Paul Gerber, Daniel Ullrich

Automatisierte nichtinvasive Emotionsmessung. Ein Erfahrungsbericht von der Vision bis zur Realisierung

Eye Tracking_Emotion Tracking_Neural Networks_FaceLAB

Affektive Reaktionen spielen eine wichtige Rolle in der Usability- und User Experience Forschung und sind häufig Gegenstand in entsprechenden Messinstrumenten. Diese Messinstrumente haben aber oft den Nachteil, dass sie den Interaktionsprozess unterbrechen, etwa wenn sie während eines Usability-Tests durchgeführt werden. Oder aber sie unterliegen retrospektiven Verfälschungen, zum Beispiel wenn sie nach Abschluss der Interaktion durchgeführt werden. Ziel unserer Arbeit war die Entwicklung eines nichtinvasiven Messinstruments, das emotionale Reaktionen während einer Interaktion erfassbar macht, ohne dass der Interaktionsprozess unterbrochen werden müsste. In dem vorliegenden Beitrag skizzieren wir den Prozess der Entwicklung dieses Messinstruments, von der ursprünglichen Vision bis zur Realisierung. Dabei gehen wir auf theoretische wie praktische Hindernisse ein und beschreiben, wie diese die finale Lösung beeinflusst haben. Abschließend berichten wir Ergebnisse zur Genauigkeit aus der Pilotstudie und diskutieren Implikationen für zukünftige Arbeiten sowie für Praktiker, die solch ein System einsetzen möchten.

1. Einleitung

Die Beschreibung und Messung vom Emotionen und affektiven Reaktionen nimmt eine wichtige Rolle im Bereich der Produktevaluation sowie in der User Experience Forschung ein. Nicht nur ein effektiver und effizienter Umgang mit Produkten soll möglich sein – auch auf die Abwesenheit von negativem Affekt wie Ärger oder Frustration, weil das System nicht funktioniert wie man möchte, wird Wert gelegt. Idealerweise wird eine Interaktion von positiven Emotionen begleitet. In der Psychologie sind Emotionen schon seit langer Zeit ein zentraler Forschungsgegenstand. Zur Messung von Emotionen gibt es zahlreiche Ansätze, die auf verschiedenen Modellen basieren. Hierbei werden unter anderem Maße des autonomen Nervensystems, des Zentralnervensystems, Verhaltensmaße als auch Selbstberichte verwendet. Eine detaillierte Übersicht über Erhebungsmethoden und deren Validität hinsichtlich verschiedener Aspekte des emotionalen Zustandes liefern beispielsweise Mauss und Robinson (2009).

In diesem Artikel beschreiben wir zunächst kurz verschiedene Ansätze zur Emotionsmessung. Daraus leiten wir unsere Vorstellung von einem idealen Messinstrument zur Emotionsmessung ab. Dieses besteht aus einem Anteil der Datenaggregation und einem Anteil der Datenintegration, welche im dritten Abschnitt genauer beschrieben werden und zusammen das Gesamtsystem der Emotionsmessung bilden. Im vierten Abschnitt schildern wir unsere Erfahrungen, die wir während der Systementwicklung gesammelt haben und welchen zentralen Problemen wir begegnet sind. Im Anschluss wird kurz die mit dem System durchgeführte Pilotstudie beschrieben und Ergebnisse zur Vorhersagegüte des Systems berichtet. Abschließend diskutieren wir offene Forschungspunkte und welches Fazit wir aus der Entwicklung und dem Einsatz des Systems ziehen.

DOI 10.1524/icom.2012.0037

2. Ansätze zur Emotionsmessung

Um die emotionale Reaktion einer Person, die mit einem Produkt interagiert, messbar zu machen, werden in Usability-Tests verschiedene Messinstrumente eingesetzt. Diese gliedern sich grob in solche, die während der Interaktion eingesetzt werden und solche, die nach der Interaktion eine retrospektive Messung durchführen.

2.1 Abfrage von Emotionen während der Interaktion

Eine Messung während der Interaktion hat den Vorteil, dass tatsächlich am Punkt der Interaktion, wenn die Emotionen „frisch" sind, gemessen wird. Es findet also keine Verfälschung über die Zeit statt. Zudem besteht keine Gefahr, dass die emotionale Bewertung über den gesamten Interaktionszeitraum integriert wird. Ein Nachteil der Messung während der Interaktion ist jedoch, dass die Versuchsteilnehmer neben ihrer eigentlichen Aufgabe – der Interaktion – von der Messung abgelenkt werden. Wird beispielsweise die Methode des lauten Denkens angewandt, um durch die Äußerungen des Versuchsteilnehmers auf dessen emotionale Reaktion zu schließen, so müssen während der Interaktion zwei Aufgaben gleichzeitig erfüllt werden: Die Interaktion selbst und eine stetige Reflektion und Artikulation des Erlebten. Auch wenn es niederschwelligere Methoden gibt (beispielsweise die Valenzmethode

nach Burmester und Kollegen, 2010), so bleibt das prinzipielle Problem bestehen, dass die Versuchsteilnehmer ihre Aufmerksamkeit nicht völlig auf den Interaktionsgegenstand bündeln können. Die gravierendste Störung stellt wohl die komplette Unterbrechung zwecks Ausfüllens eines Fragebogens dar, da hierbei die Versuchsteilnehmer völlig aus ihrer Interaktion gerissen werden.

