# Thermal imaging as a way to classify cognitive workload

John Stemberger
Department of Computer Science and Engineering
York University
jms@cse.yorku.ca

Robert S. Allison
Department of Computer Science and Engineering
York University
allison@cse.yorku.ca

Thomas Schnell
Industrial and Mechanical Engineering
The University of Iowa
thomas-schnell@uiowa.edu

## Abstract

*As epitomized in DARPA's 'Augmented Cognition' program, next generation avionics suites are envisioned as sensing, inferring, responding to and ultimately enhancing the cognitive state and capabilities of the pilot. Inferring such complex behavioural states from imagery of the face is a challenging task and multimodal approaches have been favoured for robustness. We have developed and evaluated the feasibility of a system for estimation of cognitive workload levels based on analysis of facial skin temperature. The system is based on thermal infrared imaging of the face, head pose estimation, measurement of the temperature variation across regions of the face and an artificial neural network classifier. The technique was evaluated in a controlled laboratory experiment using subjective measures of workload across tasks as a standard. The system was capable of accurately classifying mental workload into high, medium and low workload levels 81% of the time. The suitability of facial thermography for integration into a multimodal augmented cognition sensor suite is discussed.*

## 1. Introduction

The task of operating a vehicle is difficult. This can be seen from the fact that between 1989 and 2007 there were 46 fatal aircraft accidents [2] resulting in over 1700 deaths. Even when disregarding the years in which accidents were caused by illegal acts (1994 and 2001) there were over 900 deaths. This means there is a death approximately every 31,000 departures, and with an estimated 840,000 domestic departures in a typical month (May 2005) in the United States alone [7] it would be a great relief if even a portion of these accidents could be prevented.

The situation is even worse for operating an automobile where from 1994 to 2008 there were approximately 560,000 fatal road accidents [6]. Ranney has determined that in 10.5% of car accidents the driver was distracted [12].

If a method was available to detect potentially dangerous mental states such as distraction, then it is conceivable that methods for minimizing or preventing these mental states could result in fewer deaths and collateral damage. Assessment of cognitive workload can also be used to tune and monitor tasks with the goal of improved productivity and awareness [11], particularly for jobs that require high attention but are not cognitively stimulating such as security monitoring.

The current system is intended for integration into an advanced multi-modal avionics suite for cognitive workload assessment and mitigation at the Operator Performance Laboratory at the University of Iowa. The system integrates an advanced AugCog suite of sensors and software into a small aircraft. This augmented aircraft serves as a testbed for assessing the utility of intelligent autonomous systems to increase efficiency, inter-operability and safety of human-in-the-loop control in a realistic flight environment [13]. Currently the system integrates a number of sensors including gaze tracking, EEG nets, pulse oximeters and thermal cameras. The present paper describes our efforts to assess the suitability of using thermal imaging to classify mental workload as a component of the augmented cognition sensor suite.

In past research, thermal imaging has been linked to specific emotional states such as stress [11] or deception [10, 8] but never to a spectrum of mental workload levels ranging from low to high. Finally, with inputs from many varying sources it is desired that these sources can be merged together to improve accuracy. We looked at the use of artificial neural networks (ANN) to classify each thermal image

978-0-7695-4040-5/10 $26.00 © 2010 IEEE
DOI 10.1109/CRV.2010.37

into either a low, medium, or high level of workload. The ultimate goal is to prevent extremely low workload levels (which could lead to boredom and inattention) or extremely high workload levels (which could lead to decreased performance or the operator being overwhelmed).

The logic for use of facial thermography is based on the known relationships between cardiovascular physiology and mental states. Cardiovascular measures have been shown to reliably differentiate between emotional states [14]. Since the temperature of the face is directly related to the conduction of heat from the blood to the surface of the skin we believe that thermal temperature will significantly correlate with these mental states.

## 2. Background Information

One problem with the augmented cognition concept is that all sensor-based techniques to infer cognitive states are limited; there is no silver bullet solution. Limitations include practical issues such as attachment of sensors, visibility, noise, interference and artefact. Most measures are also indirect, measuring correlates of the cognitive state rather than being specific to the psychological parameter of interest. Like other techniques, detection of mental workload through the use of thermal imaging has pros and cons.

