

Die Pupillengröße als Index zur Online-Erfassung der kognitiven Belastung

Holger Schultheis

Universität des Saarlandes
D-66123, Saarbrücken, Deutschland
hosc5003@stud.uni-saarland.de

Zusammenfassung

In vielen Situationen könnte die Mensch-Computer Interaktion effizienter und für den Menschen angenehmer gestaltet werden, wenn es dem System möglich wäre, die kognitive Belastung des Subjekts zu berücksichtigen. In der Literatur ist zwar eine Fülle von Methoden zur Erfassung der mentalen Beanspruchung aufgeführt, eine Analyse der Eigenschaften der verschiedenen Maße zeigt jedoch, daß die wenigsten zur Erfassung der Belastung in konkreten Anwendungssituationen geeignet sind. Eine potentielle Ausnahme bildet die Messung der Pupillengröße, deren Zusammenhang mit der kognitiven Last durch 40 Jahre Forschung gut belegt ist. Allerdings gilt es zu überprüfen, ob die Ergebnisse aus den streng kontrollierten Laborsituationen auch auf alltägliche Anwendungen generalisieren. Zu diesem Zweck wurde ein Experiment entworfen mit dessen Hilfe das Pupillenmaß als Indikator kognitiver Belastung in einer alltäglichen Situation der Wissensakquisition (Lesen) validierbar ist.

1 Einleitung

Ein wesentlicher Aspekt des Zustands eines Benutzers ist seine kognitive Belastung. Klassischerweise wird kognitive Belastung als der Unterschied zwischen der grundsätzlichen Leistungsfähigkeit des menschlichen Informationsverarbeitungssystems und der Leistungsfähigkeit für eine aktuell zu bewältigende Aufgabe konzeptualisiert [Gopher and Donchin, 1986]. Mit anderen Worten ist kognitive Belastung ein hypothetisches Konstrukt, das die augenblicklich verwendeten bzw. benötigten mentalen Ressourcen repräsentiert. Damit ist die Belastung gleichzeitig ein Indikator für die freien Kapazitäten des Benutzers, d.h. wie viele und wie schwierige Aufgaben er bei der gegebenen Anforderungssituation zusätzlich bewältigen kann.

Ein System, das Kenntnis über die aktuelle kognitive Belastung des Benutzers hat, kann demnach Anforderungen an den Benutzer so verteilen, daß eine Überforderung - und somit auch Fehler - weitestgehend vermieden wird. Die Berücksichtigung der noch vorhandenen kognitiven Ressourcen ist also von großem Vorteil für verschiedenste Anwendungsbereiche (s. Abschnitt 2).

Berücksichtigung der kognitiven Belastung setzt allerdings voraus, daß eine reliable und valide Methode zur Erfassung derselben verfügbar ist. In Abschnitt 3 werden

verschiedene Maße vorgestellt und erläutert, warum sich der Pupillendurchmesser besser als alle anderen Indikatoren zur Online-Erfassung mentaler Anstrengung in Anwendungssituationen eignet.

2 Anwendungen

Wie schon einleitend erwähnt kann das Wissen über die kognitive Belastung des Systembenutzers dazu verwendet werden, Anforderungen an ihn optimal über die Zeit zu verteilen, um so Überforderung zu vermeiden. Letztendlich läuft dabei die Frage nach der optimalen Verteilung auf die Frage der Unterbrechbarkeit hinaus. Zu jedem Zeitpunkt steht das System vor dem Problem zu entscheiden, ob es den Benutzer mit einer oder mehreren neuen Anforderungen konfrontiert. Hierbei sollte diese Entscheidung im Idealfall sowohl von der Wichtigkeit der jeweiligen Aufgabe sowie dem aktuellen Belastungszustand geleitet sein.

