Our Emotions as Seen through a Webcam

Natalie Sommer[1], Leanne Hirshfield[2], and Senem Velipasalar[1]

[1] L.C. Smith College of Engineering and Computer Science,
Syracuse University, Syracuse, NY. 13210, USA
[2] S.I. Newhouse School of Public Communications,
Syracuse University, Syracuse, NY. 13210, USA
{nmsommer,lmhirshf,svelipas}@syr.edu

Abstract. Humanity's desire to enable machines to "understand" us drives research that seeks to uncover the mysteries of human beings and of their reactions. That is because a computer's ability to correctly classify our emotions will lead to an enhanced experience for a user. Making use of the eye of the computer, a webcam, we can acquire human reaction data through the acquisition of facial images in response to stimuli. The data of interest in this research are changes in pupil size and gaze patterns in conjunction with classification of facial expression. Although fusion of these measurements has been considered in the past by Xiang and Kankanhalli [14] as well as Valverde et al. [15], their approach was quite different from ours. Both groups used a multimodal set-up: an eye tracker alongside a webcam and the stimulus was visual. A novel approach is to avoid costly eye trackers and rely on images acquired only from a standard webcam to measure changes in pupil size, gaze patterns and facial expression in response to auditory stimuli. The auditory mode is often preferred since luminance does not need to be accounted for, unlike visual stimulation from a monitor. The fusion of the information from these features is then used to distinguish between negative, neutral and positive emotional states. In this paper we discuss an experiment (n = 15) where the stimuli from the auditory version of the international affective picture system (IAPS) are used to elicit these three main emotions in participants. Webcam data is recorded during the experiments and advanced signal processing and feature extraction techniques are used on the resulting image files to achieve a model capable of predicting neutral, positive, and negative emotional states.

1 Introduction

Pupil dilation and constriction are involuntary processes controlled by the Autonomic Nervous System (ANS). Knowing that it is also the ANS that gives us the ability to feel emotions, there is an obvious connection between changes in the diameter of our pupils and changes in emotions. Environmental factors such as lighting also change our pupils' sizes to control the amount of light that reaches the retina. Although responses to brightness levels are usually visibly apparent to the naked eye and considered primary, pupil size variations due to emotional factors are independent. While keeping the brightness level constant, pupil measurements can be an effective and

D.D. Schmorrow and C.M. Fidopiastis (Eds.): AC 2014, LNAI 8534, pp. 78–89, 2014.
© Springer International Publishing Switzerland 2014

unobtrusive way of understanding how a human feels about a visual or auditory stimulus. The sounds can be customized and delivered in a story-like fashion [8] or extracted from a database such as the International Affective Digital Sounds (IADS) [11] one. It is important to maintain the stimuli of various categories at a constant level to avoid the interference of dilation due to cognitive processing, another type of reaction of the ANS [10].

Emotions have different categories and the most common ones to be studied in the field of HCI are the valence and arousal dimensions. The valence dimension varies from negative to positive reactions whereas the arousal one varies from calm to very excited reactions. Users can rate their emotions after each stimulus using these categories and changes in pupil diameter have been able to differentiate between two groups in the arousal category: neutral and excited [1, 4, 8, 10]. Therefore, another facet to the problem presents itself. How can we devise a system to classify human emotion that is able to detect a difference between positive and negative emotions in the valence category? We want the design to be autonomous and, because of the use of a remote camera, unobtrusive. A possible solution lies in the fusion of different measurements such as pupil size and other facial features. Separately, pupil dilation has been able to differentiate between neutral and excited states whereas facial expression detection using a software such as FaceReader [17] has successfully classified emotion based on exaggerated features. Together, pupil size measurements and facial expression detection along with eye gaze can provide a finer tuned classification approach. Subtle, naturally occurring emotional states will have a better chance of being detected bringing us closer to the goal of replacing user rating input with other sets of data and giving a computer the ability to make the distinction between positive, neutral and negative emotional reactions without burdening its user.