A primary advantage of thermal imaging is the non invasive, contact-free sensing. By removing contact with participants, a greater level of comfort can be achieved. Also it is simply not efficient to have people to be connected with wires to monitoring devices in many applications. Loss of direct sensor contact also forms a problematic mode of failure for many techniques such as EEG.

Objective workload measures rely on phenomena that are not under the voluntary control of the user. Since the facial temperature of a person is directly related to the blood flow rate within the tissue of the face [8] and the blood flow (part of the cardiovasular system) is controlled by the autonomic nervous system [3], it can be classified as an objective measure in the same manner as an EEG signal. This control is exercised at a local level to direct blood flow to tissues and organs so that cognitive state is reflected in local flow patterns as well as in whole-body parameters such as heart rate. This can lead to more reliable readings than subjective measures.

Finally, with the development of uncooled thermal cameras, a facial thermography system can be implemented with cameras that are not much bigger than a standard web camera for a personal computer.

The disadvantages of thermal imaging include artifacts from environmental effects and metabolic effects of digestion [8], occlusion of ROI by eye glasses or hair bangs [8], and the current high cost of small thermal cameras. We also encountered a difficulty of tracking the ROI within the thermal image due to the nature of intensities changing not as a result of motion but as a result of changes in workload. To compensate for this we added a tracking system which would ease deployment of any commercial application. Such head tracking is currently built into the AugCog sensor suite in the aircraft so this data is readily available in the target system.

### 2.1. Thermal Imaging

Several different mental states have been correlated with changes in facial temperature.

Pavlidis et al. explored how facial temperatures changed depending on the activity being performed [9]. Using a Raytheon ExplorIR thermal camera type, six participants were imaged as they performed a battery of tasks including resting in the dark, a 60dB startle stimulus (a sudden loud sound intended to increase alertness, anxiety and fear), chewing gum, and mild physical exertion. Each thermal image collected was segmented into five ROI consisting of a section around both eyes, the left and right cheeks, the nose, chin, and neck.

Within 300 ms of the startle stimulus an increase in thermal intensity was recorded around the eyes and the carotid with an accompanying decrease in temperature of the cheeks. Similarly when chewing gum a warming of the chin area was seen. Finally, with mild exertion a slow cooling of the nasal area was observed. These results let Pavlidis et al. to conclude that unique facial thermal patterns can be associated with different activities.

Puri et al. found that the thermal intensity of a rectangular region within the forehead was shown to correlated with stress levels in 12 participants performing a Stroop colour word conflict test [11].

Finally in 2006, two applications for the detection of effective learning rates and for the detection of concealed information using thermal imaging were described.

In the first, Kang et al. showed that nose temperature of participants learning an unfamiliar arithmetic operation increased as they became more familiar with the operation [5]. This increase in nose temperature was correlated with a decreasing response time, increasing response accuracy, and a decreasing subjective rating of mental workload (as measured using a Modified Cooper Harper scale). In total, nine participants learned to verify addition between the numbers 1, 2, 3, and 4 with the numeric codes for the letters C, D, and E over a set of seven blocks consisting of 96 questions.

In the second application, Pavlidis and Levine as well as Pollina et al. looked at the use of thermal imaging to improve detection rates of polygraph tests [8, 10]. By having participants reenact a murder scenario and administering a traditional polygraph test, both were able to show a connec-

tion between knowledge of the crime and the facial temperature.

Pavlidis and Levine were able to correctly classify 84% of participants as either deceptive or non deceptive. This was done by converting the thermal intensity of the periorbital region of skin around both the left and right eyes to a blood flow rate using the algorithm described by Fujimasa et al. [4]. This represents an improvement over traditional polygraphs systems that only achieve a correct classification rate of 78%.

Even more impressive is the 91.7% correct classification over 24 participants achieved by Pollina et al. based on analysis of two 10 by 10 pixel squares below the left and right eyes.