Vor allem bei kritischen Echt-Zeit-Anwendungen, wie z.B. dem Steuern eines Fahrzeugs (Flugzeug, PKW, Schiff) oder der Überwachung von Prozessen (Flugraumüberwachung, Anzeigenüberwachung in einem Atomkraftwerk) ist eine optimale Verteilung unabdingbar. Sowohl eine Unterbrechung im falschen Moment, wenn dadurch wichtige Aufgaben nicht oder zumindest nicht korrekt zu Ende geführt werden können, als auch das Zurückhalten einer Unterbrechung, die dem Benutzer wesentliche Information vermittelt hätte, kann zu einer Katastrophe führen. Nach McFarlane und Latorella [1982] lassen sich eine Reihe von Flugunglücken zumindest teilweise auf ungenügende Unterbrechungsstrategien zurückführen.

In anderen Bereichen hat eine mangelhafte Unterbrechungsstrategie im Zweifelsfall nicht entsprechend verheerende Folgen, kann aber trotzdem zur ineffizienten, fehlerhaften und damit auch frustrierenden Aufgabenbearbeitung führen. Falsch applizierte Hilfen eines intelligenten Tutoriensystems mögen dazu führen, daß der Benutzer weniger lernt, als hätte er auf das System verzichtet und stattdessen klassischere Informationsquellen zur Wissensgewinnung verwendet. Auch im Büro am Bildschirmarbeitsplatz sind häufig genug Unterbrechungen der laufenden Arbeiten durch den Computer notwendig. Die Benachrichtigung über den Eingang einer EMail oder die an einem bestimmten Punkt der Verarbeitung eines autonomen Prozesses notwendige Entscheidung des Benutzers sind Beispiele für Situationen, in denen das System entscheiden muß, ob es den Benutzer unterbricht oder nicht.

Es sind mindestens drei verschiedene Dimensionen vorstellbar, auf denen die Unterbrechung des Benutzers durch das System variieren und eine Optimierung der Unterbrechungsstrategie möglich ist. Die erste ist die zeitliche

Dimension, also die Frage wann der Benutzer unterbrochen werden kann. Die zweite bezieht sich auf den Inhalt der Unterbrechung - was. Dies beinhaltet die Entscheidung des Systems, welche Anforderungen es selbst bewältigt und welche an den Benutzer weitergeleitet werden. Drittens kann eine zusätzliche Anpassung an den Benutzer und seinen Zustand durch die Art (wie-Dimension) der Informations-/Unterbrechungsdarbietung erreicht werden. Abhängig von der kognitiven Belastung des Benutzers kann das System die Unterbrechungen, dann geeignet auf den verschiedenen Dimensionen den noch freien Ressourcen anpassen.

Bisher lag die Betonung darauf, Überforderung des Benutzers zu vermeiden. Abschließend sei darauf hingewiesen, daß in einigen Situationen die Unterforderung ebenso problematisch sein kann. Im Zuge der Automatisierung komplexer Arbeitsumgebungen (wie z.B. ein Cockpit) sinkt die Normalfallanforderung an den Benutzer. Dies kann soweit führen, daß der Benutzer Langeweile empfindet, sich zu sehr auf das System verläßt und seine Aufmerksamkeit nachläßt. Als Folge dessen sind stille Fehler, Unzufriedenheit und eine generelle Abnahme der Leistung zu erwarten [Prinzel *et al.*, 2001]. Hier scheint es demnach sinnvoll, der Unterforderung durch eine Verringerung der Automatisierung, mit anderen Worten einer Erhöhung der Unterbrechungen, entgegenzuwirken (adaptive Automatisierung). Auch für diese Aufgabe ist ein gutes, praktisches Maß zur Online-Erfassung der kognitiven Belastung unumgänglich.

3 Die Pupillengröße als Indikator

Damit ein Maß als geeigneter Indikator für die mentale Beanspruchung gelten kann, sollte es mindestens drei Kriterien erfüllen. Zum Einen muß es interindividuelle Fähigkeitsunterschiede widerspiegeln. Zum Anderen sollte es für Schwierigkeitsunterschiede zwischen strukturell identischen Aufgaben sensitiv sein und drittens Schwierigkeitsunterschiede zwischen strukturell unterschiedlichen Aufgaben abbilden. In den vergangenen Jahrzehnten wurde eine Vielzahl von Maßen, die diese drei Kriterien erfüllen, vorgeschlagen. Dabei lassen sich drei grundsätzlich unterschiedliche Arten von Methoden identifizieren: physiologische, subjektive und Verhaltensmaße.

