

UNIVERSITÄT MÜNSTER
INSTITUT FÜR WIRTSCHAFTSINFORMATIK

Interpreting Visual Attention: A Systematic Analysis of
Eye-Tracking Data

Abschlussarbeit im Rahmen des Projekt Seminars *Eyes Wide Scroll*

eingereicht am 25.09.2025 von

David Lika, Jan Felix Deuse, Joris Engels, Maik Philipp Paulsen
Nikolaj Schlumbohm, Philip Alexander van Rickelen, Jan Schnorrenberg, Fabian Kizio
537160, 537390, 536427, 537809, 537506, 539435, 536480, 526085
david.lika@uni-muenster.de, jan.deuse@uni-muenster.de, joris.engels@uni-muenster.de, mpaulsen@uni-muenster.de,
nschlumb@uni-muenster.de, philip.van.rickelen@uni-muenster.de, jan.schnorrenberg@uni-muenster.de,
fabian.kizio@uni-muenster.de

Prof. Dr.-Ing. Grimme
FORSCHUNGSGRUPPE
COMPUTATIONAL SOCIAL SCIENCE & SYSTEMS ANALYTICS

Interpreting Visual Attention: A Systematic Analysis of Eye-Tracking Data

The eyes are not only passive receivers of visual stimuli, but active seekers of information guided by attention. *Alfred L. Yarbus (Eye Movemnet and Vision, 1967)*

Kurzfassung

Lorem ipsum dolor sit amet, consectetur adipiscing elit. Ut purus elit, vestibulum ut, placerat ac, adipiscing vitae, felis. Curabitur dictum gravida mauris. Nam arcu libero, nonummy eget, consectetur id, vulputate a, magna. Donec vehicula augue eu neque. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Mauris ut leo. Cras viverra metus rhoncus sem. Nulla et lectus vestibulum urna fringilla ultrices. Phasellus eu tellus sit amet tortor gravida placerat. Integer sapien est, iaculis in, pretium quis, viverra ac, nunc. Praesent eget sem vel leo ultrices bibendum. Aenean faucibus. Morbi dolor nulla, malesuada eu, pulvinar at, mollis ac, nulla. Curabitur auctor semper nulla. Donec varius orci eget risus. Duis nibh mi, congue eu, accumsan eleifend, sagittis quis, diam. Duis eget orci sit amet orci dignissim rutrum.

1 Einführung

Nachdem die Datenerhebung abgeschlossen war, erfolgte zunächst eine explorative Analyse des Materials. In diesem ersten Schritt lag der Fokus darauf, einen umfassenden Überblick über die erhobenen Eye-Tracking-Daten zu gewinnen, ohne bereits spezifische Annahmen oder Hypothesen zu verfolgen. Ziel war es, Strukturen, Auffälligkeiten und Muster sichtbar zu machen, die sich aus den Blickverläufen, Fixationen und Sakkaden ergeben. Dazu wurden grundlegende Kennwerte betrachtet, wie etwa die mittlere Fixationsdauer, die Verteilung der Blickpunkte über die Stimuli hinweg oder die zeitliche Dynamik des Blickverhaltens. Diese explorative Phase diente insbesondere dazu, erste Einsichten in die Daten zu gewinnen, mögliche Einflussfaktoren zu identifizieren und relevante Variablen für eine weiterführende Analyse einzuzgrenzen. Gleichzeitig konnten wir dadurch prüfen, ob die Datenqualität den Anforderungen entspricht und ob sich erwartungsgemäß interpretierbare Strukturen abzeichnen.

Aufbauend auf diesen ersten Befunden folgte eine systematische Analyse. Während die explorative Untersuchung noch eher offen und hypothesesgenerierend angelegt war, stand in diesem zweiten Schritt die gezielte Überprüfung einer klar formulierten Fragestellung im Vordergrund. Auf Basis der im explorativen Teil gewonnenen Einsichten formulierten wir eine spezifische Hypothese, die sich auf ein

bestimmtes Muster im Blickverhalten bezog. Diese Hypothese wurde anschließend mit geeigneten statistischen Verfahren überprüft. Dadurch konnten wir nicht nur unsere anfänglichen Beobachtungen absichern, sondern auch präzise Aussagen darüber treffen, ob die vermuteten Zusammenhänge tatsächlich empirisch gestützt werden können.

Durch das zweistufige Vorgehen - zunächst explorativ, dann hypothesesgeleitet - wurde sichergestellt, dass wir sowohl unvoreingenommene Einblicke in das Datenmaterial erhielten als auch die Möglichkeit hatten, spezifische Annahmen fundiert zu prüfen. Dieses Vorgehen verbindet Offenheit gegenüber unerwarteten Mustern mit wissenschaftlicher Stringenz bei der Hypothesenprüfung und ermöglicht damit eine differenzierte Interpretation der Eye-Tracking-Daten

2 Clustern

Die Clusteranalyse fasst die Fixations-Zusammenfassungen pro Bild zu Verhaltensarchetypen zusammen. Eine Kreuztabelle aus Clustern und Labels ordnet alle 152 annotierten Stimuli über drei Cluster hinweg und zeigt, wie Bilder den Lösungsraum füllen, der die weitere Interpretation steuert. Die Merkmalsvektoren kombinieren verschiedene Blickparameter: Fixationsdauern in frühen und späten Dritteln, mittlere und mediane Betrachtungszeiten, Scanpfadlängen, Blickstreuung über 68 %- und 95 %-Konfidenzellipsen, Fixationsanzahlen pro Teilnehmer, mittlere Pupillengröße und gesamte Betrachtungszeit. So entsteht ein mehrdimensionales Profil für jedes Bild.

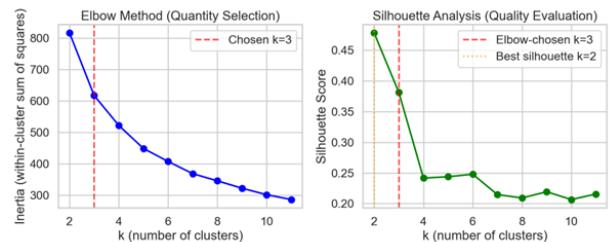


Abbildung 1

Um die optimale Anzahl an Clustern zu bestimmen, wurden sowohl das **Elbow**- als auch das **Silhouette**-Verfahren verwendet: Die Trägheit (Inertia) fiel deutlich

von 983,8 auf 695,4 zwischen k=1 und k=3 und nahm danach nur noch langsam ab. Die Silhouette-Werte waren bei k=2 mit 0,437 am höchsten, blieben bei k=3 aber stabil bei 0,231. Auf Basis des Elbow-Kriteriums wurde daher k=3 gewählt, auch wenn die Cluster bei k=2 etwas homogener waren. 1 zeigt beide Kurven nebeneinander und verdeutlicht die Entscheidungsgrundlage für drei Cluster.

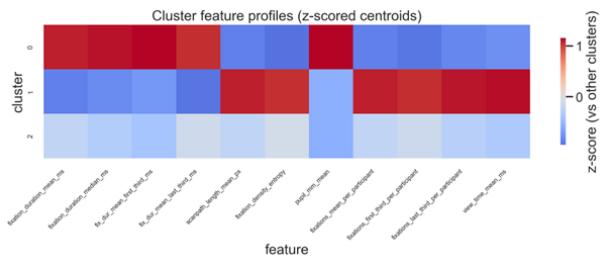


Abbildung 2

Cluster 0 repräsentiert ein bedachtes Betrachten: Fixationsdauern liegen mit durchschnittlich 299,7 ms hoch, sowohl im frühen als auch im späten Zeitfenster über 285 ms. Gleichzeitig sind Scampfade mit 2.664 px kurz und die Blickstreuung am geringsten Hinweise auf fokussierte Aufmerksamkeit auf wenige Bildbereiche. Von seinen 62 Bildern stammen 27 aus der Kategorie Person und 12 aus Ort, ergänzt durch Meme- und Politik-Overlays. Dieses Muster verknüpft konzentrierte Aufmerksamkeit mit sozial und räumlich geprägtem Bildmaterial. (siehe 3) *Cluster 1* steht für schnelles Scannen: Teilnehmende fixieren im Schnitt 38,7 Mal pro Bild und betrachten es 11,16 Sekunden, doch einzelne Fixationen dauern mit 222,3 ms kürzer, die Pupillen verengen sich auf 3,57 mm. Dies deutet auf wiederholte, schnelle Blickwechsel unter kognitiver Belastung hin. Hier dominieren textlastige Kompositionen etwa zehn Meme-Text-, neun Meme-Politik-Text- und sechs Person-Politik-Bilder, was den Aufwand für die Verarbeitung mehrschichtiger Inhalte widerspiegelt. (siehe 3) *Cluster 2* bildet einen ausgeglichenen Archetyp: Fixationsdauern liegen bei 262,9 ms, die Scampfade verlängern sich auf 3.157 px, und die Pupillen weiten sich leicht auf 3,68 mm. Diese Mischung spricht für eine integrierte, aber weniger intensive Exploration. Unter den 52 Bildern dominieren 19 Person-Szenen, doch auch Ort-Text- und Meme-Bilder sind vertreten. Das Cluster fungiert als flexibler Basisfall für verschiedene semantische Kontexte. (siehe 3)

Abschließend lassen sich für die drei Cluster folgende Schwerpunkte festhalten:

- *Cluster 0* : höchste Fixationsdauern und größte Pupillengröße
- *Cluster 1*: Spitzenwerte bei Fixationsanzahl und Gesamtbetrachtungszeit

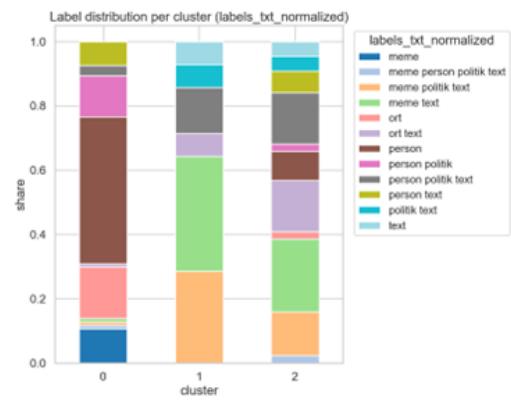


Abbildung 3

- *Cluster 2*: mittlere Werte bei Pupillengröße und Fixationsanzahl und bildet den Durchschnitt zwischen den Extremen

Abbildung 4 zeigt die standardisierten Clusterzentren als Koordinatensystem und erlaubt so den visuellen Vergleich zeitlicher, räumlicher und physiologischer Muster.