2.2 Abfrage von Emotionen nach der Interaktion

Die Messung von Emotionen nach abgeschlossener Interaktion bietet den großen Vorteil, dass die Interaktion ohne Störungen oder Unterbrechungen ablaufen kann. Somit können sich Versuchsteilnehmer ganz auf das Produkt konzentrieren, was einem natürlichen Nutzungsszenario am nächsten kommt. Problematisch bei diesen Methoden ist, dass Versuchsteilnehmer ihren Gesamteindruck in Form eines holistischen Urteils widergeben und der situationsspezifische Eindruck durch den zur Abfrage aktuellen Eindruck verzerrt wird (Levine & Safer, 2002). Dies macht es problematisch, die Messung auf einen bestimmten Teil der Interaktion zu beziehen.

Trotzdem bieten diese Methoden einen schnellen Zugang zur emotionalen Befindlichkeit und sind relativ leicht einzusetzen. Des Weiteren bietet die Vielfalt an bestehenden Methoden eine gute Anpassung an das eigene Untersuchungssetting: Fragebogenbasierte Methoden wie das Emotion Sampling Device (Roseman und Kollegen, 1996) oder die Differential Emotions Scale (Izard und Kollegen, 1974) fragen den Nutzer nach bestimmten vorkategorisierten Emotionen und wie häufig sie diese während der Interaktion erlebt haben. EMO2 (Laurans & Desmet, 2006) ist eine videobasierte Methode, bei der die Anwender bei der Interaktion gefilmt werden. Nach Abschluss bekommen sie das Video gezeigt und sollen die Gefühle beschreiben, die sie während der Interaktion hatten. Darüber hinaus gibt es diverse sprachfreie Methoden wie den SAM (Self Assessment Manikin; Bradley & Lang, 1994) oder auch Emocards (Desmet und Kollegen, 2001), die ohne Einsatz von Sprache auskommen. Hier sollen Nutzer ihre Gefühle anhand von Bildern ausdrücken, indem sie aus einer Auswahl das Bild wählen, das ihrem Gefühlszustand am nächsten kommt. Ein Merkmal der sprachfreien Methoden ist, dass sie kulturübergreifend oder auch bei Kindern angewendet werden können, da bei ihnen keine Sprachbarriere existiert.

2.3 Das Beste aus beiden Ansätzen: Nichtinvasive Emotionsmessung während der Interaktion

Da die Messung während und nach der Interaktion jeweils Vor- und Nachteile als Konsequenz haben, war das Ziel unserer Forschung die Erstellung eines Messinstruments, das die Vorteile aus beiden Ansätzen verbindet. Einerseits sollte die Messung während der Interaktion ablaufen, so dass eine direkte Zuordnung emotionaler Reaktionen auf die Interaktionsinhalte möglich war. Somit wären Usability-Probleme genauso wie besonders gelungene Interaktionen mit positiven affektiven Reaktionen leicht identifizierbar. Andererseits sollte die Messung nichtinvasiv erfolgen, d.h. ohne Zutun der Versuchsteilnehmer und ohne Ablenkung von der Interaktion durch die Messung.

3. Die Vision – Emotionen messen in Echtzeit

Um Emotionen nichtinvasiv und während der Interaktion zu messen, gibt es bereits Ansätze: Experten führen hierzu (Video-) Analysen durch und bewerten den mimischen Ausdruck der Versuchsteilnehmer. Als Bewertungsgrundlage dienen hierbei Rating-Systeme wie beispielsweise das „Facial Action Coding System" (FACS) nach Ekman & Friesen (1978). Dieses System bietet zwar nicht unmittelbar die Messung von Emotionen, aber die eindeutige Kodierung aller dem Menschen möglichen Mimikausdrücke auf Basis der Muskelbewegungen und daraus resultierenden Verschiebungen statischer (z.B. Mundwinkel) und temporaler (z.B. Stirnfalten) Gesichtsmerkmale. Auf Basis dieser Arbeit wurde daraufhin von verschiedenen Autoren versucht, Zusammenhänge zwischen emotionalen Zuständen und der Mimik zu dokumentieren. Hierbei zeigten sich beispielsweise starke Zusammenhänge zwischen dem Gesichtsausdruck und der Valenz-Dimension des emotionalen Zustandes (z.B. Mauss, Levenson, McCarter, Wilhelm & Gross, 2005). Nach unserer Vorstellung sollte unser System die während der Interaktion mit einem Produkt auftretenden mimischen Reaktionen des Nutzers messen und die so gewonnenen Daten in Anlehnung an das FACS interpretieren. Im Idealfall würde das System die Ausprägung verschiedener Emotionen in Echtzeit rückmelden, um so – optional – der Versuchsleitung die Gelegenheit zu geben, direkt auf besonders positive oder negative Emotionen reagieren zu können.Für die Erfassung der Mimik befanden wir ein kamerabasiertes System (im Gegensatz zu etwa Mimikerfassung mittels EMG via Elektroden) für angemessen, um den nichtinvasiven Charakter zu bewahren.