3.1 Verhaltensmaße

Letztere vergleichen das Verhalten bei der Bearbeitung einer Aufgabe mit dem Verhalten bei der Bearbeitung derselben und mindestens einer zusätzlichen Aufgabe. Durch den Unterschied in der Qualität der Zielerreichung in den verschiedenen Bedingungen sind Rückschlüsse über die für die Erstaufgabe benötigten Ressourcen möglich. Zur Online-Erfassung im Anwendungskontext ist diese Methode aus verschiedenen Gründen schlecht bzw. gar nicht geeignet. Einerseits wird zur Bestimmung der Belastung, die ja über das Verhalten erfolgt, eine beobachtbare Reaktion seitens des Benutzers benötigt. Je nach Anforderungskontext (z.B. bei Vigilanzaufgaben) sind solche Reaktionen jedoch äußerst selten, so daß über große Zeiträume keine Information über die aktuelle Belastung zur Verfügung stünde. Außerdem ist es bei konkreten Anwendungen in den seltensten Fälle praktikabel, oft sogar unmöglich, zu der eigentlichen Aufgabe Weitere einzuführen, da dies die Bewältigung der (wichtigen) Hauptaufgabe stört. Neben diesen praktischen Überlegungen gibt es aber auch theoriegeleitete Kritik an einem solchen Maß. Es ist nämlich

nicht klar, ob die durch die Hauptaufgabe und die Nebenaufgaben verursachte Belastungen additiv verknüpft sind oder interagieren. Daher ist der Rückschluß aus der Situation mit Mehrfachaufgaben auf die Einzelaufgabe nur eingeschränkt möglich.

3.2 Subjektive Maße

Eine weitere Möglichkeit zur Erfassung der kognitiven Beanspruchung besteht darin, den Benutzer direkt danach zu fragen wie stark ihn eine Aufgabe beansprucht hat - also ein subjektives Maß zu erheben. Allerdings ergeben sich auch mit dieser Methode Schwierigkeiten im gegebenen Kontext der Online-Erfassung. Es scheint nicht sinnvoll, den Benutzer während seiner Tätigkeit in kurzen Abständen nach einer Einschätzung seiner Belastung zu fragen, da dies in extremen Belastungssituationen zur Verschärfung des Problems und damit zum Gegenteil der angestrebten Regulierung führen würde. Eine Befragung nach Absolvierung der Aufgabe ist dagegen per se Offline und daher ohne Nutzen. Erschwerend kommt hinzu, daß Personen (a) dazu neigen nicht ihre eigene Beanspruchung, sondern eine objektive Schwierigkeit der Aufgabe anzugeben und, (b) sich nicht aller kognitiven Vorgänge bewußt sind, und dadurch zusätzlich mit Fehleinschätzungen zu rechnen ist.