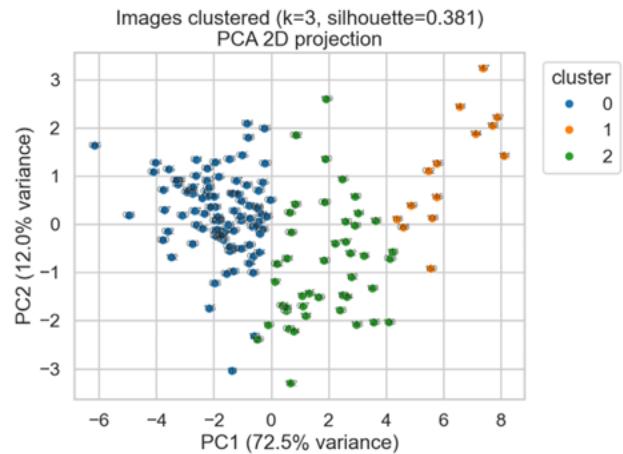


Abbildung 4

Insgesamt wandelt das Clustering-Analyse die hochdimensionalen Blickdaten in drei stabile Aufmerksamkeitsstrategien um, die eng mit den Label-Zusammensetzungen und den verwendeten Entscheidungsmaßen verknüpft sind. Durch die Kombination aus interpretierbaren Clusterzentren, verlässlichen Verfahren zur Bestimmung der Clusterzahl und label-spezifischen Mitgliedszahlen entsteht ein Workflow, der eine solide Grundlage bietet, um Blickverhalten und Inhaltskategorien in vergleichenden oder längsschnittlichen Studien systematisch miteinander zu verbinden.

3 Betrachtungsdauer

3.1 Zeitpunkt

Hypothese: Die Betrachtungsdauer von später gezeigten Bildern ist geringer als Bilder, die früh gezeigt werden.

Um diese Hypothese zu testen, wurde ein lineares Mixed-Effects-Modell verwendet, welches die Reihenfolge des Zeigens (als erstes gezeigt = 1, als zweites = 2 usw.) als Prädiktor nimmt und einen normalverteilten, zufälligen Interzept hinzuzieht, welcher unterschiedliche Grundniveaus (manche Personen schauen generell länger auf Bilder als andere) berücksichtigt. Für eine stabilere Schätzung wurden die Betrachtungsdauern log-transformiert. Die Berechnung wurde mit Python durchgeführt, mithilfe des mixedlm-Modells der statsmodels.formula.api-Bibliothek. Die Daten wurden durch `numpy.log1p()` log-transformiert.

Bei diesem linearen Mixed-Effects-Modell ist eine Steigung der Reihenfolge von $\beta \approx -0.00071$ mit einem 95%-Konfidenzintervall $[-0.000837; -0.000579]$ ermittelt worden, was einem Abfall der Betrachtungsdauer von ungefähr 0.071% pro Bild und auf 152 Bilder gesehen ca. 10.792% entspricht. Das bedeutet, dass jedes Bild (in Sekunden umgerechnet) um ca. 0.0075 Sekunden pro Schritt kürzer betrachtet wird als das vorherige. Auf 152 Bilder gesehen ergibt das einen gesamten Abfall von ca. 1.14 Sekunden vom ersten auf das letzte Bild. Weitere Daten dieses Modells sind wie folgt:

- $SE = 0.0000658$
- $t = \beta/SE = -10.75$
- Zweiseitig $p = 5.6 \cdot 10^{-27}$
- Einseitig für später = kürzer $p = 2.8 \cdot 10^{-27}$

Die sehr kleine Standardabweichung (SE) zeugt von einer sehr präzisen Schätzung der Steigung. Außerdem ist der t-Wert extrem (negativ) groß und die p-Werte beide extrem nahe an null, was auf einen hochsignifikanten negativen Trend schließen lässt. Somit bestätigen unsere Daten die oben genannte Hypothese. In Abbildung 5 ist eine lineare Approximation der Betrachtungsdauer zu sehen. Sie veranschaulicht die Bestätigung unserer Hypothese.

Diese Hypothese ist uns bereits vor der Durchführung unserer Studie in den Sinn gekommen, weshalb wir vorab bereits allen Probanden eine zufällige Reihenfolge der Bilder gezeigt haben, um bei Bestätigung dieser Hypothese keine Ergebnisse erhalten, die mit einem Bias versehen sind. Durch die zufällige Reihenfolge ist davon auszugehen, dass jedes Bild gleichhäufig an den entsprechenden Zeitpunkten gezeigt wurde, weshalb jedes Bild denselben Bias haben sollte und dieser sich somit ausgleicht. Zukünftige Studien sollten dies ebenfalls beachten.

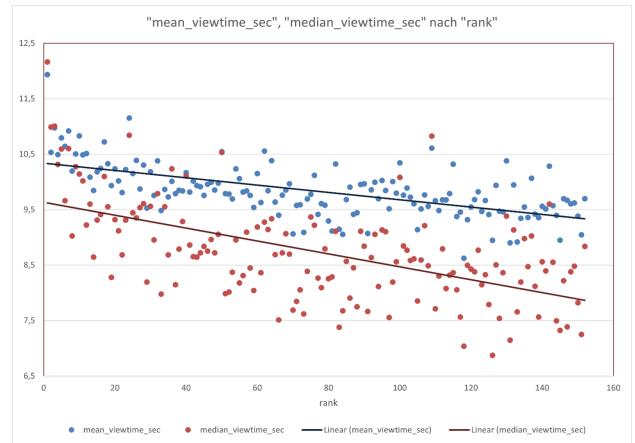


Abbildung 5 Lineare Approximation der Betrachtungsdauer in Abhängigkeit vom Zeitpunkt der Bildpräsentation. Durchschnitt (blau) und Median (rot) sind dargestellt.

3.2 Kategorien

Hypothese: Die Kategorien eines Bildes haben Einfluss auf die Betrachtungsdauer im Median

Um diese Hypothese zu testen, verglichen wir zunächst die Mediane über unsere Kategorien hinweg. Zwischen unseren Kategorien (Personen, Politik, Memes, Orte und Text) sowie ihren Kombinationen fallen bereits einige Kategorien als besonders prominent auf. Generell ist auffällig, dass Kombinationen mit Text längere Betrachtungsdauer aufweisen als der Konterpart ohne Text (Abb. 22)

Diese Tatsache hing nach näherer Betrachtung allerdings im Wesentlichen von der Menge an Text ab, den die Probanden in der Regel sorgfältig lasen. Abb. 20 zeigt die Betrachtungsdauer der einzelnen Kategorien aufgeschlüsselt auf Text und keinen Text.

Daher versuchten wir, den Einfluss der Textlänge herauszufiltern. Als erste Lösung bereinigten wir die Betrachtungsdauer um die Anzahl der Wörter mit folgender Formel:

$$\text{Dauer}_{\text{bereinigt}} = \frac{\text{Dauer}_{\text{gemessen}} - 5}{\text{Wörteranzahl}}$$

Die resultierenden relativen Betrachtungsdauern sind in Abbildung 21 dargestellt, wobei ein Vergleich mit Bildern ohne Text wegen fehlender Vergleichbarkeit außen vor gelassen wurde.

Eine Möglichkeit, um den Einfluss jeder Kategorie auf die Betrachtungsdauer zu ermitteln, ist die Regression. In unserem Fall lag eine Zensierung der Daten, also eine Beschränkung auf ein bestimmtes Intervall von 5 bis 15 Sekunden vor, weshalb eine gewöhnliche Regression ungeeignet schien. Eine Alternative stellte die Überlebenszeitanalyse und insbesondere das Modell für beschleunigte Aus-

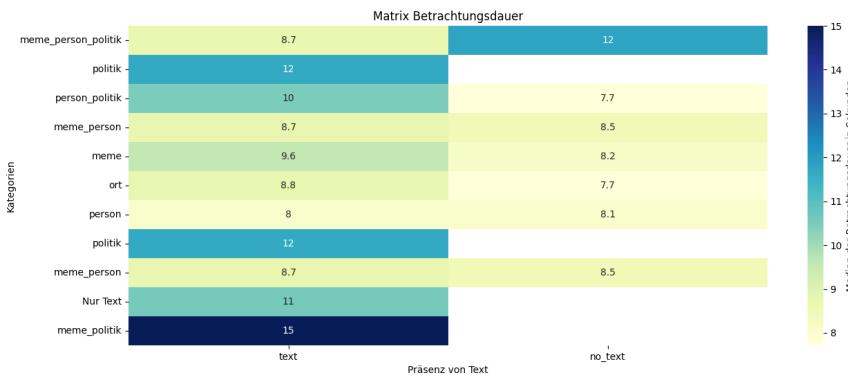


Abbildung 6 Die Kategorien mit der jeweiligen durchschnittlichen Betrachtungsdauern

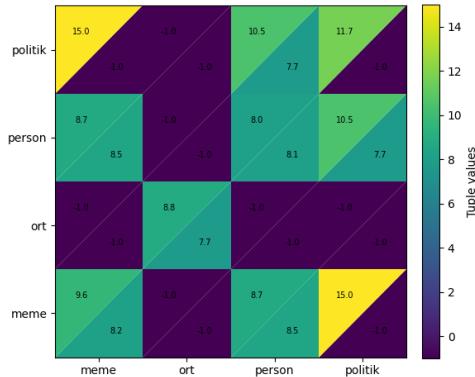


Abbildung 7 Jedes Feld zeigt die Kombination zweier Kategorien, mit dem Wert für "Text" linksoben und "Ohne Text" rechtsunten

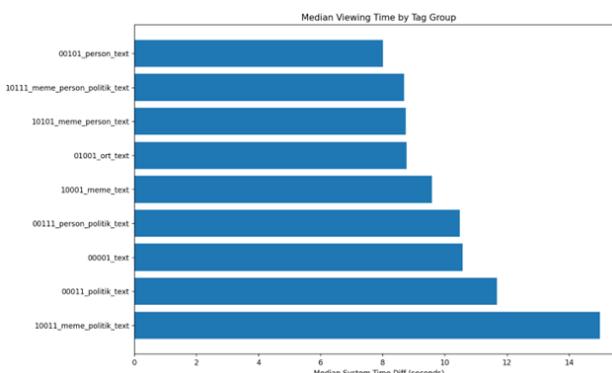


Abbildung 8 Median der Betrachtungsdauer pro Gruppe

fallzeiten (AFT) dar, das mit zensierten Werten umgehen kann.