Keeping in mind these various challenges, webcam data from 15 participants was collected while they listened to positive, neutral and negative auditory stimuli chosen from the IADS database. Participants were asked to remain as steady as possible while maintaining a constant unspecified distance from the camera. They sat within a controlled lighting environment while listening to the stimuli. These stimuli were chosen to elicit “compatible” positive and negative emotions (i.e. at the same intensity level). The neutral stimuli were chosen to be used as a frame of reference. Various image processing techniques were applied, including the use of wavelets in the extraction of pupil size, gaze patterns and parameters to analyze facial expressions. Changes in the shape of eyebrows and mouths and distances between eyebrows were the facial parameters of interest. Using a suitable classification process, results in differentiating between reactions from positive stimuli and neutral ones and ultimately from those that an HCI system would like to avoid, the negative ones, will be presented.

2 Experiment

Using the AVS Audio Editor 7.2, a fifteen minute track was created that included various types of sounds from the IADS database. These six-second sounds were

classified by the NIMH Center of the Study of Emotion and Attention according to valence, arousal and dominance ratings. Certain sounds were repeated to possibly assess whether a change of interest in the sound could be detected. As shown in Table 1, a total of 40 sounds were chosen based on similar arousal ratings in the different valence categories of neutral (N), pleasant (P) and unpleasant (U). In other words, all unpleasant sounds had high arousal ratings and all pleasant sounds except the Brook and Harp ones also had high levels of arousal. On the other hand, all sounds except for Alarm and Yawn that were rated as neutral in valence in the database had neutral arousal ratings. The soundtrack played these sounds in various sequences as shown in Table 1 to keep the participants engaged.

Table 1.

Sound #	1	2	3	4	5	6	7	8	9	10
Rating	N	P	U	N	U	P	U	N	P	U
Sound Type	Tropical	Baby	Attack1	Writing	Car Wreck	Crowd2	Scream	Rain1	Bongos	Victim

Sound #	11	12	13	14	15	16	17	18	19	20
Rating	P	N	P	N	U	P	U	N	N	N
Sound Type	Casino2	Writing	Boy Laugh	Shovel	Buzzer	Roller Coaster	Attack2	Rain1	Tropical	Radio

Sound #	21	22	23	24	25	26	27	28	29	30
Rating	P	P	P	P	U	U	U	P	N	P
Sound Type	Sports Crowd	Bongos	Baby	Baby	Buzzer	Buzzer	Alarm	Brook	Yawn	Harp

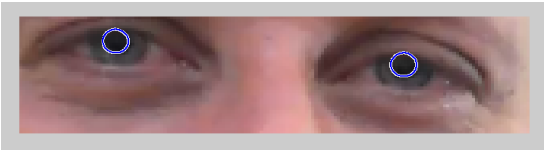
Sound #	31	32	33	34	35	36	37	38	39	40
Rating	N	P	U	P	U	P	U	N	P	U
Sound Type	Rain1	Harp	Scream	Casino2	Fem Scream2	Roller Coaster	Bees	Tropical	Brook	Alarm

Seated comfortably in front of a Microsoft LifeCam Studio 1080p HD webcam, volunteers (10 males and 5 females) between the ages of 20 and 30 were asked to look into the camera while listening to each of the first 30 sounds. They were then given 10 seconds to assess their emotions using the Self-Assessment Manikin (SAM) for 9-point valence and arousal scale ratings in between each six second sound. For the last set of 10 sounds, participants looked into the camera while listening to the sounds without having to provide ratings. This enabled us to gather data for future work on determining whether the rating process disrupts and alters reactions to sounds. Recordings of each volunteer were gathered with the latest Cyberlink UCam software and stored on an external hard drive for processing.

Some sounds were played several times in the user rating part of the experiment, the first 30 sounds. The Baby sound was played three times, Buzzer was played three times and Tropical was played twice. The Scream and Harp sounds were played once in that group. The Baby and Buzzer sounds were chosen to repeat more than once because of their respective high and low valence ratings and compatible database arousal ratings. The Tropical sound was repeated more than once as it was a good representation of a neutral sound according to the database ratings. The anticipated use of sound repetitions was not only to increase the number of samples in the three different valence categories but to also gather additional data that could be used to detect whether a loss of interest in the sounds occurred during the experiment.