3.3 Physiologische Maße

Die im Zusammenhang mit den subjektiven und Verhaltensmaßen auftretenden Probleme können durch die Verwendung physiologischer Maße umgangen werden. Physiologische Daten, d.h. Daten über körperliche Aspekte des Benutzers, sind kontinuierlich vorhanden und können ebenso kontinuierlich erfaßt werden. Insbesondere sind physiologische Methoden also nicht auf offene Reaktionen der Person angewiesen. Auch die Notwendigkeit zur Einführung zusätzlicher Aufgaben und die damit einhergehenden Probleme entfallen. Da die verwendeten Maße so gut wie gar nicht der bewußten Kontrolle der Person unterliegen, können auch keine Verzerrungen durch Top-Down Prozesse die Erfassung der kognitiven Belastung behindern. Von allen gängigen physiologischen Maßen wiederum (z.B. EEG, EKPs, Herzfrequenz und deren Variabilität, MEG, PET, Pupillendurchmesser, Hautleitwert, Atemfrequenz [Kramer, 1991; Wilson and Egge-meier, 1991]) ist der Pupillendurchmesser am besten zur anwendungsbezogenen Erfassung mentaler Belastung geeignet. Zum einen hat die Pupille eine sehr geringe Latenz, d.h. Veränderungen der Belastung führen in kürzester Zeit zu einer Veränderung der Pupillengröße. Beatty [1982] geht von 100 - 200 msec, Kramer [1991] von höchstens 600 msec aus. Demgegenüber wirken sich Belastungsveränderungen z.B. im Hautleitwert erst nach 1,4 - 2,5 sec und nach frühestens 30 sec. im PET aus [Kramer, 1991]. Außerdem ist der technische Aufwand zur Erfassung der Pupille im Vergleich z.B. zum EEG und EKPs relativ gering und - was besonders vorteilhaft für die praktische Anwendung ist - der Benutzer muß nicht über Kabel, Drähte und/oder Elektroden mit dem Meßgerät verbunden werden.

Zusammenfassend läßt sich also festhalten, daß die Pupille der beste Indikator zur Online-Erfassung von kognitiver Belastung ist, da:

- sie schnell auf Veränderungen anspricht.
- sie kontinuierlich erfaßt werden kann.
- keine ablenkenden zusätzlichen Aufgaben eingeführt werden müssen.

- die Messung auch ohne Verhalten des Benutzers möglich ist.
- die Messung nicht von der Person beeinflusst werden kann.
- keine Verkabelung für Messung notwendig ist.

4 Empirie

4.1 Bisherige Untersuchungen

Seit Mitte der 60er Jahre ist der Zusammenhang zwischen dem Pupillendurchmesser und der kognitiven Belastung empirisch fundiert (s. Rückblick von Beatty [1982]). Aber auch neuere Untersuchungen haben diese Kovariation repliziert und weitere Belege für ihre Validität geliefert [Dionisio *et al.*, 2001; Hyönä *et al.*, 1995]. Allerdings wurden alle diese Untersuchungen unter streng kontrollierten Bedingungen durchgeführt und es stellt sich die Frage, inwiefern sich die Ergebnisse auf alltäglichere Situationen übertragen lassen.

Wie alle physiologischen Maße ist auch der Pupillendurchmesser relativ stark verrauscht. Neben der kognitiven Belastung gibt es unzählige andere Faktoren, die den Pupillendurchmesser in einer gegebenen Situation determinieren [Krüger, 2000]. Besonders kritisch für die Erfassung der Belastung sind dabei jene Einflüsse, die sich ebenfalls kurzfristig und teilweise sogar deutlich stärker auf die Pupillengröße auswirken. Hierzu zählen unter anderem der Lichtreflex und der Hippus. Während die maximale bisher beobachtete Pupillenweite durch kognitive Belastung im Bereich von 0,1 - 0,9 mm liegt, können sich die durch den Lichtreflex hervorgerufenen Veränderungen im Millimeterbereich (bis zu 5 mm) bewegen. Der Hippus führt zwar nicht zu Variationen solcher Größe, tritt dafür aber scheinbar zufällig auf, ist rein endogen und damit nicht experimentell kontrollierbar.

Um solche Störeinflüsse zu eliminieren, wurden von allen bisherigen Studien spezielle Versuchsdesigns und Auswertungsverfahren eingesetzt. So wurde in allen Experimenten das visuell dargebotene Reizmaterial auf ein Minimum reduziert. Zusätzlich hatten die Versuchspersonen während der gesamten Experimentalphase einen bestimmten Punkt zu fixieren. Hierdurch konnten Licht- und Fixationseffekte im Pupillendurchmesser kontrolliert werden. Um auch den Einfluß des Hippus auf die Pupillenvariation zu kontrollieren, absolvierte jede Versuchsperson mehrere gleichartige und kurze (wenige Sekunden) Aufgaben, über die bei der Auswertung gemittelt wurde.