Eine Möglichkeit hierfür schien das System FaceLAB zu sein, ein modulares System welches ein Modul für Eye-Tracking, für Face-Tracking sowie für das Aufzeichnen physiologischer Variablen enthält. FaceLAB ermöglicht also neben der Erfassung des Blickwinkels einer Person auch die Erfassung und Aufzeichnung des Gesichtsausdrucks, beziehungsweise verschiedener markanter Punkte des Gesichtes, in kartesischen Koordinaten. Diese Daten sollen als Basis für die Vorhersage des aktuellen emotionalen Zustandes des Probanden zu einem beliebigen Zeitpunkt während der Erhebung dienen. Hierbei sollte auf Basis von im FACS-Manual (Ekman, Friesen & Hager, 2002) dokumentierten, mit Emotionen assoziierten Gesichtsausdrücken ein Kriterienkatalog entwickelt werden, wie sich die verschiedenen durch FaceLAB erfassten Mimikpunkte verändern müssen, um die verschiedenen FACS-Codes abzubilden. Diese Prototyp-Muster wiederrum sollten dann als Vergleich für den jeweils aktuellen Gesichtsausdruck genutzt werden. Für den Vergleich sollte dabei ein im Verlauf der Arbeit zu trainierendes neuronales Netzwerk genutzt werden. Es soll also die Datenaggregation, und nicht nur die Integration dieser, ebenfalls computergestützt umgesetzt werden, sodass diese deutlich zeiteffizienter zu realisieren ist. Als zu bewertende Emotionen legten wir uns auf die Basisemotionen nach Ekman und Friesen (1971) fest: Angst, Ärger, Freude, Trauer, Überraschung und Ekel.

Zusammenfassend war unser Ziel die Erstellung eines Systems, dass beispielsweise während Produktevaluationen parallel zur Interaktion mitläuft und die Reaktionen des Nutzers hinsichtlich seiner Primärreaktionen misst. Hierdurch sollte es möglich sein, für jeden Zeitpunkt der Interaktion genau feststellen zu können, wie sich der Nutzer fühlte und was ihm bei der Interaktion Spaß machte und wo er Frustrationen erlebte.

4. Die Realität – Anpassung unseres Systems an die FaceLAB-Technik

Im Zuge der Realisierung unserer Vision des Messinstruments zeigten sich Begrenzungen durch die FaceLAB-Technik, die teilweise Abstriche am Messinstrument nötig machten. Im Folgenden beschreiben wir einige dieser Problembereiche und deren Konsequenzen für unser System.

Der zentrale Kern unseres Systems sollte die Erfassung von Emotionen anhand mimischer Ausdrücke sein. Um eine möglichst gute Bewertung des emotionalen Zustands zu erreichen, ist die Erfassung einiger charakteristischer Punkte im Gesicht notwendig. FaceLAB bietet hierzu standardmäßig zahlreiche sogenannte „Facial Landmarks", also spezifische Punkte im Gesicht, wie beispielsweise die Mundaußenwinkel oder die Position der Augenbrauen. Zusätzlich zu diesen Facial Landmarks ist es in FaceLAB möglich, für die Erstellung des „Head-Models" (das Koordinatenabbild des Kopfes des aktuell getrackten Nutzers) personenspezifische Punkte zu definieren. Die Facial Landmarks sowie die selbst spezifizierten Punkte sollten eine ausreichende Basis für eine umfassende Erfassung der Mimik einer Person darstellen (siehe Bild 1a). Bei der Auswertung der Logfiles mussten wir jedoch feststellen, dass zahlreiche Facial Landmarks nicht getrackt werden. So fehlen beispielsweise die gesamte Gesichtskontur sowie eine Daten über Position und Kontur einer Brille (sofern vorhanden). Auf Nachfrage beim Hersteller stellte sich heraus, dass es im Augenblick keine Möglichkeit gibt, an diese Daten heranzukommen, obwohl sie im Manual