Wir verglichen die AICs verschiedener AFT-Modelle und wählten den LogNormal-AFT-Fitter, der den besten AIC besaß. Die Ergebnisse zeigt Tabelle 1. Es stellte sich heraus, dass die Anzahl der Wörter mit einer Sekunde je weiterem Wort und die Kategorie Politik mit zusätzlichen 1,07 Sekunden bei Auftreten signifikant auf die Betrachtungsdauer einwirken.

Kategorie	Koeffizient	p-Wert
Ort	1.02	0.32
Person	1.05	0.01
Meme	1.03	0.12
Politik	1.07	< 0.005
Wörteranzahl	1.01	< 0.005

Tabelle 1 Regressionsergebnisse: Koeffizienten und Signifikanzniveaus für verschiedene Kategorien.

4 Kontraste

Hypothese: Hoher Bildkontrast korreliert positiv mit der Fixationsanzahl und -dauer.

4.1 Ansatz & Bildmetriken

Wir prüfen, ob und wie Bildkontrast mit dem Blickverhalten zusammenhängt, also ob kontrastreichere Bilder häufiger fixiert werden und ob Fixationen dabei länger oder kürzer ausfallen. Dazu kombinieren wir zwei Bildkontrast-Metriken (RMS und Laplace) mit über Personen aggregierten Fixationsdaten.

RMS-Kontrast: RMS-Kontrast misst die globale Helligkeitsspanne eines Bildes. Mathematisch ist das die Standardabweichung der auf $[0, 1]$ normalisierten Grauwerte. Hoher RMS heißt große Unterschiede zwischen hell und dunkel, unabhängig davon, wo sie im Bild liegen [1]..

Laplace: Laplace misst Kanten/Feinstruktur/Schärfe. Pipeline: (a) optional Gauß-Glättung (σ) zur Rauschreduktion, (b) Laplace-Filter, (c) Varianz des Laplace-Bildes, geteilt durch die Varianz des (geglätteten) Intensitätsbildes. Hoher Wert bedeutet viele/kräftige Kanten, feine Details [2].

RMS misst den globalen Helligkeits-Kontrast eines Bildes (wie stark die Pixelwerte insgesamt um ihren Mittelwert schwanken), während Laplace den lokalen Kanten-/Detail-Kontrast erfasst (zweite Ableitung vom Helligkeitsgradienten: starke Reaktion genau dort, wo Intensität schnell wechselt).

4.2 Fixationsmetriken

Als Fixationsmetriken verwenden wir pro Bild über alle Personen robust aggregierte Kennwerte: Die Blickdaten werden mit einem getrimmten Mittel (95%) zusammengefasst, um Ausreißer zu dämpfen. $n_{fix_mean_trim}$ bezeichnet dabei die mittlere Anzahl der Fixationen pro Bild, $tot_dur_mean_trim$ die mittlere Gesamtdauer aller Fixationen in Millisekunden und $med_dur_mean_trim$ den mittleren Median der Fixationsdauer (ebenfalls in ms), der besonders unempfindlich gegenüber extremen Einzelwerten ist.

4.3 Korrelationen

Zur Einordnung prüfen wir mittels Pearson-Korrelation, ob und wie stark die Kontrastmaße unserer Bilder (RMS bzw. Laplace) linear mit den Metriken (Fixationsanzahl und -dauer) zusammenhängen. Die wichtigsten Zusammenhänge mit Signifikanzniveau von 0,05: siehe Abbildung 9 und 23 bis 27

4.4 Interpretation & Ergebnisse

Bilder mit hohem RMS-Kontrast oder hohen Laplace-Werten ziehen den Blick häufiger an und führen trotz kürzerer Einzelfixationen zu einer höheren Gesamtdauer des Betrachtens. Das Muster passt zu einem Scanning-beziehungsweise Explorationsverhalten: Viele visuelle Ankerpunkte wie Kanten, Texturen oder Beschriftungen werden nacheinander kurz inspiziert. Dass die Laplace-Metrik stärker mit der Anzahl der Fixationen korreliert als RMS, legt nahe, dass Kanten und Feinstruktur ein besonders wirksamer Treiber der visuellen Aufmerksamkeit sind. Diese Befunde sind mit Ergebnissen aus der Salienz-Forschung vereinbar und knüpfen an die Literatur zu vorhersagbaren Blickbewegungen an [3].

Wichtig ist eine methodische Einordnung: Eine Korrelation beweist keine Kausalität. Kategorien wie Text oder

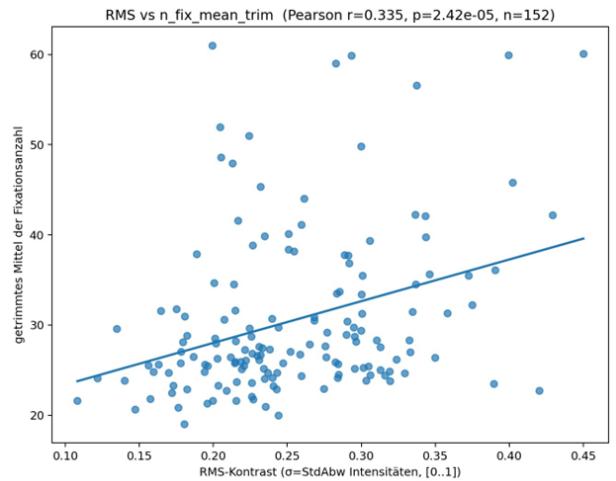


Abbildung 9 Zwischen dem RMS-Kontrast und der mittleren Fixationsanzahl zeigte sich eine signifikante positive Korrelation ($r \approx 0.34$; $p < 0,01$). Das bedeutet: Je höher der globale Bildkontrast, desto mehr Fixationen werden im Mittel auf dieses Bild gerichtet.

Personen können zugleich den Kontrast und das Blickverhalten beeinflussen und so den Zusammenhang verzerrern. Sinnvoll sind deshalb Analysen innerhalb homogener Kategorien, multiple Regressionen mit RMS, Laplace und Kategorie-Indikatoren, robuste Rangkorrelationen (Spearman) sowie Sensitivitätsanalysen für die gewählte Glättung beim Laplace-Maß.

Als zentrale Mitnahme lassen sich beide Kontrastmaße komplementär nutzen. RMS erfasst den globalen Helligkeitskontrast des Bildes, während Laplace den Kanten- und Detailkontrast quantifiziert. In Kombination erklären diese Maße plausibel, warum bestimmte Bilder insgesamt häufiger und länger fixiert werden, obwohl einzelne Fixationen tendenziell kürzer ausfallen. Ein naheliegender nächster Schritt ist eine reguläre, vorzugsweise kreuzvalidierte lineare Regression, um Fixationskennzahlen aus RMS- und Laplace-Werten vorherzusagen und den jeweiligen Beitrag beider Prädiktoren sauber zu trennen.

5 Wiederholte Betrachtung von Bildelementen

Hypothese: Bildkategorien haben Einfluss auf die Rekurrenz der Fixationen

5.1 Methodik

Recurrence quantification analysis (RQA) nach Anderson u. a. Kurzfassung: Zu jeder Bildbetrachtung wird aus der zugehörigen Fixationssequenz eine Rekurrenz Matrix berechnet. Aus dieser wiederum lassen sich die Metriken *recurrence*, *determinism*, *laminarity* (jeweils in Prozent) und

center of recurrence mass (corm) bestimmen. Diese wurden zu allen erhobenen Eye-Tracker Daten berechnet, auf Korrelation überprüft und zu Histogrammen (nach unterschiedlichen Bildkategorien) aggregiert.

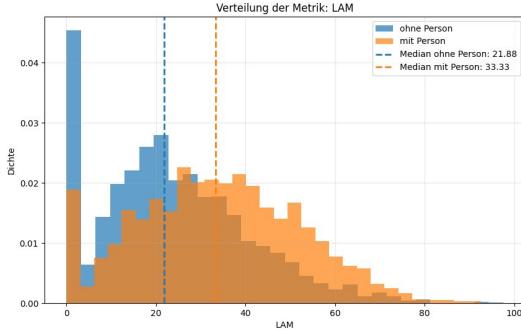


Abbildung 10 Histogramm Vergleich nach Person und laminarity LAM

5.2 Ergebnisse und mögliche Interpretation

Abbildung 31a und 31b dienen als Beispiel dafür, wie sich die Metriken ergeben. Abbildung 11 zeigt die Verteilung der Metriken als Histogramm. Auffällig ist die hohe Anzahl an Null-Werten bei *determinism* und *laminarity*. In der Korrelationsmatrix in Abbildung 28 lässt sich erkennen, dass diese zum Teil mit einer niedrigen *recurrence* korrelieren (mehr dazu im Vergleich zwischen Bildkategorien nach Ort). Denn: Je weniger sich Fixationen wiederholen, desto seltener wiederholen sich Fixationssequenzen (*determinism*) und desto seltener werden AOIs im Detail gescannt (*laminarity*).

Hohe *determinism* Werte können sich andererseits bspw. aus dem wiederholten Lesen von Text ergeben (Abb. 29a und reffig:Bild13b). Zusätzlich können hohe Ausreißer entstehen, wenn sich der Proband nur eine AOI anschaut (Abb. 30a und 30b).