Offensichtlich steht ein solches Vorgehen im Widerspruch zum Einsatz des Pupillendurchmessers als Online-Indikator in konkreten Anwendungssituationen. So kann man in der Praxis nicht davon ausgehen, daß der Benutzer über einen längeren Zeitraum ein und denselben Punkt fixiert und seine Augen nicht bewegt. Ganz abgesehen davon, daß die Annahme einer Folge von gleichartigen, kurzen Aufgaben in den wenigsten Alltagssituationen plausibel ist, scheint eine Mittelung über viele Aufgaben hinweg zur Online-Erfassung nicht praktikabel. Aus diesem Grund wurde ein experimentelles Design entworfen, mit dem die Validität der Pupille in einer stärker dem Alltag angenäherten Situation überprüft werden kann.

4.2 Versuchsaufbau

In diesem Experiment haben die Versuchspersonen die Aufgabe, verschiedene Texte nacheinander zu lesen und nach jedem Text einige Verständnisfragen zu beantworten.

Somit stellt die Untersuchungssituation eine Annäherung an den schon in Abschnitt 3 erwähnten Büroalltag dar: Auf der Suche nach benötigter Information (den Antworten auf Fragen) betrachtet und liest der Benutzer verschiedene Dokumente - z.B. im Internet.

Die in der Studie verwendeten Texte unterscheiden sich in ihrem Schwierigkeitsgrad so, daß es zwei leichte und zwei schwere Texte zu lesen und zu verstehen gibt. Alle Texte sind etwa gleich lang (um die 1800 Zeichen). Hierdurch wird nicht nur die Abstrahlungsintensität der Texte auf dem Bildschirm kontrolliert, sondern auch ausgeschlossen, daß sich allein aufgrund der Textlänge Pupillenunterschiede bei der Bearbeitung der verschiedenen Leseaufgaben ergeben.

Ein Beispiel für den Anfang eines leichten Textes ist: *„Vor einigen Jahren fuhrn bei einem großen Hotel in Brüssel drei Herren in vornehmer Kleidung vor und gaben an, sie kämen aus Paris. Der eine gab sich als Diplomat aus, er habe einen Spezialauftrag bei der EG. Der zweite sagte, er sei Kommissar und müsse einige Tage hier bleiben, er habe nämlich Attentäter aufzuspüren, die einen Anschlag auf die Botschaft planten.“*. Einer der schwierigen Texte beginnt dagegen wie folgt: *„Ohne daß wir deshalb gerade dem Traume vor dem Wachen, dem Närrisch seyn vor der Besonnenheit einen Vorzug geben wollen, dürfen wir uns doch nicht läugnen: daß jene Abkürzungen- und Hieroglyphensprache, der Natur des Geistes in vieler Hinsicht angemessener erscheine, als unsre gewöhnliche Wortsprache.“*.

Der augenscheinliche Schwierigkeitsunterschied wird dabei durch die Quellen der Texte objektiviert. Beide einfachen Geschichten sind einem Schulbuch für die Jahrgangsstufen 5 und 6 entnommen, wohingegen die schwierigen aus einem Schulbuch für die Sekundarstufe II stammen. Zur weiteren Kontrolle, ob die Manipulation der Textschwierigkeit erfolgreich ist, wird jeder Proband nach der Lektüre eines Textes um Auskunft darüber gebeten, wie schwierig er den Text fand, und wie unangenehm bzw. belastend es für ihn gewesen wäre, beim Lesen unterbrochen zu werden. Außerdem wird mit einem speziellen Test die Lesespanne der Versuchspersonen bestimmt. Sie ist ein Maß für die Arbeitsgedächtnis- und Verarbeitungskapazität für textuelle Inhalte, und ist eventuell bezogen auf den Zusammenhang von Textschwierigkeit und Pupillenweite eine moderierende Variable: Im Zweifelsfall sind die Texte nur für Personen mit geringerer Lesespanne unterschiedlich schwierig, daher auch unterschiedlich belastend und führen nur für solche Probanden zu unterschiedlichen Pupillenweiten.

