post-evaluation Emogram are unrelated; (2) happiness and survey are unrelated; (3) anxiety and survey are unrelated; and (4) sadness and survey are unrelated.

The difference between the baseline Emotional Quality (EQ1) measure and the post-evaluation Emotional Quality (EQ2) was calculated and called EMDIFF. The difference between EQ1 and EQ2 measures (EMDIFF) was used for some of the analysis. For the data analysis, the usability survey score was named SURVEY.

Analysis of Results of Usability Tests

Using Pearson's Correlation, EMDIFF was compared to SURVEY and with a coefficient of 0.728 was found to be significantly correlated at the 0.01 level. This result appears to be based on the relationship of EQ2 and SURVEY. EQ2 and SURVEY had a coefficient of 0.738 which was significant at the 0.01 level. A mapping of the data is shown in Figure 3 below. EQ1 and SURVEY were found to have a negative correlation, but it was not statistically significant.

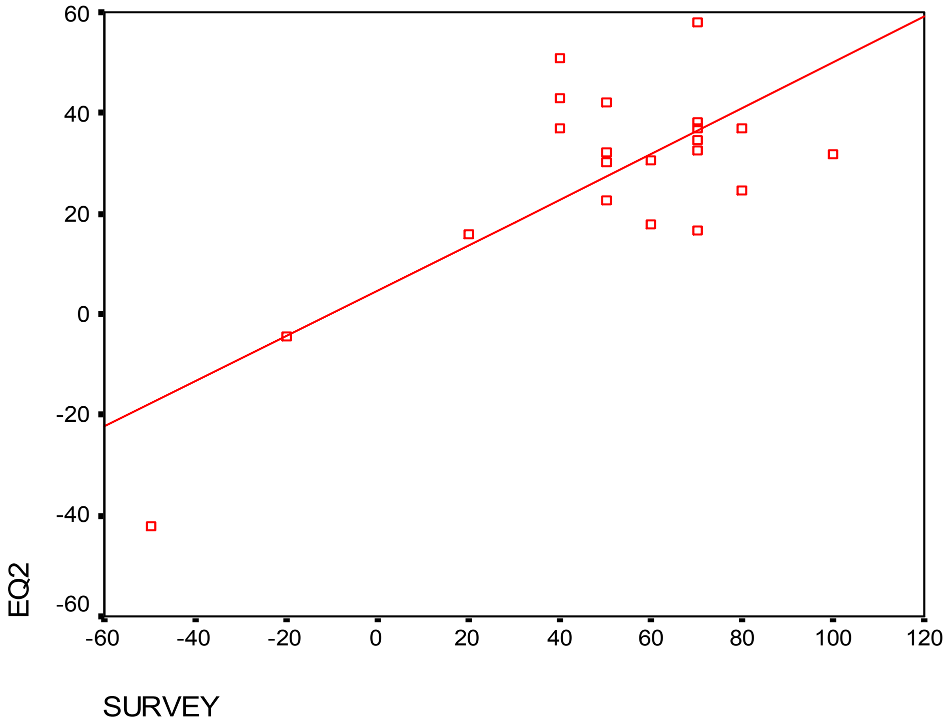


Figure 3. Survey Results and Post-Evaluation Emotional Quality Relationship

SURVEY was compared to happiness (HAPPINES) and they were found to have a positive correlation of 0.87, significant at the 0.01 level. Sadness (SADNESS) and SURVEY were found be negatively correlated, but this relationship was not found to be statistically significant. SURVEY and anxiety (ANXIETY) were found be to have a negative correlation of -0.87 which was statistically significant at the 0.01 level.

Three of the four statements of the hypothesis were rejected. There appears to be a relationship between (1) survey results and post-evaluation Emogram; (2) happiness and survey; and (3) anxiety and survey. However, the hypothesis that sadness and survey are unrelated could not be rejected.

Detection of Emotions in Real-Time

It is obvious from the discussion above that the user's emotional satisfaction with a Web site was measured after the conclusion of the actual experiment or use of the system. The next logical step of this research is to begin to move to real-time assessment of emotions or user satisfaction. The ideal evaluation scenario would permit researchers to be able to detect a user's reaction to a test scenario while he or she was actually in the process of using the interaction system.