While all of these studies show significant results, only two of them demonstrate potential in a real-world application, in this case polygraph testing. Although Kang et al. demonstrated a link between nose temperature and learning levels, they did not attempt to estimate current level of learning achieved given a nose temperature. In contrast, we are interested in estimating workload levels. Thus, we looked not only at the correlation of cognitive workload with thermal intensity, but also at how that thermal intensity could be used to classify an operators' current mental state.

## 3. Methods

In order to evaluate our system we had 12 participants (6 male and 6 female) take part in a cognitive stress test (CST) consisting of three blocks. After the session each block was given a post-hoc subjective rating of mental workload by the participants on a scale of 1 to 7, 1 being extremely low, and 7 being extremely high.

The three blocks of the CST were randomly ordered, to compensate for any learning or order effects, resulting in six different orders (each of which was tested with 2 participants). There was at least a four-minute rest period between blocks in which the participants were able to rest and lower their cognitive load to a resting state. During this period subjects were allowed to remove the tracking headset and move around freely.

The cognitive stress test was based on Berka et al. [1] and consists of three workload levels. In their work Berka et al. demonstrated the ability to categorize cognitive workload using a six-channel wireless EEG system called the B-Alert system. Berka et al. varied cognitive workload over blocks consisting of 250 trials in which a digit between 1 and 8 was presented at a rate of 1.6 digits per second. For the present experiment, we used similar techniques to provide three levels of relative workload: low, medium and high.

For our experiments, the first level of workload (low workload) asked the participants had to press the mouse button when they saw the number 5. This simple recognition paradigm was fairly trivial to perform.

For the second level of workload (moderate workload) the participants had to press the mouse button only when three even numbers were displayed consecutively. This increased the cognitive load because the size of the set to be recognized by the participant increased from just the number 5 to the numbers 2, 4, 6, and 8. Also, requiring the participants to recall whether the last two digits were even added a working memory component to the task.

The final level of workload (high workload) required the participants to press the mouse button when they saw a digit that was identical to the the digit displayed two trials earlier (2-back). In this way we expanded the set of digits of interest to be all eight digits as well as requiring the use of working memory to store the specific value of the last 2 trials rather than just an odd/even classification.

For our experiment each participant performed 600 trials for each of the workload blocks in a random counterbalanced order. The probability for any given digit appearing was identical with each digit presented 75 times per block. The frequency of the target condition was equated across the three tasks so that the target sequence occurred 75 times for each of the thee workload blocks. Due to random system delays on the recording computer not all stimuli were presented for exactly 1.6 seconds. Only trials that were presented for less than 1.7 seconds were analyzed for performance.

While performing the stress test blocks, each participant's face was imaged using an Indigo A10 thermal camera capturing frames at a rate of 15 frames/sec. Each frame of the video was separated into ROI similar to four of the five ROI defined by [9] as well as a forehead region similar to [5, 8, 11]. We also distinguished between left and right sides similar to [10] in the hopes that this could be used at a later date to compensate for any possible environmental effects such as a warming of a part of the face due to direct light. Thus, the ROI were the forehead, nose, eyes (peri-orbital), left cheek, right cheek and chin (Figure 1 and Figure2). The neck region was not used due to an inconsistent ability to view the region within the small field of view of the camera ($25^\circ$ x $19^\circ$). To facilitate separation of the ROI from the background an InterSense 900 hybrid ultrasonic inertial tracking system headset was worn by the participants. This allowed a generic 3D head model (Figure 1) to be used to track the movement of and separate each of the ROI. Unfortunately the added tracker negated the non-invasive benefits of using thermal imaging. In the target system this is not a concern since the system was designed to allow for the easy swapping of the InterSense 900 tracking system with the SmartEye tracker (a non-contact infra red tracking system) present in the AugCog suite of the flight platform.