Weiterhin ist zu erwarten, daß die unterschiedliche Schwierigkeit der Texte nur dann einen Einfluß auf die kognitive Belastung hat, wenn die Personen sich bemühen den Inhalt zu verstehen. Um dies zu motivieren und zu kontrollieren, müssen die Versuchsteilnehmer nach jedem Text 4 Verständnisfragen in multiple-choice Form beantworten. Zur zusätzlichen Vermeidung ungewollter Effekte wird die Reihenfolge der Texte über die Versuchspersonen permutiert und als Teilnehmer nur Muttersprachler zugelassen.

Ist die Pupille in diesem experimentellen Aufbau ein geeigneter Indikator für die kognitive Belastung, so bietet sie sich auch zum Einsatz in konkreten Anwendungssituationen an, da:

- die Probanden im Versuch die Augen bewegen
- die einzelnen Aufgaben einer längerfristigen Bearbeitung bedürfen

- weder über die Versuchspersonen noch über einzelne Aufgaben hinweg die Pupillendaten gemittelt werden. (-> individuelle Online-Erfassung)

4.3 Stand der Arbeit

Die erste Phase der Datenerhebung steht kurz vor dem Abschluß. In der zweiten Phase, die voraussichtlich im Juli stattfinden wird, soll als Maß für die kognitive Belastung beim Lesen der Texte neben dem Pupillendurchmesser auch EEG bzw. EKPs verwendet werden. Durch dieses zweite Maß der mentalen Beanspruchung erhält man zusätzliche Information über die Validität des Pupillendurchmessers als Online-Indikator.

Im Rahmen des Workshopvortrags werde ich einige der Ergebnisse und Erkenntnisse, die aus den beiden Erhebungsphasen gewonnen wurden, vorstellen.

Literatur

- [Beatty, 1982] Jackson Beatty. Task-Evoked Pupillary Responses, Processing Load, and the Structure of Processing Resources. *Psychological Bulletin*, 91(2):276–292.
- [Dionisio *et al.*, 2001] Daphne P. Dionisio, Eric Granholm, William A. Hillix and William F. Perrine. Differentiation of deception using pupillary responses as an index of cognitive processing. *Psychophysiology*, 38:205–211.
- [Gopher and Donchin, 1986] Daniel Gopher and Emanuel Donchin. Workload: An Examination of the concept. In K. Boff, L. Kauffmann and J. Thomas (Eds.), *Handbook of perception and human performance* (vol. 2). New York: Wiley.
- [Hyönä *et al.*, 1995] Jukka Hyönä, Jorma Tommola and Anna-Mari Alaja. Pupil Dilation as a measure of Processing Load in Simultaneous Interpretation and Other Language Tasks. *The Quarterly Journal of Experimental Psychology*, 48A(3):598–612.
- [Kramer, 1991] Arthur F. Kramer. Physiological metrics of mental workload: A review of recent progress. In Damos, D. L. (Ed.), *Multiple-task performance*. London: Taylor and Francis.
- [Krüger, 2000] Frank Krüger *Coding of temporal relations in semantic memory: cognitive load and task-evoked pupillary response*. Münster: Waxmann.
- [McFarlane and Latorella, 1982] Daniel C. McFarlane, Kara A. Latorella. The Scope and Importance of Human Interruption in Human-Computer Interaction Design. *Human-Computer Interaction*, 17:1–61.
- [Prinzel *et al.*, 2001] Lawrence J. Prinzel, Alan T. Pope, Frederick G. Freeman, Mark W. Scerbo and Peter J. Mikulka. Empirical Analysis of EEG and ERPs for Psychophysiological Adaptive Task Allocation. *NASA/TM-2001-211016*.
- [Wilson and Eggemeier, 1991] Glenn F. Wilson and F. Thomas Eggemeier. Psychophysiological assessment of workload in multi-task environments. In Damos, D. L. (Ed.), *Multiple-task performance*. London: Taylor and Francis.