5.2.1 Vergleich zwischen Bildkategorien

In Abbildung 32 lässt sich erkennen, dass Bilder mit Text eine höhere Anzahl an Fixationen haben, dies wird besonders deutlich ab mehr als 50 Fixationen. Die *recurrence* dagegen ist für Bilder mit Text deutlich niedriger (Abb. 33). Im Kontext der Bildbetrachtung bedeutet dies, dass sich Texte je länger sie sind, seltener mehrmals durchgelesen werden. Schaut man auf die Korrelationsmatrix in Abbildung 28, bestätigt sich der Trend der sinkenden *recurrence* mit steigender Fixationsanzahl. Der niedrigere corm Wert (Abb. 34) ergibt sich vermutlich daraus, dass Textelemente nur einmal komplett am Anfang der Bildbetrachtung gelesen werden. Das reicht, um den Wert trotz

nachfolgender normaler Betrachtung der Bildelemente zu senken (siehe Abb. 35a und 35b).

Beim Vergleich von Bildern mit und ohne Person, sticht eine deutlich höhere laminarity heraus (Abb. 10). Dies liegt vermutlich daran, dass vor allem Gesichter im Verlauf der Bildbetrachtung erneut genauer betrachtet werden.

Beim Vergleich von Bildern nach Ort ist auffällig, dass Bilder mit Ort bei *determinism* und *laminarity* prozentual mehr als die doppelte Anzahl an fast Null-Werten haben als ohne Ort (Abb. 36a und 36b), obwohl die *recurrence* nur um 1,52% sinkt. Dies könnte daran liegen, dass die AOIs bei reinen Orts-/ Landschaftsbildern oft verstreuter und unklarer sind.

Für Memes lassen sich die Unterschiede darauf zurückführen, dass diese oft Text, während politische Bilder oft Personen enthalten.

5.3 Limitationen

Die RQA-Ergebnisse sind sehr umfangreich und hochaggregiert, weshalb diese Analyse und Interpretation der Ergebnisse nicht alle Auffälligkeiten und Zusammenhänge behandelt.

6 Segmentation

Hypothese: Probanden zeigen eine höhere Anzahl von Fixationen auf den erkannten Personenmasken im Vergleich zu anderen Bildbereichen.

6.1 Methodik

Dieses Teil widmet sich der Blickverhaltensanalyse innerhalb der Personen Segmentation mit Hilfe des Segmentation Models YOLOv8 verbunden mit den gesammelten Eye-Tracking Daten. Aus den validen 7.443 aufgezeichneten Versuchen haben wir das Bincode-Suffix jedes Identifikators entschlüsselt und nur Stimuli beibehalten die Teil der Personenkategorie waren. Die Versuche wurden anschließend auf nicht leere Masken überprüft; 48 Bilder mit leeren Vorhersagen wurden verworfen, sodass 4.593 Versuche aus 104 einzigartigen Stimuli übrig blieben. Diese Filterung stellte sicher, dass nachgelagerte Schätzungen echte Personeninhalte abfragten und die Interpretation maskenabhängiger Metriken geschützt wurde.

Für die verbleibenden Versuche haben wir die Blickkoordinaten auf Segmentierungspixel abgebildet, die Anzahl und Dauer der Fixierungen innerhalb und außerhalb der Masken gezählt und diese Anteile nach Maskenfläche normalisiert. Die durchschnittliche Maskenabdeckung betrug 0,378 der Stimulusebene, was als Null-Erwartung bei gleichmäßiger Betrachtung diente. Dennoch richteten die

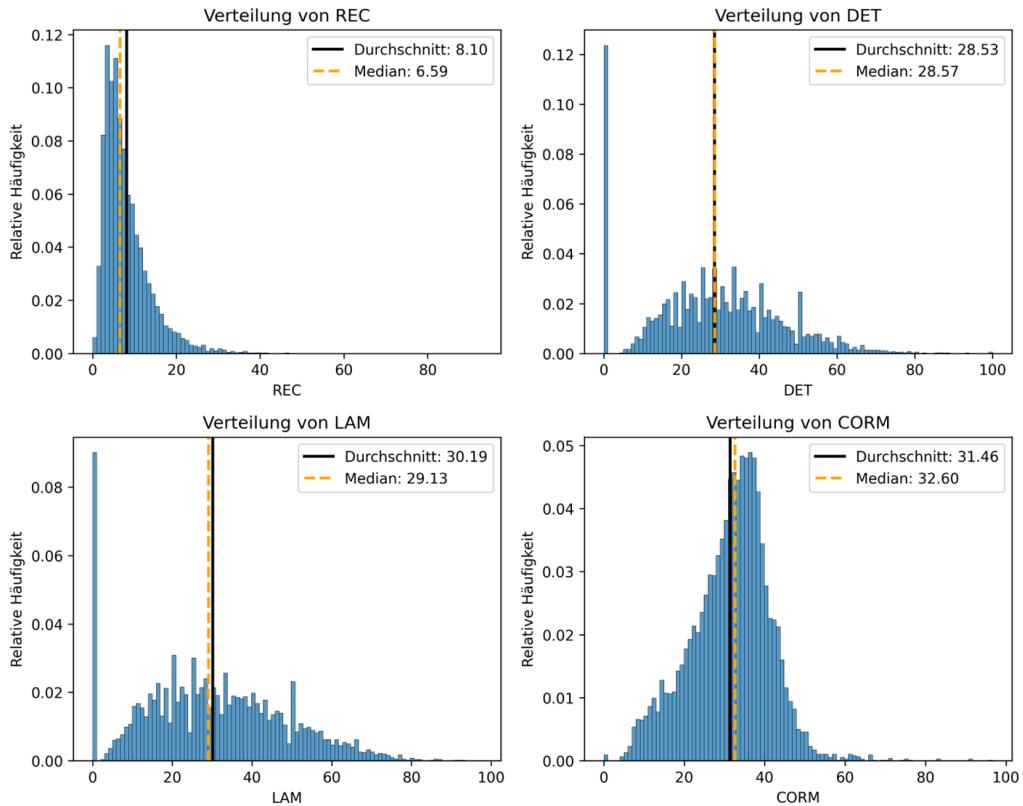


Abbildung 11 Histogramme der Metriken aller Daten

Beobachter 61,2 Prozent ihrer Fixationen auf die Personenbereiche, was zu einem mittleren Überschuss von 0,234 über der Flächenbasislinie führte. Zeitbasierte Messungen spiegelten dieses Muster wider: Die Teilnehmer verbrachten 65,5 Prozent ihrer Verweildauer auf den Masken, 0,263 mehr als der oberflächengewichtete Referenzwert. Selbst das untere Quartil der Teilnehmer verzeichnete Dichte-verhältnisse von mehr als eins, was die Konsistenz der Aufmerksamkeitsverzerung belegt.

6.2 Ergebnisse

Abbildung 12 fasst die Verteilung der Fixationsüberschüsse zusammen. Das Histogramm liegt überwiegend rechts von Null, da 3.943 der 4.593 Versuche einen positiven Überschuss ergeben. Der lange rechte Teil spiegelt Szenen wider, in denen kompakte Personensegmente den Blick monopolisierten. Defizite, die in 650 Versuchen auftreten, entstehen in der Regel, wenn Masken große Hintergrundbereiche abdecken, die die Bereichsnormalisierung verwässern, doch selbst in diesen Fällen wird den Personen immer noch beträchtliche absolute Aufmerksamkeit geschenkt.

Abbildung 13 stellt den Fixationsanteil gegenüber der Maskenabdeckung dar. Die meisten Punkte liegen über der Identitätslinie, was zeigt, dass die Fixationsanteile über das gesamte Abdeckungsspektrum hinweg die Maskenflä-

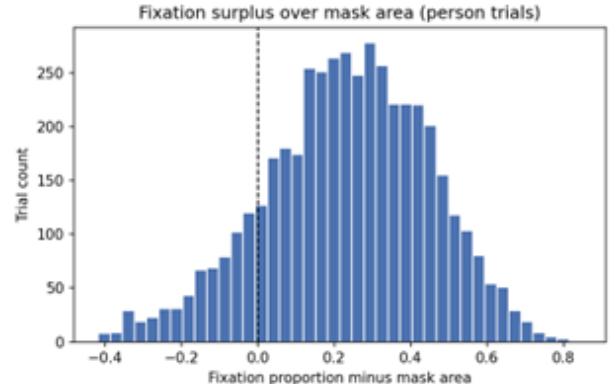


Abbildung 12 Verteilung der Fixationsüberschüsse

chen übersteigen. Masken, die weniger als 20 Prozent der Pixel einnahmen, zogen immer noch mehr als 40 Prozent der Fixationen auf sich, während Masken, die mehr als die Hälfte des Bildschirms bedeckten, überproportional vertreten blieben. Die Streuung bestätigt somit, dass die Segmentierungspipeline semantisch reichhaltige Bereiche isoliert, die den Betrachter unabhängig von ihrer Größe leiten.

Um die Stärke dieser Abweichungen zu formalisieren, haben wir Einstichproben-Z-Tests auf die Fixations- und Dauerüberschüsse angewendet. Standardfehler von

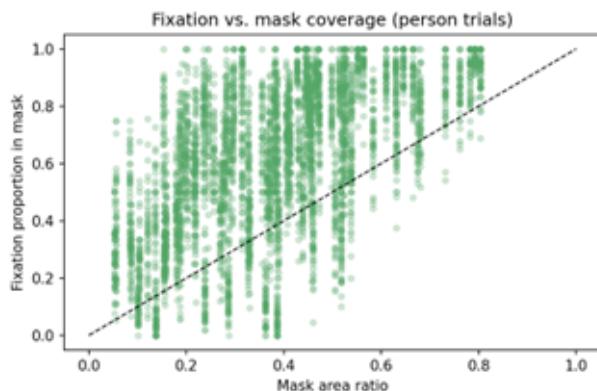


Abbildung 13 Fixationsanteil gegenüber der Maskenabdeckung

0,00318 und 0,00335 ergaben Z-Statistiken von 73,67 und 78,65, was unter der Normalapproximation zu p-Werten führte, die effektiv gleich Null waren. Auf Bildebene erreichten 37 der 94 analysierten Personenbilder durchschnittliche Fixationsüberschüsse von mindestens 0,300, während nur neun leicht unter Null fielen, typischerweise wenn Masken diffuse Hintergründe umhüllten oder Figuren teilweise verdeckt waren. Diese Ergebnisse bestätigen, dass computergestützt erkannte Personen eng mit den Blickprioritäten des Menschen übereinstimmen und bringen eine kleine Gruppe von Ausreißern hervor für weitere qualitative Folgeuntersuchungen. Insgesamt zeigt die Integration von Segmentierung bei Eye-Tracking, dass Masken den fokalen Inhalt der betrachteten Szenen erfassen und dass die Betrachter diesen Bereichen wesentlich mehr Aufmerksamkeit schenken, als die Oberflächenabdeckung allein vermuten lässt.