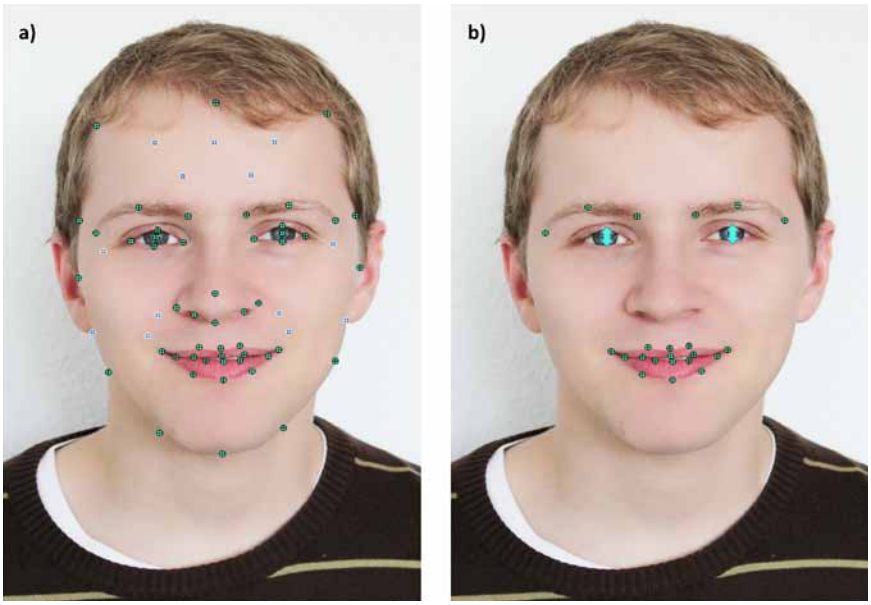

Bild 1: In der Planung vorgesehene (a) und tatsächlich verwendete (b) Facial Landmarks

erwähnt werden. Ähnlich verhielt es sich mit den personenspezifischen Punkten, da auch für diese keine Möglichkeit zur Speicherung vorhanden war.

Auf Grund der fehlenden individualisierten Trackingpunkte, waren wir gezwungen, von unserer ursprünglichen Idee einer automatisierten Reproduktion des Ratingvorgangs beim FACS abzurücken und mit einer vereinfachten Form zu arbeiten. Im Rahmen unserer Pilotstudie entwickelten wir ein vereinfachtes Kodierungssystem, welches die mit FaceLAB erfassbaren Mimikveränderungen erfasst. Die hierdurch abgeleitete Beschreibung der Mimik sollte ebenfalls auf Zusammenhänge mit dem vom Probanden geäußerten emotionalen Zustand untersucht und in ihrer Nützlichkeit zur Vorhersage dieses mit der Leistung eines neuronalen Netzwerkes, welches mit den von FaceLAB gelieferten Rohdaten arbeitet, vergleichen werden. Für eine detaillierte Vorstellung des Vorgehens bei der Erstellung dieses Ratingverfahren siehe auch Gerber (2012). Bild 1b zeigt im Vergleich zu Bild 1a, welche Punkte der menschlichen Mimik von FaceLAB zuverlässig verarbeitet werden können und somit in unsere Arbeit eingeflossen sind. Es fällt auf, dass nicht nur die oben angesprochenen Punkte der Gesichtskontur sowie Brille herausfallen, sondern auch jene, welche im Manual ein mögliches Tracking der Nase sowie der Augenlider implizieren. Während unserer Analyse der generierten Logfiles fiel uns bei diesen Punkten auf, dass sie keine Variablen darstellen, sondern nur statische Werte repräsentieren. Diese werden augenscheinlich von FaceLAB als Referenzpunkte für das Tracking genutzt. Im Falle der Facial Landmarks, die die Augen repräsentieren, scheinen die Datenpunkte nicht die Position der Lider, sondern nur die der Augäpfel darzustellen. Da alle Werte im Face-Log relativ zu einem Koordinatensystem des Kopfes angegeben werden, welches zwischen beiden Augäpfeln seinen Ursprung hat, verändert sich deren Position natürlich nicht. Auf Nachfrage des Herstellers erhielte wir den Tipp, dass im sogenannten Eye-Log, ein weiteres der insgesamt fünf von FaceLAB erstellten Logfiles, welches u.a. Daten zur Pupillometrie enthält, eine Variable existiert, welche die durch die Augenlider verdeckte Fläche des Augapfels in Prozent repräsentiert. Diese nutzten wir als Maß für die Öffnung der Augen.

Um die emotionale Reaktion der Versuchsteilnehmer zu dokumentieren, war es uns wichtig, den Videostream der Kameras aufzuzeichnen. Leider ist es in keiner bisher erschienenen Softwareversion von FaceLAB (in der vorliegenden Arbeit wurde Version 5.0.4 genutzt) möglich, die Videodaten der Kameras auch zu speichern. Um an den Stream direkt heranzukommen, müsste also zunächst eine eigene Lösung entwickelt werden. Ein erster Workaround ist das Aufzeichnen des Bildschirminhaltes des FaceLAB-PCs während der Erhebung. Dieser hat jedoch

zwei gravierende Nachteile. Zum einen, stellt die Aufzeichnung eine nicht zu unterschätzende Zusatzbelastung für den PC dar, welche unter Umständen dazu führen kann, dass die Trackingperformance leidet. Zum anderen war es uns während unserer Arbeit nicht möglich, eine bestehende Softwarelösung zu finden, welche mehr als 30 Frames pro Sekunde aufzeichnet. Hierdurch kommt es dazu, dass nicht für jedes Frame im Logfile ein entsprechendes Frame im Video vorhanden ist. FaceLAB hat eine zeitliche Auflösung von 60 Frames pro Sekunde, die wir auch nicht aufgeben wollten, da wir auch im Falle von Tracking-Aussetzern immer eine ausreichende Zahl auswertbarer Frames verfügbar haben wollten. Ein weiteres Problem bei der Videoaufnahme war die Inkonsistenz der Aufnahme. So konnte es passieren, dass 10 Logfileframes lang kein neues Bild aufgenommen wurde, danach aber für einige Zehntelsekunden für jedes Frame eine Aufnahme entsteht. Dieses Problem konnten wir auf eine erhöhte Ressourcenauslastung des FaceLAB-PCs zurückführen. Eine Portierung der Software auf ein leistungsstärkeres System war aus Gründen des Herstellersupports nicht möglich.