In order to be able to accomplish emotion detection in real-time mode, there is a need to initially perform facial detection. This requires one to study some aspects of biometrics, the science of reading measurable, biological characteristics of an individual in order to identify them to a computer or other electronic system[2]. In recent years advancements in technology have made researching biometrics less expensive. Facial recognition, in particular, has become a field in which establishing a research environment has become much simpler. Current technology allows a real-time facial detection and recognition system to be created with very inexpensive hardware. Digital cameras with high resolution can be purchased for a marginal sum. Modern personal computers have enough processing capability to execute real-time detection algorithms.

The most difficult task to establishing the research environment is either creating the software or finding the software to perform the detection and recognition functions. Commercial off the shelf (COTS) software is often expensive and too inflexible for research. Intel has created a library, Open Computer Vision Library (OpenCV), to aid researchers in advancing the fields related to computer vision.[11] Facial detection and recognition falls into this category. Intel's library provides various algorithms to perform object detection and recognition. This library simplifies interfacing with digital cameras and allows a researcher the flexibility to modify standard facial recognition algorithms. It is open source and works on both Linux and Windows platforms.

Another difficulty with facial recognition is collecting enough images of human faces to acquire valid results from any experiments that are performed. The Facial Recognition Technology (FERET) database from the National Institute of Standards and Technology (NIST) solves this problem[4]. This database contains of 10,000 images of 1000 people that are used to test the accuracy of different facial recognition systems[12]. This database is free to researchers studying this field. One advantage of the FERET Database is that it provides many different images of the same individual. These images include different angles, and lighting conditions. In addition, some of the images are grayscale, while others are color images. Having various poses allows the training set to be composed of images with the same angles and lighting conditions, producing the most accurate eigen faces to be used in comparisons.

Haar-Based Detection

One of the difficult tasks of face recognition is to differentiate the face from the background. Each frame from a real time image is a collection of color and/or light intensity values. Analyzing each of the pixels to determine where the human face is located can be very difficult due to the wide variation in pigmentation and shape of the human face. Viola and Jones devised an algorithm, called Haar Classifiers, to rapidly detect any object, including human faces, using AdaBoost classifier cascades that are based on Haar-like features and not pixels[16]. Haar-based detection simplifies the detection of objects within a digital image.

Haar Features

Instead of using intensity values, Haar detection uses changes in contrast values between adjacent rectangular groups of pixels. The contrast variances between the pixel groups are used to find relative light areas and dark areas. Areas with contrast variances form features, as shown in Figure 4, which are used to detect the desired objects within the image[11]. These features can be easily scaled by increasing or decreasing the area of the pixels being examined. This allows features to be used to detect objects of various sizes.

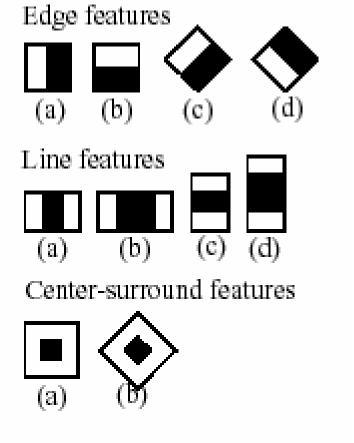


Figure 4. Haar features

The simple features of an image are calculated using an intermediate representation of an image, called the integral image[17]. The integral image is an array containing the sums of the pixel values to the left of a pixel and above a pixel at location (x, y) inclusive. So if **A[x,y]** is the original image and **AI[x,y]** is the integral image then:

$$AI[x,y] = \sum_{x' \leq x, y' \leq y} A(x',y')$$

The integral image array can be calculated with one pass through the original image. Using the integral image, only six to eight array references are required to compute a feature. Thus calculating a feature is extremely fast and efficient. It also means calculating a scaled feature requires the same effort as a non-scaled feature. Thus detection of various sizes of the same object requires the same amount of effort and time as objects of similar sizes.

Haar Classifiers

Although calculating features is efficient, computing the over 180,000 different rectangle features associated within a 24 × 24 sub-window is not feasible. However, only a small number of features are actually needed to determine if the sub-window contains the desired object[10,17]. The goal is to select the one feature that distinguishes the desired object, like a face, from another object. If a sub-window does not have that feature, then it can be eliminated. All non-eliminated sub-windows are then analyzed for a more features, creating a slightly more complex classifier. This process continues until the desired detection rate is achieved. This method of identifying an object is fast and efficient. Viola and Jones[16] were able to achieve a 95% detection rate of a human face with only 200 features. It took less than one second to identify the faces within a 384 × 248 pixel image. It is possible to achieve a higher degree of accuracy with more features with only a minute increase in detection speed.