7 Scanpaths

Hypothese: Scanpaths von Bildern einer Kategorie sind ähnlicher untereinander als Scanpaths von Bildern anderer Kategorien.

Die MultiMatch-Methode (Jarodzka et al., 2010; Dewhurst et al., 2012) dient dem Vergleich zweier Scanpaths untereinander. Dazu werden die Scanpaths als Reihen von Sakkadenvektoren dargestellt, Fixationen sind die Endpunkte der Vektoren. Die Ähnlichkeit wird in fünf Dimensionen dargestellt: Form, Länge, Position, Richtung und Betrachtungsdauer.

Der Algoirthmus besteht aus fünf Schritten:

1. Vektorisierung: Die Fixationen werden in Sakkadenvektoren mit Polarkoordinaten umgewandelt.

2. Vereinfachung: Kleine oder kollinare Sakkaden werden zusammengefasst, um Rauschen zu reduzieren.
3. Ausrichtung: Scanpaths werden über einen gewichteten Graphen abgeglichen. Dabei werden ähnliche Sakkaden unter Beibehaltung der zeitlichen Reihenfolge zugeordnet.
4. Selektion: Der Dijkstra-Algorithmus bestimmt die Ausrichtung mit den geringsten Gesamtkosten.
5. Ähnlichkeitsberechnung: Die gepaarten Sakkaden werden verglichen und median-normalisierte Werte (01) für jede Dimension berechnet.

Alle Scanpaths wurden paarweise verglichen. Dadurch ergab sich eine etwa 7.000 × 7.000 × 5 große Ähnlichkeitsmatrix für den gesamten Datensatz, was einen Vergleich aller produzierten Scanpaths ermöglichte. Es zeigte sich, dass Stimulus id086 über alle Probanden hinweg (intra-Stimulus) das unterschiedlichste und Stimulus id144 das konstanteste Blickverhalten aufwies. Dies deckt sich mit der Erwartung, dass textbasierte Inhalte vorhersehbarere Blickmuster produzieren. Ein Vergleich der Stimuli untereinander (inter-Stimulus) ergab, dass Stimulus id008 hinsichtlich des Blickmusters am einzigartigsten war, während Stimulus id012 das durchschnittlichste und somit über den Datensatz repräsentativste Blickmuster zeigte. Dies kann vermutlich ebenfalls auf den hohen Textanteil, der fast das gesamte Bild umspannt, zurückgeführt werden.

Figur ?? zeigt eine zweidimensionale Darstellung der relativen Ähnlichkeit aller Scanpaths und somit den gesammelten Datensatz, eingefärbt nach Stimulus. Die Größe der Punkte stellt die Gesamtdauer der Fixationen dar. Hier lässt sich erkennen, dass gleiche Stimuli über Probanden hinweg erwartungsgemäß oft ähnliche Blickmuster hervorgerufen haben.

8 Text

8.1 Hypothese 1

Text auf Bildern zieht Aufmerksamkeit auf sich

8.1.1 Methodik

Um die Hypothese zu beantworten, muss zuerst bestimmt werden, was überhaupt Text ist. Dazu wurde der Text auf jedem Bild händisch markiert. Aus diesen Daten wurden nun für jedes Bild zwei Werte berechnet, die als Grundlage für viele der folgenden Analysen dienen: Der Textanteil des Bildes und der Prozentteil der Fixationen auf Text.



Abbildung 14 Aggregierte Heatmap des Bildes mit ID 008

8.1.2 Ergebnisse

Basierend auf diesen beiden Werten kann nun untersucht werden ob Texte die Aufmerksamkeit der Probanden auf sich gezogen haben. Dazu wurden auf Abbildung 17 alle untersuchten Bilder geplottet. Die X-Achse beschreibt den Textanteil, während die Y-Achse den Fixation-auf-Text Anteil beschreibt.

Sollte Text keinen Einfluss auf das Blickverhalten haben sollte der Fixation-auf-Text Anteil im Mittel etwa gleich dem Textanteil sein. Sollte der Fixationsanteil auf Text allerdings erkennbar höher sein, als der Textanteil kann ein Zusammenhang angenommen werden. Erkennbar ist, dass die Punkte (außer einzelne Ausreißer im Bereich der niedrigen Textanteile) alle deutlich links von der Linie liegen. Demnach kann ein Zusammenhang angenommen werden.

8.2 Hypothese 2

Verschiedene Arten von Bildern haben Einfluss auf das Blickverhalten

8.2.1 Methodik

Dazu wurden die Bilder in verschiedene Gruppen unterteilt:

- Nur Text: Text im Vordergrund. Der Hintergrund ist generisch und häufig einfarbig.
- Text Hauptbestandteil: Text im Vordergrund. Der Hintergrund hat nichts mit dem Inhalt zu tun. Häufig handelt es sich um Landschaftsbilder.

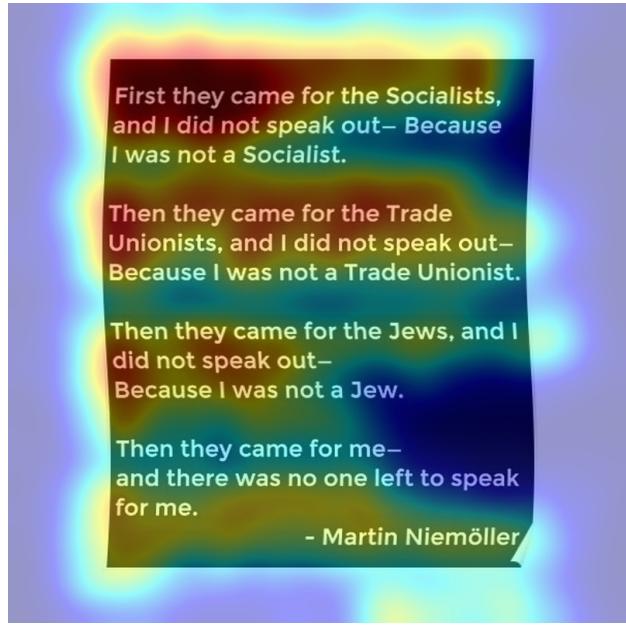


Abbildung 15 Aggregierte Heatmap des Bildes mit ID 144

- Text/Bild Kombination: Text und Bild hängen zusammen. Da die Bilder der Gruppe große Unterschiede aufweisen wird der Abschnitt weiter unterteilt in Memes und Zitate mit Person (Nur Bilder mit Zitat und zugehöriger Person).
- Text Hintergrund: Der Text spielt nur eine Untergeordnete Rolle.

8.2.2 Ergebnisse

In Abbildung 18 sind die beiden Variablen Text Anteil und Fixation-auf-Text Anteil für jede Gruppe geplottet. Die Kategorien unterscheiden sich deutlich.

Die nur Text Gruppe hat mit 51,46% den Höchsten Textanteil, gefolgt von Text Hauptbestandteil mit 26,42%. Trotz dieses Unterschiedes liegt der Fixation-auf-Text Anteil bei der Gruppe nur Text mit 89% nur um etwa 5% höher als bei Text Hauptbestandteil. Bei beiden Gruppen ist der Text das primäre Bildelement und der inhaltslose Hintergrund lenkt offenbar nur wenig ab.

Um das Verhältnis zwischen den Variablen weiter zu untersuchen ist in Abbildung 19 die Rechnung (Fixationen-auf-Text Anteil) / (Text Anteil) für die verschiedenen Gruppen aufgeschlüsselt. Hier ist vor allem der Wert von Text Hintergrund auffällig. Diese Kategorie hat den niedrigsten Text Anteil, allerdings ist der Fixationen-auf-Text Anteil prozentual am höchsten. Das heißt, dass überproportional viel Zeit auf dem Wenigen Text dieser Bilder geschaut wird und das, obwohl Text in diesen Kategorien



Abbildung 16 Aggregierte Heatmap des Bildes mit ID 012

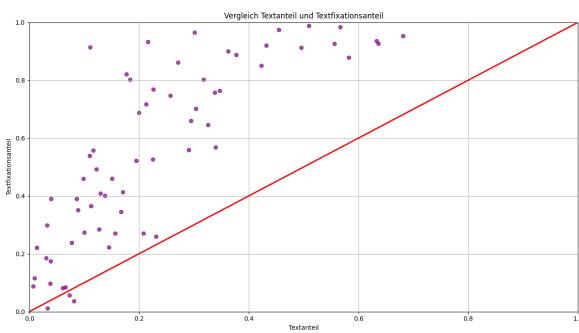


Abbildung 17 Vergleich Text Anteil mit Anteil Fixationen auf Text

lediglich eine Untergeordnete Rolle spielt. Auch auffällig ist der vergleichsweise niedrige Wert von Memes. Diese haben einen nur minimal geringeren Textanteil als zum Beispiel Zitate, dennoch ist der Fixationen-auf-Text Anteil deutlich niedriger. Das spricht dafür, dass das Bild hier wichtiger ist als bei den meisten anderen Kategorien.

8.3 Hypothese 3

Über den Zeitlichen Verlauf werden Bilder anders betrachtet

8.3.1 Ergebnisse

Zuletzt wird das Blickverhalten über den Zeitlichen Verlauf betrachtet. In Abbildung 37 kann der Anteil der Fixationen-auf-Text für jeden Zeitpunkt betrachtet werden. Das heißt, dass zum Beispiel für Memes 40% aller Gaze Points an der Stelle 200 auf Text lagen.