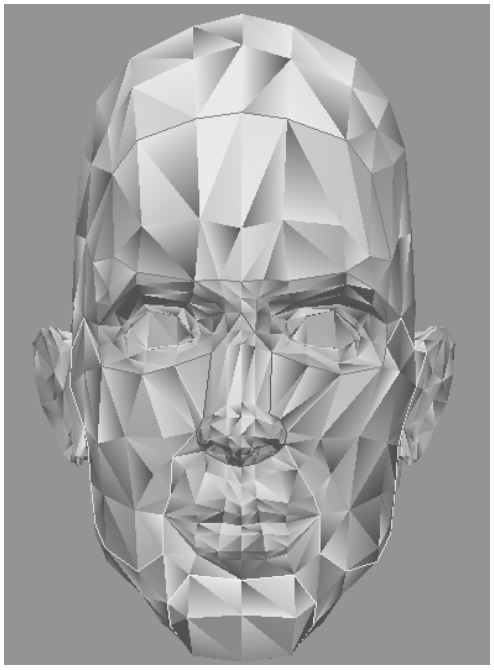

**Figure 1. Head model used for all participants**

**Figure 2. Approximate ROI as seen on a thermal image**

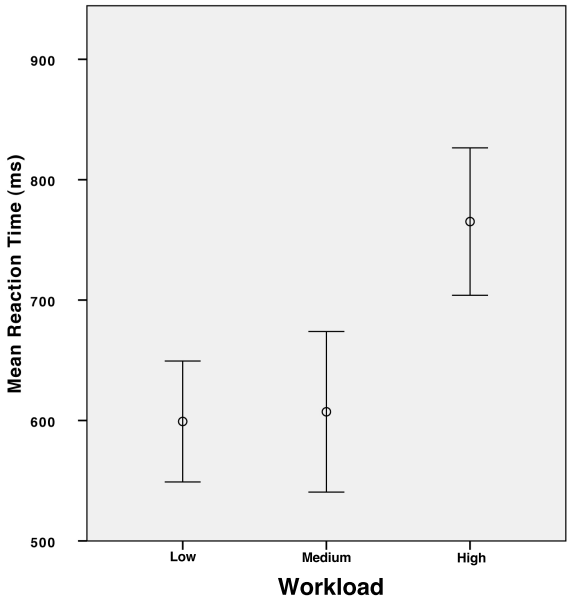


**Figure 3. Mean reaction time across all subjects versus condition. Error bars show 95% Confidence Intervals**

## 4. Results

To determine the relative workload entailed for each block, the reaction time for each trial in which a correct response was made was recorded and analyzed as a function of block. Figure 3 depicts mean reaction time (in milliseconds) broken down by workload condition type. A one-way repeated-measures ANOVA indicated a significant change in reaction time as a function of workload level ($F(2) = 25.659$, $p < 0.001$). Post-hoc analysis (with Bonferroni correction) revealed that participants' reaction times were significantly higher in the high workload block (M = 765.21, SD = 108.25) than in the low (M = 599.20, SD = 88.76) or medium (M = 607.24, SD = 117.83) workload blocks ($p < 0.001$).

Figure 4 shows mean percentage of trials in which a correct response was made within each condition. A chi-squared test of independence indicated that the mean percentage of correct responses varied as a function of workload level ($\chi^2(2) = 255.139$, $p < 0.001$). Participants made fewer correct responses in the high workload workload (M =54.75, SD = 9.94) than in either the low (M=68.33,SD=2.87) or medium (M=68.58, SD =8.03) workload conditions ($p < 0.001$). Correlation analysis confirmed the negative relationship between task difficulty and percentage of correct responses ($r(33) = -0.64$, $p < 0.001$).

These results imply a definite increase in the difficulty (and thus mental workload) from the low and medium con-