The open computer vision library (OpenCV) provides a library to utilize Haar classifiers. In addition to the ability to use the classifiers it also provides several classifiers for face detection as shown in Figure 5. This detection classifier is capable of detecting a face in real-time on a personal computer running a 1 gigahertz or better processor using a fifteen frames per second video stream with a resolution of 320 × 240 pixels. OpenCV also has a program that has the ability to create new classifiers with a desired accuracy by providing images that contain the desired object and images that lack the object. This allows one to easily expand or modify a detection routine to be more effective or more precise. One could build classifiers to identify facial features as well as a human face.

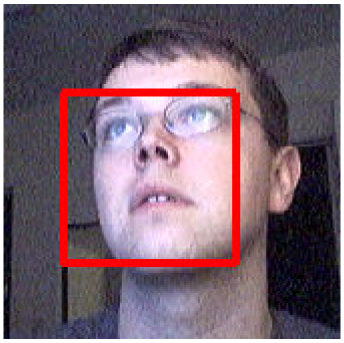


Figure 5. Detection of face in Web Camera Stream

Training Classifiers for Facial Features

Detecting human facial features, such as the mouth, eyes, and nose require that Haar classifier cascades first be trained. In order to train the classifiers, the AdaBoost algorithm and Haar feature algorithms must be implemented. Intel's OpenCV source library eases the implementation of computer vision related programs by providing for Haar classifier detection and training[11].

To train the classifiers, two set of images are needed. One set contains an image or scene that does not contain the object, in this case a facial feature, which is going to be detected. This set of images is referred to as the negative images. The other set of images, the positive images, contain one or more instances of the object. The location of the objects within the positive images is specified by: image name, the upper left pixel and the height, and width of the object[1]. For training facial features 5,000 negative images with at least a mega-pixel resolution were used for training. These images consisted of everyday objects, like paperclips, and of natural scenery, like photographs of forests and mountains.

Three separate classifiers were trained, one for the eyes, one for the nose, and one for the mouth. Once the classifiers were trained, they were used to detect the facial features within another set of images from the FERET database. The accuracy of the classifier was then computed as shown in Table 1. With the exception of the mouth classifier, the classifiers have a high rate of detection. However, as implied by Cristinacce et al.[3], the false positive rate is also quite high.

Facial Feature	Positive Hit Rate	Negative Hit Rate
Eyes	93%	23%
Nose	100%	29%
Mouth	67%	28%

Table 1 Accuracy of Classifiers

Regionalized Detection

Since it is not possible to reduce the false positive rate of the classifier without reducing the positive hit rate, a method besides modifying the classifier training attribute is needed to increase accuracy[3]. The method proposed to is to limit the region of the image that is analyzed for the facial features. By reducing the area analyzed, accuracy will increase since less area exists to produce false positives. It also increases efficiency since fewer features need to be computed and the area of the integral images is smaller.

This research is just starting so details of the work will be presented later. However, regionalization appears to be greatly increasing the accuracy of the detection of facial features. See Figure 6 below. Preliminary results indicate that all false positives were eliminated, giving a detection rate of around 95% for the eyes and nose. The mouth detection has a lower rate due to the minimum size required for detection. By changing the height and width parameter to more accurately represent the dimensions of the mouth and retraining the classifier the accuracy should increase the accuracy to that of the other features.

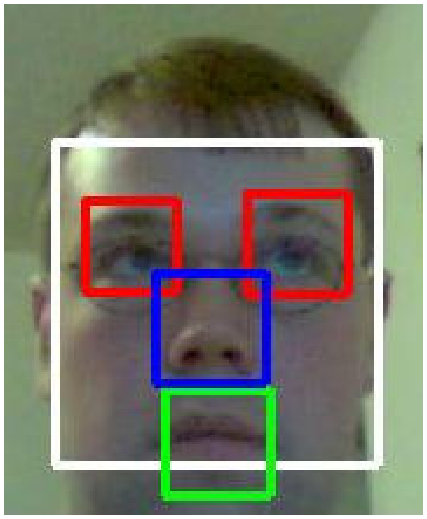


Figure 6. Detected Objects: Face (white),
Eyes (red), Nose (blue), and Mouth (green)

Conclusion and Future Work

This study shows that Emogram can be used to assess a user's emotional satisfaction with a Web site or any software system or product. To the extent that satisfaction is based on emotions, the Emogram scores can supplement traditional cognitive measures and provide useful information for improving software designs. A limitation of the Emogram is that it must used prior to and after the actual evaluation of a system or product. However, experimental results indicate that very beneficial data can be obtained with this assessment tool.