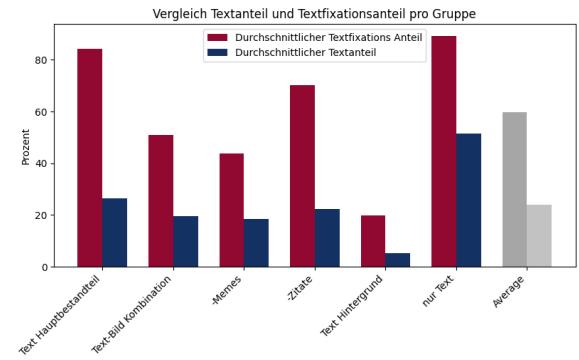


Abbildung 18 Vergleich Fixationen-auf-Text Anteil und Text Anteil

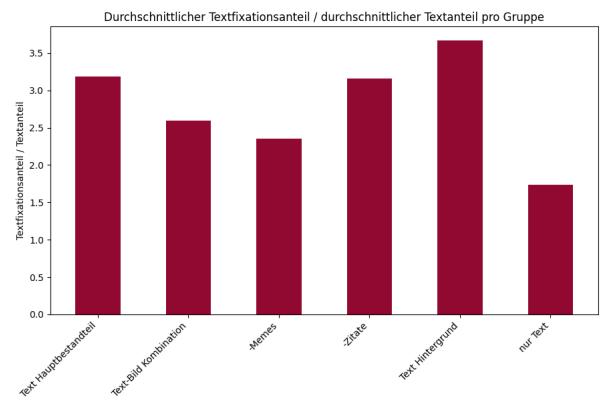


Abbildung 19 Fixationen-auf-Text Anteil / Text Anteil

Im Zeitlichen Verlauf ist ein starker Trend zu erkennen. Anfangs steigt der gaze-point-auf-Text Anteil stark an und erreicht bereits in den ersten Sekunden seinen Hochpunkt. Anschließend sinkt der Anteil wieder relativ schnell. Diese Ausprägung ist in verschiedenen Stärken zu beobachten. Besonders auffällig ist hierbei die Kategorie Memes. Bereits bei Gaze-point 50 erreicht diese den Hochpunkt von etwa 70%. Im Verlauf der nächsten Sekunden mehr als halbiert sich dieser Anteil auf nur noch etwa 30%. Das liegt vermutlich daran, dass zuerst der Text gelesen und anschließend das dazugehörige Bild betrachtet wird. Bei Zitaten dauert der Hochpunkt länger an, bevor auch hier der Anteil wieder stark abfällt. Die verlängerte Hochphase kann mit der größeren Menge an Text und zusammenhängenden Sätzen erklärt werden. Einzig die nur Text Kategorie verzeichnet kaum Schwankungen, was Sinn ergibt, da es außer Text keinen Inhalt gibt, der Betrachtet werden kann.

9 Fazit

Die durchgeführten Analysen haben gezeigt, dass Eye-Tracking ein vielfältiges und leistungsfähiges Werkzeug

ist, um Blickverhalten systematisch zu erfassen und Hypothesen über die Wahrnehmung und Verarbeitung visueller Inhalte zu überprüfen. Unser zweistufiges Vorgehen zunächst explorativ, dann hypothesesgeleitet erwies sich dabei als gewinnbringend: Während die explorative Phase Strukturen, Auffälligkeiten und erste Muster sichtbar machte, konnten in der hypothesenorientierten Phase diese Beobachtungen gezielt überprüft und quantifiziert werden.

Zentrale Befunde lassen sich wie folgt zusammenfassen:

- Die Betrachtungsdauer nimmt im Verlauf der Bildpräsentation signifikant ab, was auf Ermüdungseffekte oder eine zunehmende Sättigung bei den Teilnehmenden hindeutet.
- Bildkategorien insbesondere das Vorhandensein von Text oder Personen beeinflussen das Blickverhalten maßgeblich. Text zieht die Aufmerksamkeit in besonderem Maße auf sich, wobei die Wirkung je nach Kontext (Hintergrundtext, Meme, Zitat) variiert.
- Kontrastmerkmale wie RMS und Laplace korrelieren signifikant mit Fixationsmustern und deuten auf ein exploratives Scanning-Verhalten bei detailreichen Bildern hin.
- Rekurrenzanalysen zeigen, dass Bildinhalte wie Text und Gesichter zu charakteristischen Mustern wiederholter Fixationen führen, während landschaftliche Szenen eher durch eine zerstreute Betrachtung gekennzeichnet sind.
- Segmentierungsbasierter Analysen bestätigen, dass Personen in Bildern den Blick der Teilnehmenden besonders stark binden unabhängig von der relativen Flächenabdeckung.

Insgesamt verdeutlichen die Ergebnisse, dass Blickverhalten kein zufälliger Prozess ist, sondern systematisch durch Bildmerkmale, Präsentationsbedingungen und Kategorien geprägt wird. Die Kombination aus klassischen Kennwerten, statistischen Modellierungen und innovativen Verfahren wie der Rekurrenzanalyse oder der Segmentierung eröffnet dabei neue Perspektiven für die Aufmerksamkeitsforschung.

Für zukünftige Arbeiten erscheint es vielversprechend, die Ansätze durch größere Datensätze, eine stärkere Berücksichtigung individueller Unterschiede sowie multimediale Methoden (z. B. EEG, Verhaltensmaße) zu erweitern. So ließe sich noch präziser untersuchen, wie Menschen visuelle Informationen wahrnehmen, verarbeiten und bewerten.

Unsere Studie zeigt: Eye-Tracking ermöglicht nicht nur die Erfassung von Blickbewegungen, sondern liefert tiefe Einblicke in kognitive Prozesse der Aufmerksamkeit und Wahrnehmung und trägt damit zu einem besseren Verständnis menschlicher Informationsverarbeitung bei.

Literatur

- [1] H. Kukkonen, J. Rovamo, K. Tiippuna und R. Näsänen, „Michelson contrast, RMS contrast and energy of various spatial stimuli at threshold“, *Vision research*, Jg. 33, S. 1431–6, Aug. 1993. doi: 10.1016/0042-6989(93)90049-3
- [2] A. Starnes, A. Dereventsov und C. Webster, *Gaussian smoothing gradient descent for minimizing functions (GSmoothGD)*, 2023. eprint: [arXiv:2311.00521](https://arxiv.org/abs/2311.00521).
- [3] T. Foulsham und G. Underwood, „What can saliency models predict about eye movements? Spatial and sequential aspects of fixations during encoding and recognition“, *Journal of Vision*, Jg. 8, Nr. 2, S. 6, Feb. 2008, ISSN: 1534-7362. doi: 10.1167/8.2.6 Adresse: <http://dx.doi.org/10.1167/8.2.6>
- [4] N. C. Anderson, W. F. Bischof, K. E. W. Laidlaw, E. F. Risko und A. Kingstone, „Recurrence quantification analysis of eye movements“, en, *Behav. Res. Methods*, Jg. 45, Nr. 3, S. 842–856, Sep. 2013.

A Abbildungen

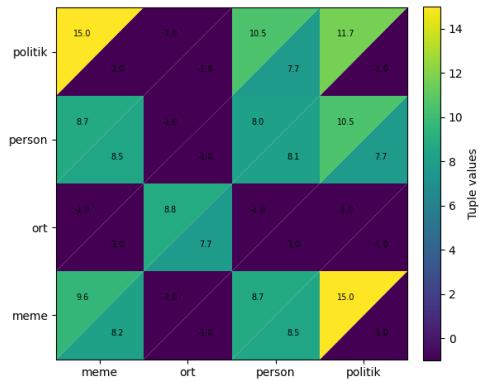


Abbildung 20 Jedes Feld zeigt die Kombination zweier Kategorien, mit dem Wert für "Text"linksoben und "Ohne Text"rechtsunten

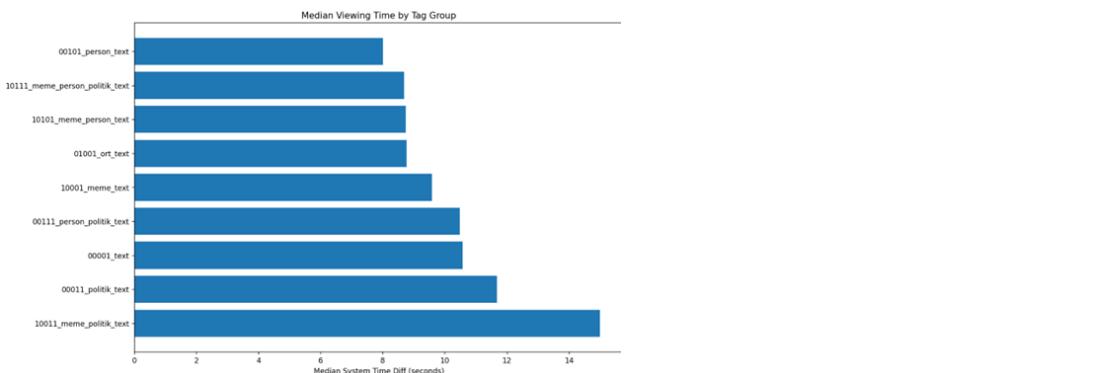


Abbildung 21 Median der Betrachtungsdauer pro Gruppe

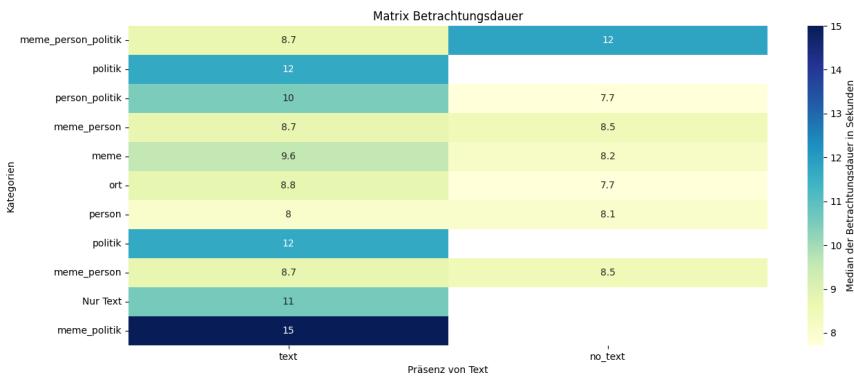


Abbildung 22 Die Kategorien mit der jeweiligen durchschnittlichen Betrachtungsdauern

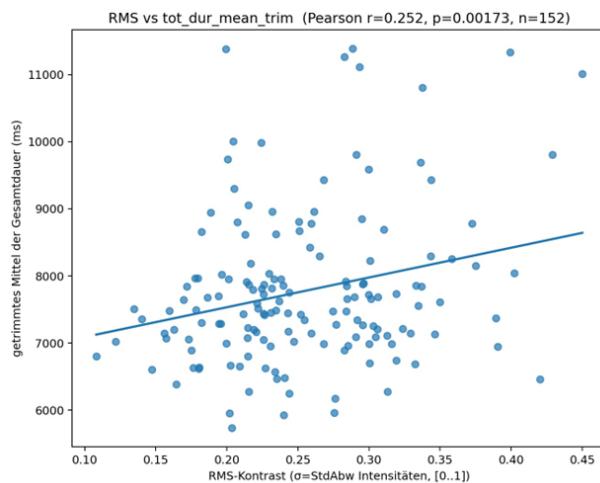


Abbildung 23 Zwischen dem RMS-Kontrast und der mittleren Gesamtdauer der Fixationen zeigte sich eine signifikante positive Korrelation ($r \approx 0.25$; $p < 0,01$). Das heißt Bilder mit höherem globalen Kontrast werden insgesamt länger betrachtet (summierte Fixationsdauer steigt).