Ein weiterer Stolperstein, der ebenfalls der hohen zeitlichen Auflösung von FaceLAB geschuldet ist, sind die bei Nutzung von FaceLAB entstehenden Datenmengen. FaceLAB zeichnet für alle erfassten Variablen 60 Datensätze pro Sekunde auf, was dazu führt, das in einer 30-minütigen Erhebung bereits ca. 100.000 Datensätze entstehen. Eine komplette Versuchsreihe mit mehreren Teilnehmern resultiert so schnell in einigen Millionen Datensätzen bzw. einigen Gigabyte an Daten. Da die Auswertung und Interpretation bei solchen Systemen typischerweise komplett dem Anwender überlassen bleibt, liegt hier eine weitere Hürde. Für viele Standardprogramme (MS Excel, Access oder auch SPSS) sind die angefallenen Datenmengen schlicht zu groß, was Spezialprogramme oder Eigenlösungen erforderlich macht. Unsere Lösung bestand aus einer Datenbanklösung mit MS Access, in der wir für jeden Versuchsteilnehmer eine eigene Auswertung durchführten, um unter der Restriktion für SQL-Abfragen von maximal 2 GB zu bleiben. Die Einzelauswertungen wurden später in SPSS zusammengeführt und deskriptiv und inferenzstatistisch ausgewertet. Für einen routinemäßigen Einsatz des Systems ist diese Form der Auswertung jedoch sehr umständlich, da sie einiges an „Handarbeit" und Detailkenntnisse des Systems verlangt. Hier ist ein größerer Grad an Automation wünschenswert, um den Einsatz effizienter zu gestalten.

5. Unser System im Einsatz: Die Pilotstudie

Unser realisiertes Messinstrument wurde im Laufe der Realisierung den Möglichkeiten angepasst. Dabei musste auf einige Features wie beispielsweise die exakte Replikation des FACS verzichtet werden. Da es sich hierbei nur um eine Pilotstudie zur Abklärung der generellen Machbarkeit handelte, wurde auf die Umsetzung weniger harter Anforderungen, wie die On-the-fly-Auswertung, verzichtet. Der im vorigen Kapitel beschriebenen Hürden zum Trotz gelang es uns, ein System umzusetzen, das nichtinvasiv und kontinuierlich die emotionalen Reaktionen von Nutzern misst. Die zentralen mimischen Kennwerte, auf denen die Auswertung basiert, sind die Koordinaten der Mundinnen- und -außenkontur, der Augenbrauen, den Verschluss der Augen durch die Lider sowie den Durchmesser der Pupille. Die gemessenen Koordinaten werden jeweils in Beziehung mit einem neutralen Referenzbild (das zu Beginn der Untersuchung aufgenommen wird) gesetzt und die Differenzen ergeben somit Rückschluss auf den emotionalen Zustand des Nutzers. Die „Interpretation" übernimmt ein neuronales Netzwerk, das im Zuge der ersten Testerhebungen trainiert wurde und vielversprechende Ergebnisse lieferte.

In einer Pilotstudie (N=37) wurde die Leistungsfähigkeit unseres Systems zur Emotionsmessung geprüft. Hierzu wurden die von FaceLAB gelieferten Rohdaten, welche durch ein neuronales Netzwerk hinsichtlich des emotionalen Zustandes interpretiert wurden mit den Selbstauskünften der Probanden sowie mit Fremdbewertungen unbeteiligter Personen (naiver Rater) verglichen. Jedem Versuchsteilnehmer wurde jeweils die gleiche Art und Anzahl verschiedener Kurzfilme präsentiert, die jeweils spezifische emotionale Reaktionen (z.B. Freude, Ärger, etc.) auslösen sollten. Die Assoziation der Filme mit spezifischen Emotionen wurde in einer Vorstudie validiert. Parallel zur Präsentation der Kurzfilme ließen wir FaceLAB mitlaufen, um die Veränderungen der Mimik bei den Probanden aufzuzeichnen. Bevor die Wiedergabe des ersten Films begann, markierten wir für jeden Teilnehmer ein Referenzbild, welches zum Vergleich mit Aufnahmen während der Wiedergabe der verschiedenen Filme genutzt wurde. Danach startete der eigentliche Versuch. Zeigte sich eine sichtbare Veränderung in der Mimik, pausierten wir kurz die Wiedergabe, und baten den Probanden, eine Selbstbewertung ihrer aktuellen Befindlichkeit hinsichtlich der Basisemotionen (Angst, Ärger, Freude, Trauer, Überraschung und Ekel) vorzunehmen. Die Skala, auf der alle Teilnehmer ihren jeweiligen emotionalen Zustand beschreiben sollten reichte in stetiger Form von 0 (beispielsweise „gar nicht ängstlich") bis 7 („sehr ängstlich").