3 Data Analysis and Results

At the beginning of the analysis, this sample group of five sounds (Baby, Buzzer, Harp, Scream and Tropical) was chosen and each participant's video recordings corresponding to these sounds were extracted to be analyzed. Matlab, along with its Image Processing and Computer Vision Toolboxes, was used to obtain key information from the frames and pixels of the videos. Using the vision.CascadeObjectDetector system object which uses the Viola Jones algorithm along with an eye pair classifier, frame images were cropped around eyes. Through a succession of image processing techniques that included sharpening, grayscale conversion and subsequent binary representation based on a threshold obtained from histogram analysis, pupils were represented with small groups of black pixels on a white background. The built-in function based on the circular Hough transform, imfindcircles, fit these groups of pixels into circles and their sizes (in a pixel to metric converted measurement) and centers (in vertical/horizontal location) were sent to an Excel spreadsheet as shown in Fig. 1.



BASED ON VIEWER'S PERSPECTIVE:		
PUPIL SIZE		
RIGHT:	0.24844	
LEFT:	0.240659	
PUPIL CENTER	X	Y
RIGHT:	120.38	17
LEFT:	30.95	9.38

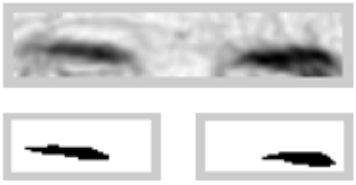
Fig. 1.

The same vision.CascadeObjectDetector system object along with the mouth classifier was used to locate this facial feature. Similar image processing techniques were used to detect the gap between the two lips. The pixels representing this gap were used in a curve fitting algorithm. The measure of mouth curvature was obtained from this calculation. Examples are shown in Fig. 2.

Fig. 2.

Additional features: eyebrow curvature and distance between brows were obtained. At first, an unsuccessful attempt was made at building an eyebrow classifier due to its unavailability in the current Computer Vision Toolbox in Matlab and in the Haar Classifiers of Open CV. Therefore, the eye detector along with a vertical movement of a cropping rectangle to the eyebrow region was used. Using a filter based on Gabor wavelets, eyebrows were segmented due to their different texture. Once the image was converted to a binary representation based on a histogram threshold, morphological erosion had to be used to thin out the eyebrows in certain cases.

Like the mouth, pixels for each brow were fit to a curve to obtain a curvature measurement. They were also used in calculating the distance between the brows (in number of pixels). A snapshot of this is shown in Fig. 3.



Volunteer's Right Brow				
Baby	Buzzer	Harp	Scream	Tropical
-0.01905	-0.00476	0.011696	0.005848	0.013054
Volunteer's Left Brow				
Baby	Buzzer	Harp	Scream	Tropical
0.016856	0.014286	-0.01818	0.010212	-0.00357
Distance between brows				
Baby	Buzzer	Scream	Tropical	Harp
15	17	16	16	13

Fig. 3.

As measurements were gathered for the pupils and eyebrows, results which indicated blinks were tagged and removed during analysis. Specifically, if the eyes could not be detected, the Matlab codes set pupil sizes and center locations to 0. Since eyebrow detection was based on eye detection, the curvatures of the brows and distances between them were also set to 0 during a blink. Mouth curvature readings did not suffer any discontinuity.

The hypothesis that the facial parameters acquired would be different in response to the various sounds was supported by ANOVA test results. Statistical difference was ascertained on the data collected for right pupil sizes, mouth curvature, left brow curvature and distance between brows for each participant for all five sounds. This analysis did not include repetitions of sounds. In other words, it was based on the first Baby sound heard, the first Buzzer sound, the first Tropical sound and the Scream and Yawn.

Upon consideration of the means of this experiment's Manikin ratings for these five sound samples, it is clear that the Baby sound evoked the most combined positive emotion (greater than average valence and arousal mean ratings) and the Scream evoked the most combined negative emotion (smaller than average valence and larger than average arousal mean ratings). This is illustrated in Fig. 4 (Note: Based on 9-point valence and arousal scales, the average values are 4.5).