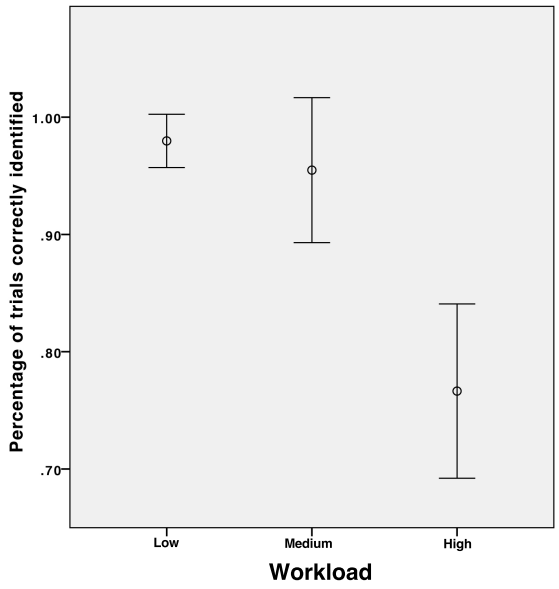


**Figure 4. Mean percentage of correct identification versus condition. Error bars show 95% Confidence Intervals**

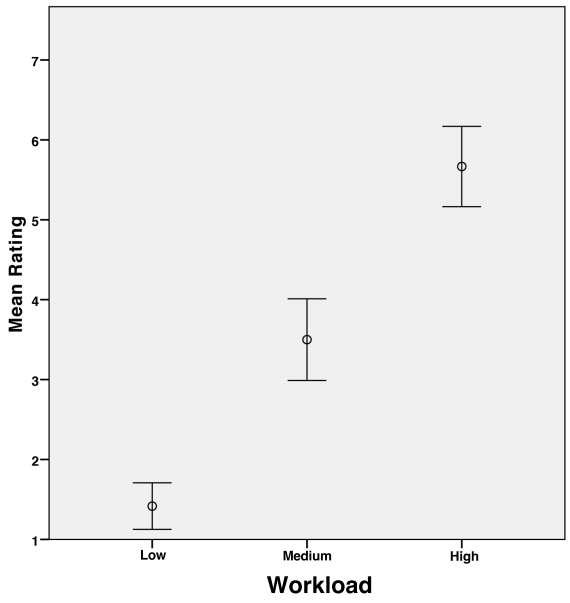


**Figure 5. Mean subjective rating versus condition. Error bars show 95% Confidence Intervals**

ditions to the high condition. On the other hand performance measures did not show a difference between low and medium workload blocks. However, participants can work harder to perform one task compared to another without a degradation in performance if they have spare capacity. Thus as performance measures can suffer from plateau effects caused by spare capacity, a subjective workload measure (using a 7-point Likert scale) was also obtained. The mean subjective workload rating broken down by condition is presented in Figure 5. Again a one-way repeated-measures ANOVA indicated participants' subjective ratings varied as a function of workload ($F(2) = 76.374$, $p < 0.001$). Here we can clearly see that despite no significant differences in performance between the low and medium workloads, participants felt their mental workload was significantly higher for medium workload than low workload.

The above results indicate that participants' mental workload differs across the conditions. However, are we able to detect these changes in the patterns of facial temperature? To determine if changes in facial temperature were indicative of changes in mental workload, we took the average temperature for each of our seven ROI and fed them into an SPSS PASW multilayer perception network [15]. The input layer took in each of the seven ROI and rescaled the values by subtracting the mean of each input and dividing by the standard deviation. The hidden layer used an hyperbolic tangent function (see equation 2) where input values where from the interval (-1, 1). The number of neurons in the hidden layer was determined by PASW using an estimation algorithm on a random sample of 50% of all video frames collected for all participants resulting in the seven hidden layer neurons (See Figure 6). 20% of the training data used to estimate the optimal architecture was used to prevent over training. Finally, the output layer used a softmax activation function (equation 2)

$$\gamma(c) = tanh(c) = \frac{e^c - e^{-c}}{e^c + e^{-c}} \quad (1)$$

$$\gamma(c_k) = \frac{exp(c_k)}{\Sigma_j exp(c_j)} \quad (2)$$

Using the remaining 50% of the data not used for training, the network achieved an 81% correct classification rate (76.9%, 79.2%, and 86.8% correct classification for low, medium, and high workload respectively; See Table 1 for details).

A second network was trained using the same proportions of data but from only a single randomly selected subject resulting in the architecture seen in Figure 7 and achieving an overall correct classification rate of 98.9% (97.6%, 99.8%, 99.3% respectively for each low, medium and high workloads; See Table 2 for details).