Diese Zeitpunkte wurden dann aus den Gesamtlogfiles extrahiert und zusammen mit den Fragebogendaten zum Training eines neuronalen Netzwerkes genutzt. Hierbei nutzten wir ca. 50% der Daten als Trainingsstichprobe, 20% als Teststichprobe sowie 30% als Holdout-Stichprobe. Die Daten der Holdout-Stichprobe werden nicht für das Training des Netzwerkes genutzt, sondern dienen nach Abschluss des Trainings der Validierung der Gesamtleistung des Netzwerkes, indem diese Datensätze dem Netzwerk vorgelegt und die Vorhersagen mit den entsprechenden Fragebogenwerten verglichen werden. Hierdurch wird sichergestellt, dass sich das Netzwerk nicht „selbst vorhersagt", wenn man die Güte des Netzwerks ermitteln möchte. Alle Angaben zur Vorhersagegüte beziehen sich auf die Holdout-Stichprobe.

Trotz der beschriebenen Beschränkungen von FaceLAB erreichte das Netzwerk eine zufriedenstellende Vorhersagegenauigkeit. Das Netzwerk erreichte hierbei in der Vorhersage der Intensität der verschiedenen Emotionen einen mittleren Fehler von 0,472 Skalenpunkten. Hierbei war Überraschung mit einem mittleren Fehler von 0,546 Punkten am schwersten in der Intensität vorherzusagen, Trauer mit 0,382 Punkten am einfachsten. Hinsichtlich der primären Emotion, also der, die am stärksten in diesem Moment ausgeprägt ist, erreichte

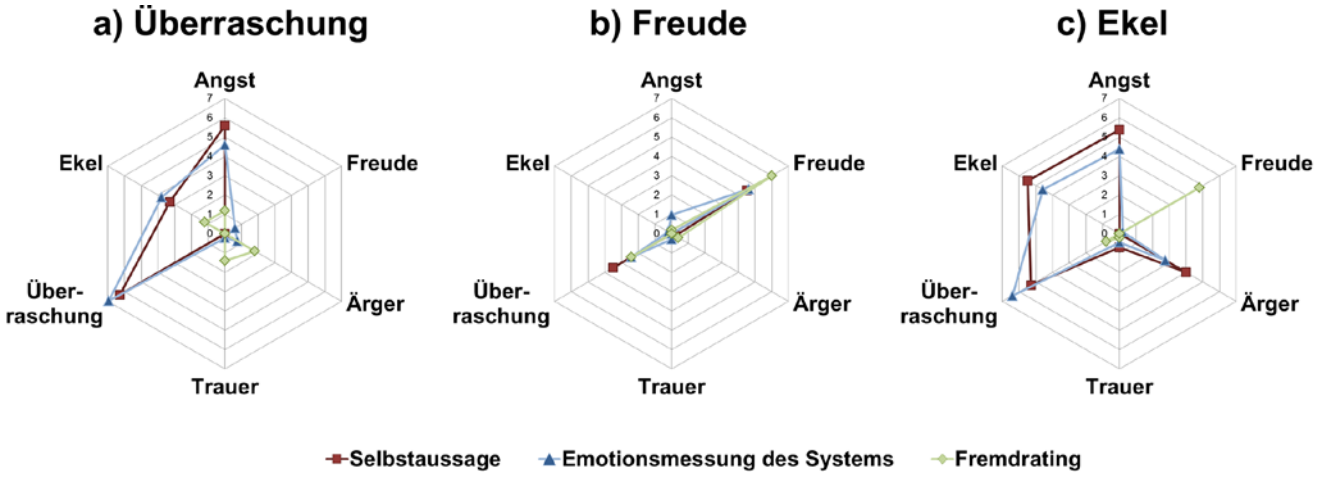

Bild 2: Vergleich von Selbstauskunft, Systemvorhersage und Fremdrating für die durch die Manipulation intendierten Primäremotionen Überraschung (a) Freude (b) und Ekel (c)

das Netzwerk auf Basis dieser Daten eine Kategorisierungsleistung von im Mittel 82% korrekt eingeordneten Emotionen. Das heißt, dass in 82% der Fälle, über alle Emotionen hinweg, das von uns trainierte Netzwerk auf Basis der von FaceLAB gelieferten Daten die Emotion als primäre Emotion vorhersagte, welche auch in der Selbstaussage der Probanden als die primäre Emotion deklariert wurde. Hierbei sind die Werte je nach Emotion aber sehr verschieden. So konnte die Emotion Angst nur sehr schlecht kategorisiert werden, wohingegen Freude, Trauer, Überraschung und Ekel mit 83% – 94% deutlich besser erkannt wurden. Dies lag augenscheinlich vor allem in der Tatsache begründet, dass es für die Emotion Angst, nur sehr wenig Trainingsmaterial gab, unser Stimulusmaterial also nicht hinreichend gut geeignet war, bei unseren Probanden Angst zu evozieren. Für eine detaillierte Beschreibung unseres Vorgehens und der Ergebnisse vergleiche auch Gerber (2012). Die Gesichtsausdrücke, auf denen die Systembewertung erfolgte, wurden per Screenshot gespeichert. Diese legten wir später Ratern vor, welche sie ebenfalls hinsichtlich der sechs Basisemotionen bewerten sollten. Abbildung 2 zeigt beispielhaft für drei verschiedene Probanden und drei verschiedene Primäremotionen die Selbstaussage zum emotionalen Zustand, die Vorhersage der Rater sowie die Vorhersage durch das neuronale Netzwerk auf Basis der von FaceLAB aufgezeichneten Daten.