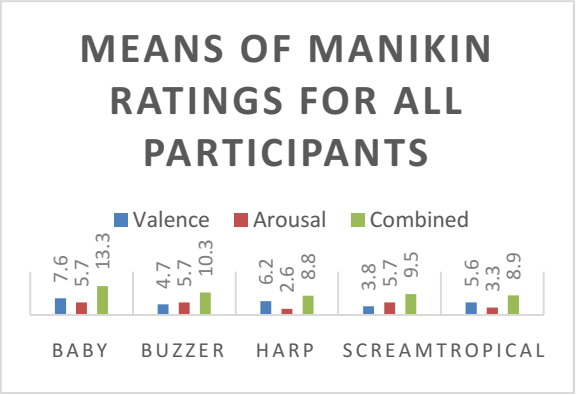


Fig. 4.

Observing that three sounds, namely, Baby, Buzzer and Scream had the same mean arousal ratings and that the Buzzer sound had a close to average valence rating, categories were assigned to them. The Baby sound was categorized as a pleasant one by the volunteers which was consistent with the IADS' rating. The Scream's unpleasant rating was also consistent with the database. The Buzzer was rated neutrally in contrast to the unpleasant rating given in the database. A factor worthy of consideration in future research is whether this difference can be attributed to the fact that this experiment's group of volunteers are city dwellers who are exposed to the daily sounds of ambulance sirens and buzzers of elevator doors rendering them somewhat cognitively immune to the Buzzer sound. It will be interesting to observe in the analysis portion of this experiment whether parameter measurements will end up supporting this neutral rating.

The analysis of parameters focused on these three sounds and the first times they were heard. When considering mouth curvature, the means obtained for each sound, when plotted, were able to make them distinguishable for most volunteers as illustrated in Fig. 5.

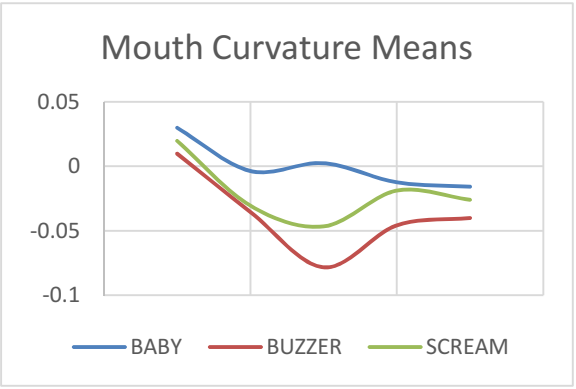


Fig. 5.

The second parameter that was considered was brow curvature. Before analysis, T-tests were performed to determine the statistical difference between each volunteer's left and right brow curvature. In all but two cases, there was no statistical difference between them. Therefore, a mean of the two curvatures was used in the special cases and only information for the left brow was used for the rest. Another interesting trend resulted in that most volunteers exhibited distinguishable reactions to these three sounds based on the maximum eyebrow curvature values (Fig. 6).

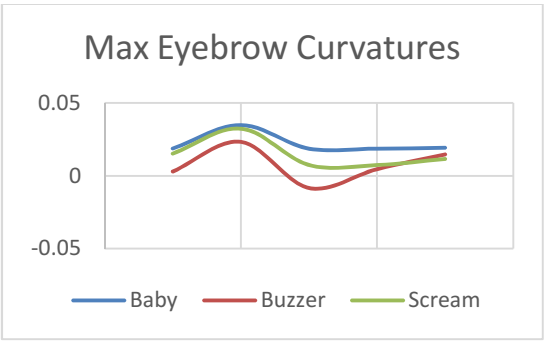


Fig. 6.

The minimum eyebrow curvatures were not as informative when plotted but did exhibit a distinction in their standard deviations (Table 2).

Table 2.

Standard Deviations of Minimum Curvatures		
Baby	Buzzer	Scream
0.008044	0.00666	0.003012

The third parameter which was considered was the distance between the eyebrows. Considering that the distance will decrease with a frown and will increase with surprise, the minimum and maximum of the distances were calculated and plotted in Fig. 7 to gain insight into the possible impact of this measurement.

The minimum distances show that the smallest distance occurred with the buzzer sound. Otherwise, even though visual distinction between the three sounds was difficult to make based on the plots, small differences existed making it a parameter worthy of consideration.

The final two parameter measurements, pupil size and x-y center coordinates were the most difficult to obtain. They both depend on a high level of precision which was difficult to obtain accurately if the color of the iris was close to pupil color.