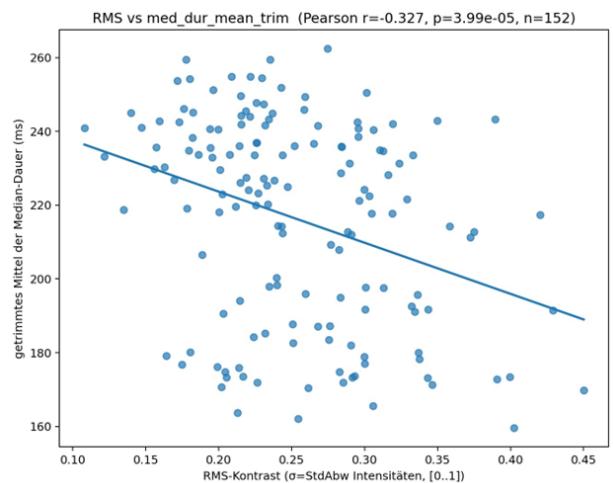


Abbildung 24 Zwischen dem RMS-Kontrast und der mittleren Median-Fixationsdauer zeigte sich eine signifikante negative Korrelation ($r \approx -0.33$; $p < 0,01$). Das heißt bei höherem globalen Kontrast fallen die typischen (medianen) Fixationen kürzer aus.

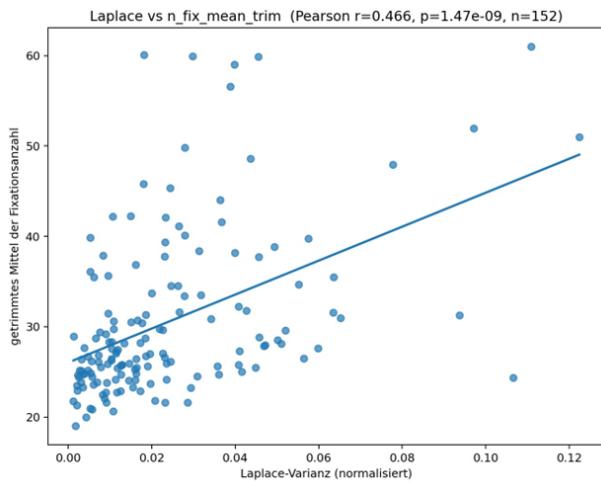


Abbildung 25 Zwischen der Laplace-Metrik und der mittleren Fixationsanzahl zeigte sich eine signifikante positive Korrelation ($r \approx 0,47$; $p < 0,001$). Das ist der stärkste beobachtete Zusammenhang - Bilder mit ausgeprägten Kanten und feiner Detailstruktur werden deutlich häufiger fixiert.

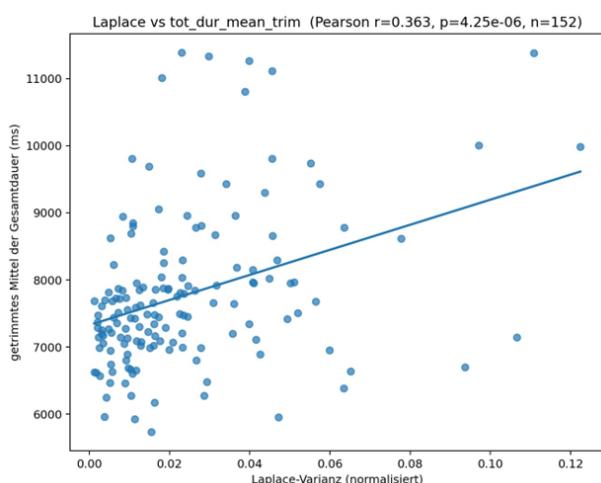


Abbildung 26 Zwischen der Laplace-Metrik und der mittleren Gesamtdauer der Fixationen zeigte sich eine signifikante positive Korrelation ($r \approx 0,36$; $p < 0,001$). Mit zunehmender Kanten- und Detailstärke steigt die insgesamt auf ein Bild verwendete Fixationszeit.

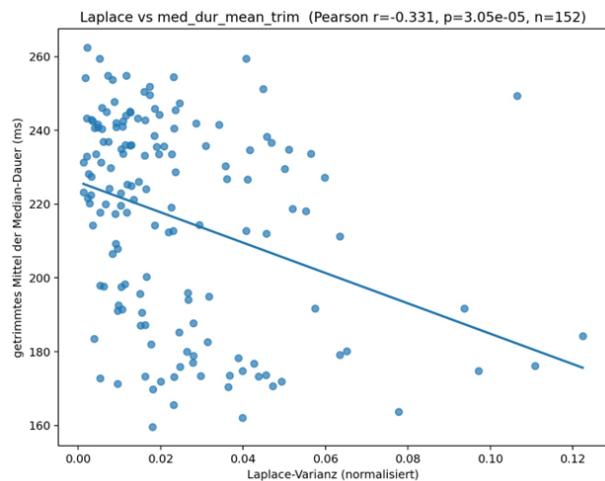


Abbildung 27 Zwischen der Laplace-Metrik und der mittleren Median-Fixationsdauer zeigte sich eine signifikante negative Korrelation ($r \approx -0,33$; $p < 0,01$). Mit mehr Kanten- und Detailstärke werden die typischen Einzel-Fixationen kürzer.

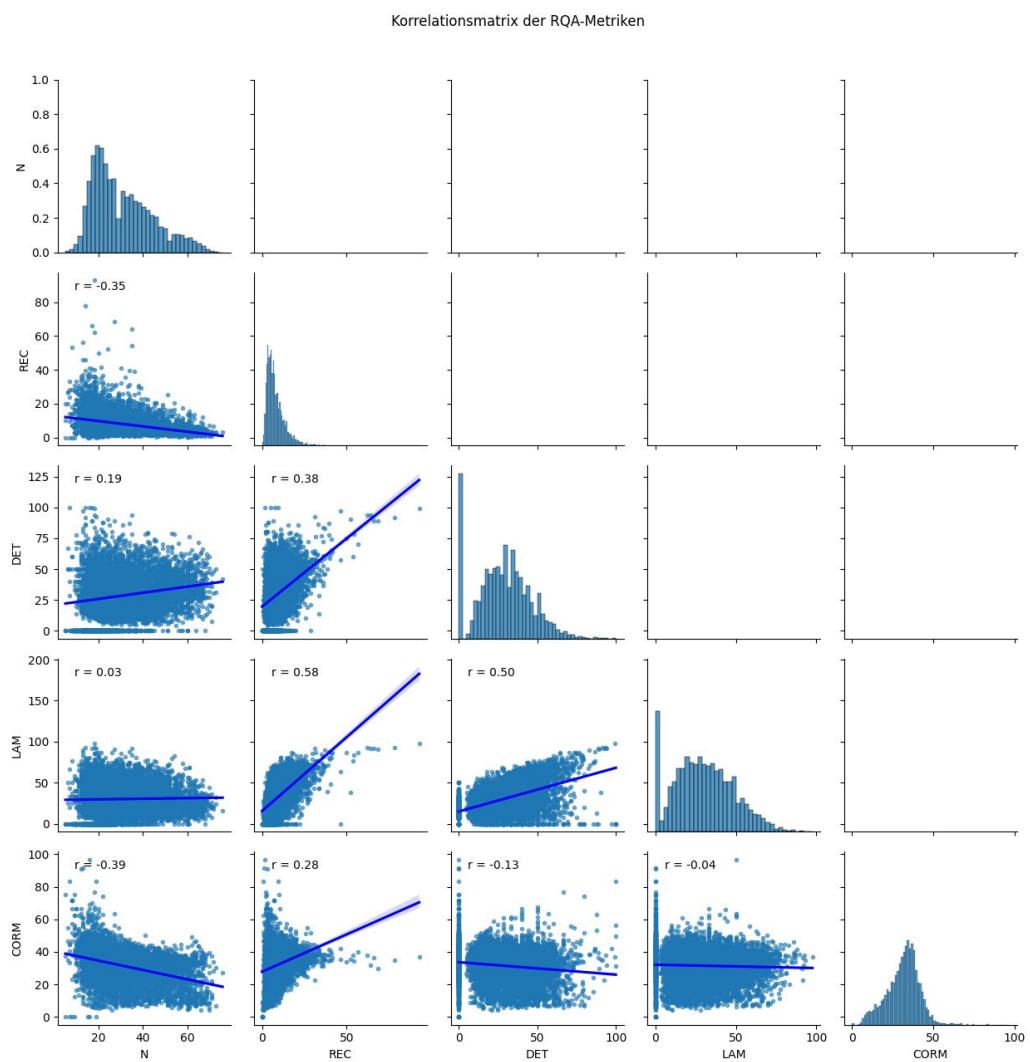
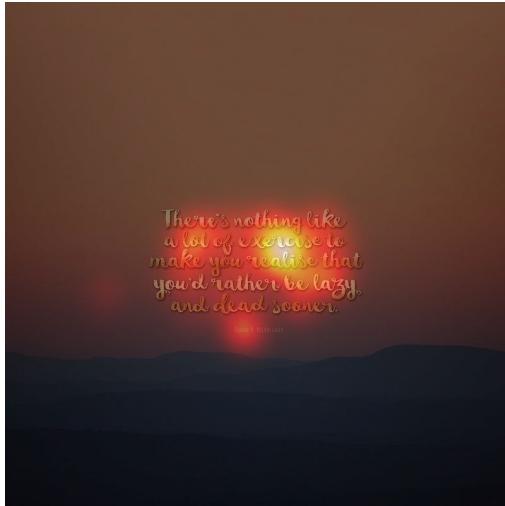
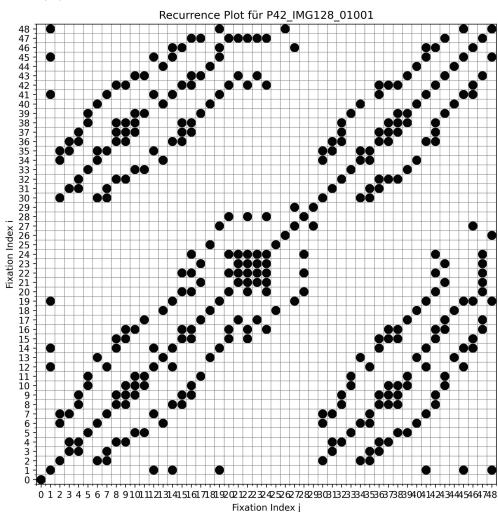