Im Vergleich zur Leistung des neuronalen Netzes fiel die Leistung der Rater signifikant schwächer aus. Die Korrelationen zwischen der Vorhersage der Emotionsintensität durch die naiven Rater und der Selbstaussage der Probanden lag in jedem Fall signifikant unter dem Zusammenhang zwischen der Selbstaussage der Probanden und der Vorhersage durch unser Netzwerk ($z < -4{,}207$; $p < 0{,}000$). Die Rater erreichten im Mittel bei der Intensitätsbewertung einen Fehler von 1,365 Skalenpunkten, wobei hier die Emotion Angst mit 1,097 Fehlerpunkten am besten, die Emotion Überraschung mit 1,647 Punkten am schlechtesten vorhergesagt wurde. Hinsichtlich der Kategorisierungsleistung gab es bei den Ratern ebenfalls große emotionsspezifische Unterschiede. So wurde Freude mit 60% korrekt klassifizierten Fällen am besten erkannt, Ärger und Überraschung jedoch mit 0% bzw. 15,38% korrekt deutlich schlechter. Im Mittel über alle Emotionen lagen die Rater in 41,67% mit ihrer Einschätzung hinsichtlich der primären Emotion richtig, was in etwa der Hälfte der vom neuronalen Netzwerk korrekt klassifizierten Fälle entspricht.

6. Fazit und Implikationen

Die Idee eines einfach zu benutzenden und dennoch validen „Emotionstrackers" klingt verlockend. Im Laufe unserer Arbeit zur Umsetzung des Systems begegneten wir einigen Problemen, die zu einigen Einschränkungen führten. So mussten wir von der Idee der Live-Auswertung Abstand nehmen, da nicht genug technische Ressourcen auf dem FaceLAB-System zur Verfügung standen, um die anfallenden Datenmengen während des Versuchs auszuwerten. Die andere große Einschränkung betrifft die verwendeten Facial Landmarks. Hier waren weit weniger verfügbar, als vorab von uns angenommen. Dennoch erzielt unser System eine gute Erkennungsleistung der vorherrschenden Emotion und ist menschlichen Ratern überlegen. Ein Vergleich mit speziell auf Emotionserkennung geschulten Ratern steht noch aus.

Dennoch steht unser System, auch im Vergleich zu ähnlichen bestehenden Softwarelösungen, die mit selbstgedrehten Filmen als Datenquelle arbeiten, gut da. Letztere leiden häufig unter Restriktionen hinsichtlich Bildqualität und Beleuchtung. So konnten Kim, Kolbe und Eisele (2012) in einem ersten Praxistest der Software FaceReader3 nur fünf von 30 Aufnahmen auswerten. Unser System, das FaceLAB zur Mimikerkennung nutzt, zeigte sich hier deutlich robuster, da es selbst für die nötige Ausleuchtung sorgt und einzig starke Verschattungen durch direkte Sonneneinstrahlung die Qualität des Trackings nennenswert beeinflussen.

Auch wenn die Technik zwar eine objektive Messung bietet und somit eine Überlegenheit gegenüber beispielsweise Fragebogenerhebungen suggeriert, bleibt auf der anderen Seite das Problem der Konstruktvalidität bestehen: Emotionen spiegeln sich zwar in der Mimik wider, aber nicht ausschließlich. So ist unser System eine indirekte Messung, die als Prämisse enthält, dass von der Mimik hinreichend auf die Emotionen der

Nutzer geschlossen werden kann. Die Varianz in den emotionalen Zuständen der Nutzer, die sich nicht in der Mimik niederschlagen, können von dem System nicht erfasst werden. Ebenso besteht das Problem, dass die Mimik auch bewusst beherrscht und kontrolliert werden kann, was die Ergebnisse verfälschen könnte. Die Untersuchung der Korrelationen zu anderen Messinstrumenten sollte Gegenstand von zukünftigen Untersuchungen sein.