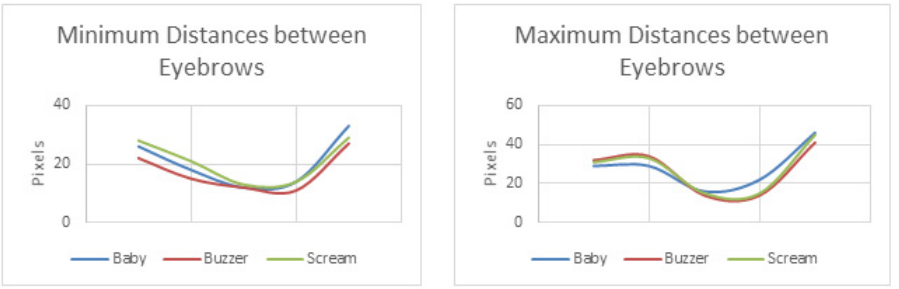


Fig. 7.

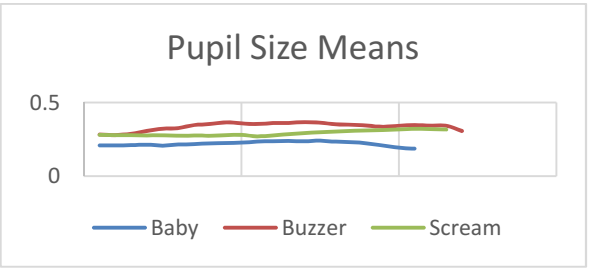


Fig. 8.

Therefore, the analysis shown below was based only on the subjects with light eye colors (n = 5). T-tests confirmed that the set of measurements for the right pupil were not statistically different to the set of measurements for the left pupil. Therefore, the left pupil's measurements were used for analysis. A moving average for each participant's left pupil size was taken and corresponding data point means were calculated and plotted in Fig. 8.

The differences in the means and reactions to the different types of sound could be established. Keeping in mind that data acquisition was based on images from a high resolution webcam and that a moving average was applied, the plots did not show the uniformity exhibited by Parlata and Surakka's [4] eye tracker plots. Nonetheless, it appears that the pleasant sound (Baby) did not elicit as great a change in pupil size as the other two sounds. The Buzzer, unlike the neutral rating it received in the experiment, seemed to startle the listener and kept him/her on high alert and the Scream triggered a delayed response, perhaps due to the volunteer's delayed reaction in understanding whether the scream was a joyful one or an unpleasant one.

Finally, the vertical and horizontal changes in eye position were considered. Although volunteers were instructed to look at the camera, we considered whether the gaze remained direct or if it became averted. According to Adams and Kleck [18], a direct gaze can be attributed to joy and an averted one is associated with fear. This means that the Baby sound would most likely evoke a direct gaze and the Scream, an

averted one. Using a moving average filter, saccades were filtered out and horizontal and vertical eye positions were plotted. Unfortunately, the results were inconclusive. An averted gaze which would necessitate a similar change in position for the right and left eye over several frames was not displayed in the generated analytical plots.

4 Classification Results

Using Matlab's Neural Network Toolbox, the Pattern Recognition Tool was used to train a classifier using the parameters presented, specifically, mouth curvature means, maximum eyebrow curvatures, pupil size means along with minimum and maximum distances between eyebrows. Vertical and horizontal locations of pupil centers were not included in the classification. The pool of data included five feature measurements for each volunteer for the first Baby sound, all three Buzzer sounds and the Scream. An additional volunteer's data was also included resulting in a total of 80 sets of facial and pupil measurements. Ten sets were saved for testing and the rest was used in a three output classification model representing the three valence emotions of interest. Utilizing ten hidden neurons and twenty-eight iterations, a successful pattern recognition classifier with a Mean Squared Error of $1.77761\ x\ 10^{-3}$ was built based on the majority of data collected. A small group was kept for testing after training. The resulting confusion matrix for testing, shown in Fig. 9, shows promising results.