Abbildung 28 Korrelationsmatrix der Metriken + Histogramm

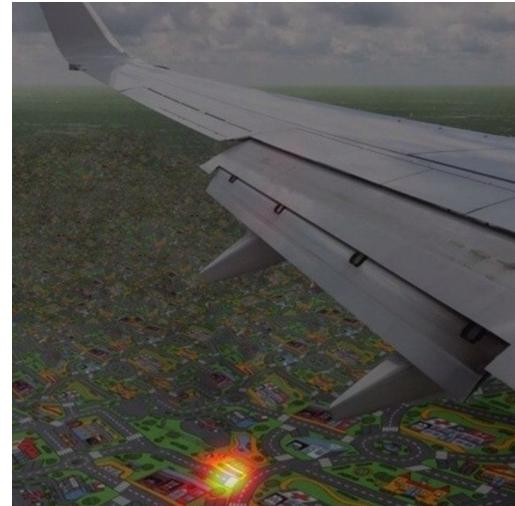


(a) Proband 42 mit Bild 128 und Heatmap

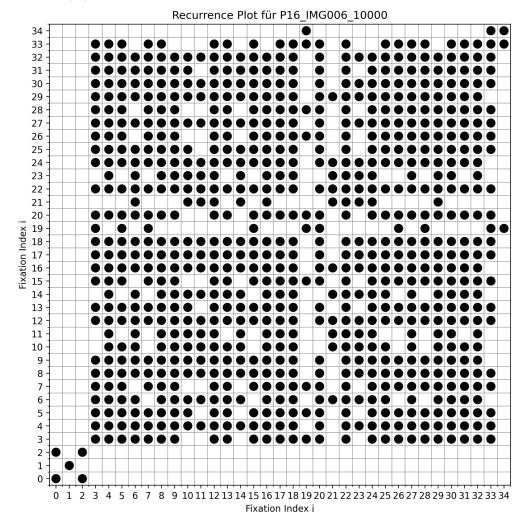


(b) Rekurrenz Plot zu Abbildung 29a

Abbildung 29 N = 49, R = 154, REC = 13.10%, DET = 70.%, LAM = 43.51%, CORM = 35.42%



(a) Proband 16 mit Bild 6 und Heatmap

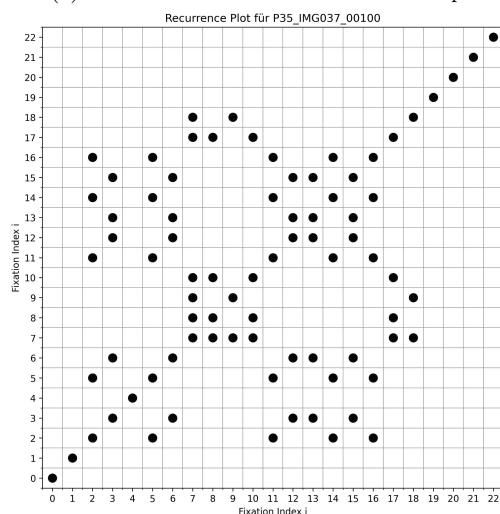


(b) Rekurrenz Plot zu Abbildung 30a

Abbildung 30 N = 35, R = 381, REC = 64.03%, DET = 93.96%, LAM = 91.99%, CORM = 31.80%



(a) Proband 35 mit Bild 37 und Heatmap



(b) Rekurrenz Plot zu Abbildung 31a

Abbildung 31 N = 23, R = 29, REC = 11.46%, DET = 62.07%, LAM = 25.86%, CORM = 29.78%

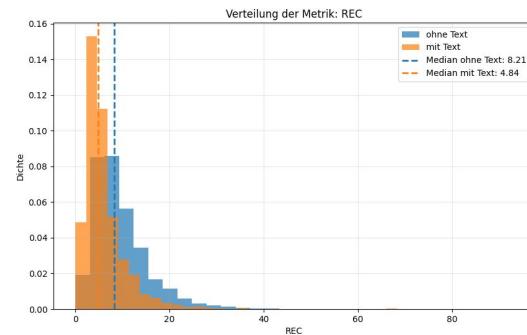


Abbildung 33 Histogramm Vergleich nach Text und reccurence REC

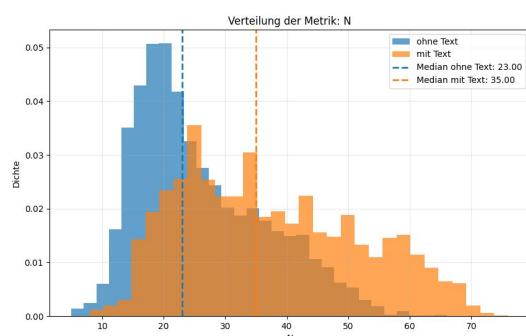


Abbildung 32 Histogramm Vergleich nach Text und Fixationsanzahl N

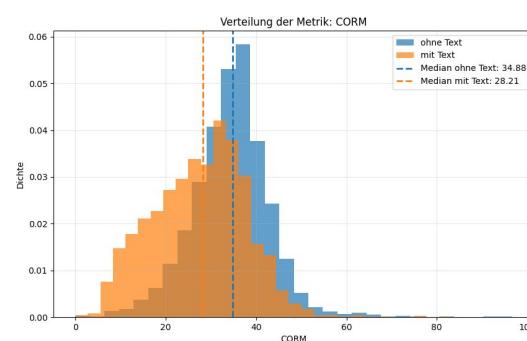
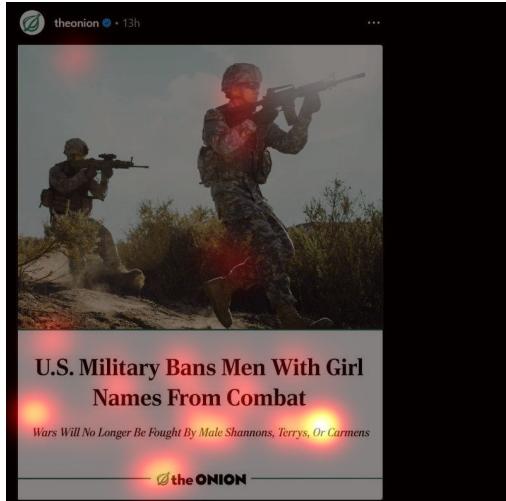
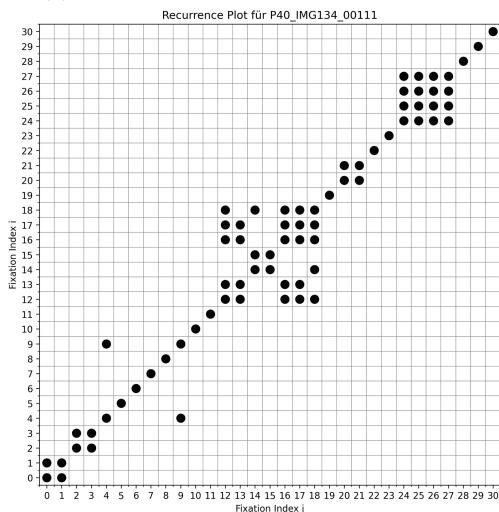


Abbildung 34 Histogramm Vergleich nach Text und corm

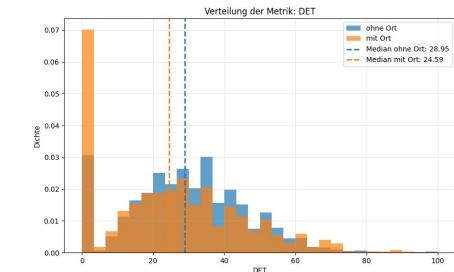


(a) Proband 40 mit Bild 124 und Heatmap

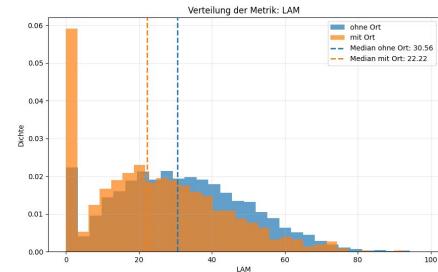


(b) Rekurrenz Plot zu Abbildung 35a

Abbildung 35 $N = 31$, $R = 21$, $REC = 4.52\%$, $DET = 47.62\%$, $LAM = 54.76\%$, $CORM = 7.94\%$



(a) Histogramm Vergleich nach Ort und determinism DET



(b) Histogramm Vergleich nach Ort und laminarity LAM

Abbildung 36 Vergleich der Histogramme mit Ort

PROBAND:INNEN GESUCHT!



HILF MIT - MIT DEINEM BLICK.

Nimm an einer spannenden Studie und unterstützen die Forschung in der Aufmerksamkeitsanalyse.

- Dauer ca. 30 Minuten
- Faire Vergütung inklusive

SCANNE JETZT DEN
QR-CODE
und mach mit!

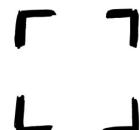


Abbildung 37 Fixationen-auf-Text Anteil/Text Anteil