Unser System stellt nun einen vielversprechenden Ansatz für eine nichtinvasive Emotionsmessung dar, gleichzeitig gibt es einige Punkte für Verbesserungen. Die Stichprobe für das Training des neuronalen Netzes war mit 37 Personen noch recht klein. Hier könnte eine größere und heterogenere Stichprobe Potential für Verbesserungen bieten, um die Vielfalt der menschlichen Mimik umfassend abzudecken. Auch die Implementation der Live-Auswertung ist ein lohnenswertes Ziel, da diese neue Anwendungen wie direktes Reagieren von Systemen auf emotionale Zustände der Nutzer ermöglichen würde. Ein Tracking-System, dass weitere Facial-Landmarks auswerten könnte, wäre ebenfalls wünschenswert, auch wenn mit der eingeschränkten Zahl bereits sehr gute Vorhersageleistungen möglich sind.

Aber auch schon jetzt ist die Leistung, die diese Technik aktuell bereits bietet, beeindruckend und verspricht in Zukunft einen leichteren und anwendungsorientierten Zugang zu dieser spannenden Datenquelle, die nicht nur in der UX-Forschung und Produktevaluation, sondern auch für praktische Anwendungen von Nutzen sein kann.

Literatur

Bradley, M. M., & Lang, P. J. (1994). Measuring emotion: The self-assessment manikin and the semantic differential. Journal of Behavioral Therapy and Experimental Psychiatry, 25, 49–59.

Burmester, M. Jäger, K., Mast, M., Peissner, M. & Sproll, S. (2010). Design verstehen – Formative Evaluation der User Experience. In: H. Brau, S. Diefenbach, K. Göring, M. Peissner & K. Petrovic (Hrsg.), Usability Professionals 2010 (S. 206–214). Stuttgart: Fraunhofer.

Desmet, P.M.A., Overbeeke, C.J. & Tax, S. J. E. T. (2001). Designing Products with Added Emotional Value: Development and Application of an Approach for Research through Design. The Design Journal, 4(1), 32–47.

Gerber, P. (2012). Computergestützte Erfassung und Auswertung spontaner, emotionaler Mimikausdrücke. Technische Universität Darmstadt.

Ekman, P., & Friesen,W. V. (1971). Constants across cultures in the face and emotion. Journal of Personality and Social Psychology, 17 (2), 124–129.

Ekman, P., & Friesen,W. V. (1978). Facial action coding system: A technique for the measurement of facial movement. Palo Alto, CA: Consulting Psychologists Press.

Ekman, P., Friesen, W. V., & Hager, J. C. (2002). Facial Action Coding System - The Manual on CD-ROM. Salt Lake City: Research Nexus division of Network Information Research Corporation.

Izard, C. E., Dougherty, F. E., Bloxom, B. M., & Kotsch, N. E. (1974). The Differential Emotions Scale: A method of measuring the meaning of subjective experience of discrete emotions. Nashville: Vanderbilt University, Department of Psychology.

Kim, K., Kolbe, K., & Eisele, S. (2012). Es steht Dir ins Gesicht geschrieben! i-com, 11(1), 63–67. doi:10.1524/icom.2012.0016.

Laurans, G., & Desmet, P.M.A. (2006). Using self-confrontation to study user experience: A new approach to the dynamic measurement of emotions while interacting with products. In: P.M.A. Desmet, M.A. Karlsson, and J. van Erp (Eds.), Design & Emotion 2006; Proceedings of The International Conference on Design and Emotion, September 27–29. Gothenburg, Sweden: Chalmers University of Technology.

Levine, L. J., & Safer, M. a. (2002). Sources of Bias in Memory for Emotions. Current Directions in Psychological Science, 11(5), 169–173. doi:10.1111/1467-8721.00193

Mauss, I. B., Levenson, R. W., McCarter, L., Wilhelm, F. H., & Gross, J. J. (2005). The tie that binds? Coherence among emotion experience, behavior, and physiology. Emotion, 5 (2), 175–190.

Mauss, I. B., & Robinson, M. D. (2009). Measures of emotion: A review. Cognition & emotion, 23(2), 209–237. doi:10.1080/02699930802204677.

Roseman, I.J., Antoniou, A.A., Jose, P.E. Appraisal determinants of emotions: constructing a more accurate and comprehensive theory, Cognition and Emotion 10:3, (1996), 241–277.

Paul Gerber

ist wissenschaftlicher Mitarbeiter der Forschungsgruppe Arbeits- und Ingenieurpsychologie der Technischen Universität Darmstadt. Seine Tätigkeitsschwerpunkte liegen in der Produktevaluation mittels Eye- und Emotiontracking im Kontext der User Experience sowie der Produktentwicklung. Paul Gerber hat an der Technischen Universität Darmstadt Psychologie mit dem Nebenfach interdisziplinäre Produktentwicklung studiert.
E-Mail:
gerber@psychologie.tu-darmstadt.de

Daniel Ulrich

ist wissenschaftlicher Mitarbeiter im Bereich Arbeits- und Ingenieurpsychologie an der Technischen Universität Darmstadt. Seine Forschungsinteressen liegen im Bereich User Experience, hier der Wahrnehmung intuitiver Interaktion und deren Komponenten sowie in der intuitiven Entscheidungsforschung. Daniel Ullrich hat an der Technischen Universität Darmstadt Psychologie mit Nebenfach Informatik studiert.
E-Mail:
ullrich@psychologie.tu-darmstadt.de