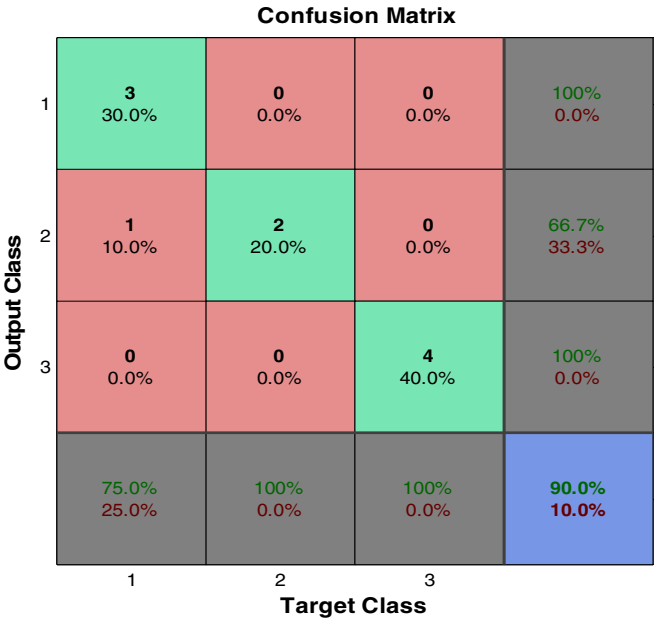


Fig. 9.

Class 1 represented the Baby sound, class 2 represented the Buzzer and class 3, the Scream. One out of ten groups of parameters was incorrectly classified into the Buzzer sound category. A successful classification rate of 90% was attained.

5 Additional Remarks

It is interesting to revisit a comparison of the ratings for the three sounds provided by the IADS database and the ratings given by the participants of this experiment. Although the database also shows the Baby and Scream sounds as pleasant and unpleasant, a significant difference lies with the Buzzer sound ratings. The database assigned a valence mean of 2.42 (arousal mean: 7.98) which would categorize it as unpleasant. Volunteers for this experiment rated it as a close to neutral sound (valence mean: 4.7; arousal mean: 5.7) as seen in Fig. 4. Experimental results, particularly, pupil size means seem to indicate greater than neutral reactions. Although reactions to sounds can be quite subjective, the inaccuracy in the pupil size measurements might be a factor. This will need further probing.

A preliminary analysis was conducted on the parameters of a small group of volunteers to determine whether hearing a sound more than once resulted in a different reaction. The Baby sound which was played three times during the soundtrack, although rated the same way, evoked different reactions indicative of a loss of interest. That is the reason only the first instance of this sound was included in the data used for classification. Reactions to the Buzzer sound did not exhibit this type of change and was included in the data. This outcome supports a neutral type of reaction by the volunteers. An analysis of the reactions to repetitive sounds requires further investigation.

6 Future Work

In addition to the previously mentioned study of loss of interest detection, analysis based on the other sounds used in the experiment will need to be explored. An attempt will be made at improving the understanding of the relationship between user rating and respective reactions modeled by facial and pupil size measurements. This means that the algorithm for pupil size and center location will need to be improved to better accommodate participants with dark eye colors. A second attempt at building an eyebrow detector will be made so as to avoid relying too much on the eyes to locate them. This way, eyebrows, will still be detected during blinks. Also, the data collected during the last ten sounds, which did not include user rating, will be compared to the reactions of similar sounds during the first thirty sounds which incorporated participant rating. It would be interesting to assess whether asking for user rating somehow interrupts emotional reactions. Finally, another parameter, such as Heart Rate Variability, will be explored as an additional unobtrusive measurement that can contribute to the assessment of emotion. This experiment will be repeated to increase the sample size and strive for a classification rate larger than 90%.

Acknowledgements. Thank you to the students and faculty of DeVry College of New York in midtown Manhattan who volunteered for this experiment.