As a final step in our analysis we attempted to interpret the classifier in terms of the underlying physiology. General trends across the different workloads were looked at by

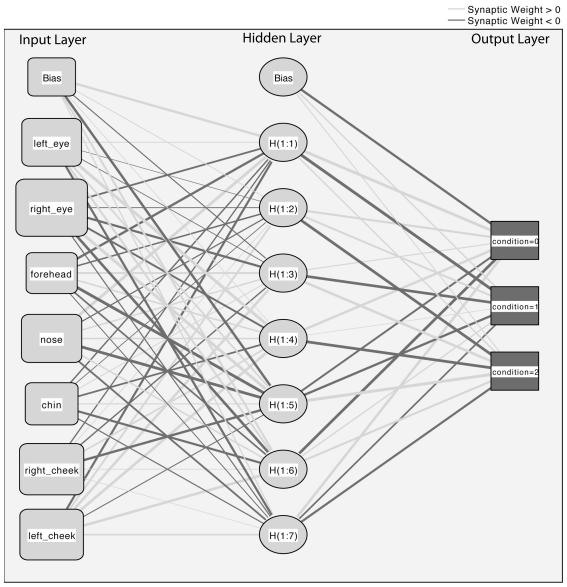


**Figure 6. Architecture for the ANN used to classify workload**

**Table 1. Confusion Matrix showing how often the trained ANN (all data sets) mistook one workload for a different workload**

| | | Actual | | |
|---|---|---|---|---|
| | Workload | Low | Medium | High |
| Predicted | Low | 59937 | 9070 | 3700 |
| | Medium | 9820 | 62486 | 7010 |
| | High | 8231 | 7300 | 70343 |
| Percent Correct | | 76.9% | 79.2% | 86.8% |

**Table 2. Confusion Matrix showing how often the trained ANN (single participant data set) mistook one workload for a different workload**

| | | Actual | | |
|---|---|---|---|---|
| | Workload | Low | Medium | High |
| Predicted | Low | 6604 | 7 | 42 |
| | Medium | 3 | 6688 | 5 |
| | High | 159 | 5 | 6722 |
| Percent Correct | | 97.6% | 99.8% | 99.3% |

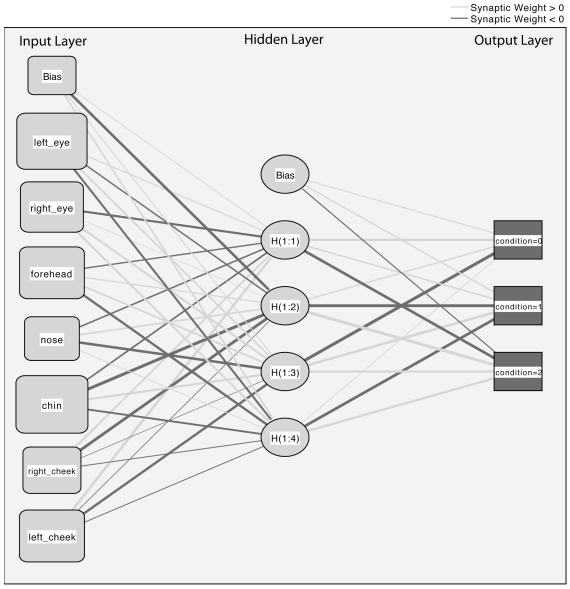


**Figure 7. Architecture for the ANN used to classify a single participants workload**

taking the average temperature across the entire face and across the entire session (see Figure 8). In these diagrams, each line corresponds to one of the 12 subjects and each data point is the average infra-red intensity recorded across the face for each of the three workloads. As several participants had a decreased thermal intensity from low workload to medium workload, and others have an increased thermal intensity, a consistent pattern was not observed. Similar results were seen when we looked at individual regions across the entire session (see Figure 9 and Figure 10).

## 5. Discussion

It is apparent from our study that our cognitive stress test did produce an appropriate change in mental workload across the experimental conditions. Furthermore these changes in subjective and objective workload produced reliable facial thermography signatures that could be detected by an uncooled thermal camera and analyzed with an ANN.