The research direction for studying user satisfaction is to build a method for detecting emotions while an evaluator is testing a system or product. This work is proceeding with the establishment of a facial feature detection environment, as discussed above, in order to determine the emotions of a user while engaged in running test scenarios on a system. By synchronizing the test scenario and the facial feature detection, more specific indications of user satisfaction can be obtained. One component of this work is to compare the real time analysis with the Emogram metrics. This will provide useful information about the staying power of emotions felt during testing which are evaluated with the Emogram after the fact.

Much related research can be pursued with real-time facial feature and emotion detection. Training and support systems for users may be able to adjust the speed and content of information presented based on the feedback of real-time emotions. The future holds much promise in this area of research.

References

1. Adolf, F., How to build a cascade of boosted classifiers based on Haar-like features. http://robotik.inflomatik.info/other/opencv/OpenCV_ObjectDetection_HowTo.pdf, June 20 2003.
2. Bishop, M., *Computer Security, Art and Science,.* Massachusetts: Pearson Education Inc., 2003.
3. Cristinacce, D. and Cootes, T., Facial feature detection using AdaBoost with shape constraints. *British Machine Vision Conference*, 2003.
4. The Facial Recognition Technology (FERET) Database, *National Institute of Standards and Technology*, 2003. http://www.itl.nist.gov/iad/humanid/feret/
5. Fernandez, J.D., Human-computer interaction closes the software engineering gap, *Computers in Education Journal,* vol. XV, no. 3, July – September 2005, 96-100..
6. Fernandez, J.D., Fernandez, M.A., & Priesmeyer, R., Experimenting with an emotions measurement instrument in usability testing, *Proceedings of 2005 ASEE Annual Conference*, Session 2658, June 15-17, 2005, Portland Oregon.
7. Krug, S., *Don't Make Me Think: A Common Sense Approach to Web Usability*, New Rider Press, Indianapolis, IN, 2000.
8. Marcus, A., The emotion connection, *Interactions,* November-December, 2003, 28-34.
9. McCracken, D. and Wolfe, R., *User-Centered Website Development: A Human-Computer Interaction Approach,* Pearson Education Inc., Upper Saddle River, NJ, 2004.
10. Menezes, P., Barreto, J.C. and Dias, J., Face tracking based on Haar-like features and eigenfaces, *5th IFAC Symposium on Intelligent Autonomous Vehicles*, Lisbon, Portugal, July 5-7, 2004.

11. *Open Computer Vision Library Reference Manual*, Intel Corporation, USA, 2001.
12. Phillips, P.J., Martin, A., Wilson, C.L. and Przybocki, M., An introduction to evaluating biometric systems, *IEEE Computer*, pp. 56-63, February, 2000.
13. Preece, J., Rogers, Y., and Sharp, H., *Interaction Design: Beyond Human-Computer Interaction*, John Wiley & Sons, Inc., New York, NY, 2002.
14. Pressman, R. S., *Software Engineering: A Practitioner's Approach*, *6th Ed.,* McGraw Hill, New York, NY, 2005.
15. Priesmeyer, H. R., Axiomakaros, P., & Murria, M. A., Human emotions and purchase intent: How to measure the emotional quality shift or products, services, and advertising, from proprietary workshop titled *Marketing Emotions,* a product of Chaotics International, 2003.
16. Viola, P. and Jones, M., Rapid object detection using boosted cascade of simple features. *IEEE Conference on Computer Vision and Pattern Recognition*, 2001.
17. Weyrauch, B., Huang, J., Heisele, B. and Blanz, V., Component-based face recognition with 3D morphable models, *IEEE Workshop on Face processing in Video*, FPIV04, Washington, D.C., 2004.