References

1. Klingner, J., Tversky, B., Hanrahan, P.: Effects of visual and verbal presentation on cognitive load in vigilance, memory and arithmetic tasks. J. of Psychophysiology 48(3), 323–332 (2011)
2. Petridis, S., Giannakopoulos, T., Spyropoulos, C.: Unobtrusive Low Cost Pupil Size Measurements using Web cameras. In: Proceedings of the 2nd International Workshop on Artificial Intelligence and NetMedicine, pp. 9–20 (2013)
3. Schwarz, L., Gamba, H., Pacheco, F., Ramos, R., Sovierzoski, M.: Pupil and Iris Detection in Dynamic Pupillometry using the OpenCV Library. In: 2012 5th Annual Congress on Image and Signal Processing, pp. 211–215 (2012)
4. Parlata, T., Surakka, V.: Pupil Size Variation as an Indication of Affective Processing. International Journal of Human-Computer Studies, 185–198 (2003)
5. Canento, F., Fred, A., Gamboa, H., Lourenco, A.: Multimodal Biosignal Sensor Data Handling for Emotion Recognition. In: Proceedings of the IEEE Sensors Conference (2011)
6. Xu, G., Wang, Y., Li, J., Zhou, X.: Real Time Detection of Eye Corners and Iris Center from Images Acquired by Usual Camera. In: 2009 Second International Conference Proceedings on Intelligent Networks and Intelligent Systems, pp. 401–404 (2009)
7. Zhai, J., Barreto, A.: Stress Detection in Computer Users Based on Digital Signal Processing of Noninvasive Physiological Variables. In: Proceedings of the 28th IEEE EMBS International Conference, pp. 1355–1358 (2006)
8. Baldaci, S., Gockay, D.: Negative Sentiment in Scenarios Elicit Pupil Dilation Response: An Auditory Study. In: 2012 International Conference on Multimodal Interaction, pp. 529–532. ACM (2012)
9. Wang, W., Li, Z., Wang, Y., Chen, F.: Indexing Cognitive Workload Based on Pupillary Response under Luminance and Emotional Changes. In: 2013 International Conference on Intelligent User Interfaces, pp. 247–256. ACM (2013)
10. Partala, T., Jokiniemi, M., Surakka, V.: Pupillary Responses to Emotionally Provocative Stimuli. In: 2000 Eye Tracking Research & Applications Symposium, pp. 123–129. ACM (2000)
11. Babiker, A., Faye, I., Malik, A.: Non-conscious Behavior in Emotion Recognition. In: 2013 IEEE 9th International Colloquium on Signal Processing and its Applications, pp. 258–262 (2013)
12. Delibasis, K.K., Asvestas, P., Matsopoulos, G.K., Economopoulos, T., Assimakis, N.: A Real Time Eye-Motion Monitoring System. In: 16th International Conference on Systems, Signals and Image Processing, pp. 1–5 (2009)
13. Viola, P., Jones, M.: Robust Real-Time Object Detection. International Journal of Computer Vision 57(2), 137154 (2004)
14. Xiang, X., Kankanhalli, M.S.: A Multimodal Approach for Online Estimation of Subtle Facial Expression. In: Lin, W., Xu, D., Ho, A., Wu, J., He, Y., Cai, J., Kankanhalli, M., Sun, M.-T. (eds.) PCM 2012. LNCS, vol. 7674, pp. 402–413. Springer, Heidelberg (2012)
15. Valverde, L., DeLera, E., Fernandez, C.: Inferencing Emotions Through the Triangulation of Pupil Size Data, Facial Heuristics and Self-Assessment Techniques. In: 2010 Second International Conference on Mobile, Hybrid, and On-Line Learning, pp. 147–150. IEEE Computer Society (2010)

16. Bousefsaf, F., Maaoui, C., Pruski, A.: Remote Assessment of the Heart Rate Variability to Detect Mental State. In: 2013 7th International Conference on Pervasive Computing Technologies for Healthcare and Workshops, pp. 348–351. IEEE (2013)
17. Terzis, V., Moridis, C., Economides, A.: Measuring Instant Emotions Based on Facial Expressions During Computer-Based Assessment. Personal and Ubiquitous Computing 17(1), 43–54 (2013)
18. Adams, R., Kleck, R.: Effects of Direct and Averted Gaze on the Perception of Facially Communicated Emotion. American Psychological Association's Emotion 5(1), 3–11 (2005)
19. Bradley, M., Lang, P.: Affective Ratings of Sounds and Instruction Manual. In: The International Affective Digitized Sounds (2nd edn., IADS-2). Technical Report B-3, University of Florida, Gainesville, FL