One interesting observation is that while the thermal signatures appear to be differentiable they don't appear to be consistent across participants. With workload being a complex psychological constant it is possible that each individual has a unique signature. It is also a possible explanation for why the single person neural network only requires 4

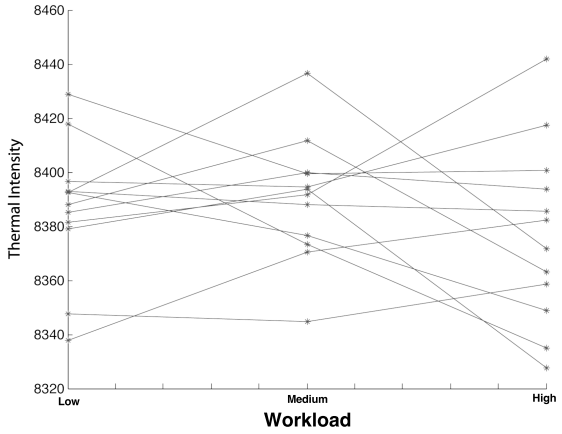


**Figure 8. The average raw intensities recorded across each session (all ROI) for each participant.**

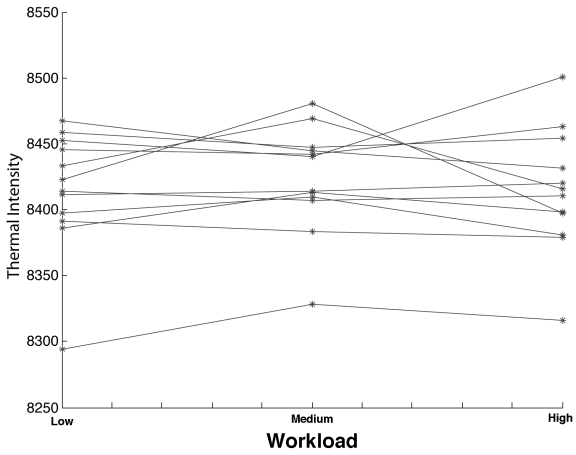


**Figure 10. The average raw intensities recorded across each session (left eye) for each participant. The eye temperature has previously been shown to be in indicator of stress levels [9] and deception [10]**

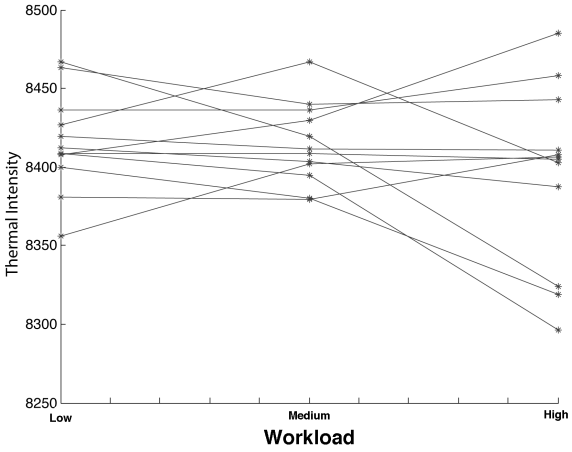


**Figure 9. The average raw intensities recorded across each session (nose) for each participant. The nose temperature has previously been shown to decrease as learning levels increase.**

hidden layer neurons instead of the 7 that are required for all participants and achieved a superior level of classification. This leads us to believe that for any commercial application an ANN trained for each user would be required to minimize failed classifications.

This system provides scientists with a framework for posing research questions and basing future studies. This base will allow for studies that test the generalization to real work tasks. Part of this generalization to real work tasks will be to look at what, if any, effects the environment has on thermal imaging. It is trivial to imagine difficult scenarios such as when a pilot is flying perpendicular to the sun so that half of his or her face in being heated by solar radiation and the other half is in shade. Also, the effects of biological operations such as digestion can be investigated and would allow for an improved level of detection in scenarios where pilots are flying for extended periods and eating meals or drinking coffee.

If the neural network classifier provides similar discrimination power between workload levels in the cockpit then flight operations could be safer and more reliable. This technology could also be applied to other applications such as automobile drivers, heavy equipment operators, and security guards. The detection of underload could be used to help maintain vigilance.

Another area in which this system will be of benefit is the study of mitigation strategies. If the automated workload analysis provided by the present system was merged with the data collection processes then a complete real-time workload estimation system could be implemented. With

such a tool mitigation studies would be able to get real-time feedback enabling study of the effects of different mitigation strategies on mental workload.

Once a real-time system has been implemented and integrated with the other physiological systems to make a multimodal and robust system and appropriate mitigation strategies have been developed, air travel will become a safer mode of transportation as well as a less stressful operation for the pilot.

## 6. Conclusions

We have been able to demonstrate that facial thermography can reliably quantify participants' workload in various cognitive tasks. Through the use of an artificial neural network our system was able to correctly classify the majority of thermal images into multiple levels of workload. This provides the foundation for future work in multimodal augmented cognition and demonstrates the potential of facial thermography for the estimation of cognitive state.

## References


[1] C. Berka, D. J. Levendowski, M. M. Cvetinovic, M. M. Petrovic, G. Davis, M. N. Lumicao, V. T. Zivkovic, M. V. Popovic, and R. Olmstead. Real-time analysis of eeg indexes of alertness, cognition, and memory acquired with a wireless eeg headfset. *International Journal of Human-Computer Interactions*, 17(2):151–170, 2004.

[2] National Transportation Safety Board. NTSB Aviation Accident Statistics, *http://www.ntsb.gov/aviation/Table6.htm* Accessed: January 2010.

[3] N. R. Carleson, W. Buskist, M. E. Enzle, and C. D. Heth. *Psychology: The Science of Behaviour*. Pearson Education Canada Inc, Needham Heights, MA, 3rd edition, 2005.

[4] I. Fujimasa, T. Chinzei, and I. Saito. Converting far infrared image information to other physiological data. *IEEE Engineering in Medicine and Biology Magazine*, 19(3):71 – 76, 2000.

[5] J. Kang, J. McGinley, K. Babski-Reeves, and G. McFadyen. Determining learning level and effective training times using thermography. In *Proceedings of Army Science Conference*, Orlando, Florida, USA, September 2006.

[6] U.S. Department of Transportation. Fatality Analysis Reporting System Encyclopedia, *http://www-fars.nhtsa.dot.gov/Main/index.aspx.* Accesssed: January 2010.

[7] U.S. Bureau of Transportation Statistics. May airline traffic: Five-month domestic traffic up 5.8 percent from 2004, *http://www.bts.gov/press_releases/2005/bts036_05/html/bts036_05.html#table_05*. Accessed January 2010.

[8] I. Pavlidis and J. Levine. Thermal image analysis for polygraph testing. *IEEE Engineering in Medicine and Biology Magazine*, 21(6):56 – 64, 2002.

[9] I. Pavlidis, J. Levine, and P. Baukol. Thermal imaging for anxiety detection. *Computer Vision Beyond the Visible Spectrum: Methods and Applications, 2000. Proceedings. IEEE Workshop on*, pages 104–109, 2000.

[10] D. A. Pollina, A. B. Dollins, S. M. Senter, T. E. Brown, I. Pavlidis, J. A. Levine, and A. H. Ryan. Facial skin surface tempurature changes during a "concealed information" test. *Annals of biomedical Engineering*, 34(7):1182 – 1189, 2006.

[11] C. Puri, L. Olson, I. Pavlidis, J. Levine, and J. Starren. Stresscam: non-contact measurement of users' emotional states through thermal imaging. In *CHI '05: CHI '05 extended abstracts on Human factors in computing systems*, pages 1725–1728, New York, NY, USA, 2005. ACM.

[12] T. A. Ranney. Driver distraction: A review of the current state-of-knowledge. Final Report DOT HS 810 787, National Highway Traffic Safety Administration, 2008.

[13] T. Schnell, T. Macuda, P. Poolman, and M. Keller. Workload assessment in flight using dense array eeg. *25th Digital Avionics Systems Conference, 2006 IEEE/AIAA*, pages 1–11, Oct. 2006.

[14] R. Sinha, W. R. Lovallo, and O. A. Parsons. Cardiovascular differentiation of emotions. *Psychosomatic Medicine*, 54:422 – 435, 1992.

[15] SPSS. *SPSS Neural Networks 17.0*, 2009. *http://support.spss.com/ProductsExt/SPSS/ESD/17/Download/User%20Manuals/English/SPSS%20Neural%20Network%2017.0.pdf*. Accessed March 2010.