Technische Universität Berlin

Untersuchung zu MOS als geeignetes Maß für die subjektive Bildqualität der LIVE2 Datenbank

Thesis zur Erlangung des akademischen Grades
Bachelor of Science (B. Sc.)
im Studiengang Informatik

Annalena Schillen
Fakultät IV - Elektrotechnik und Informatik

Gutachter:

Dr. Guillermo Aguilar
Computational Psychology

Prof. Dr. Marc Alexa
Computer Graphics

Technische Universität Berlin
Eingereicht: 12. August 2022

Selbstständigkeitserklärung

Hiermit erkläre ich, dass ich die vorliegende Arbeit selbstständig und eigenhändig sowie ohne unerlaubte fremde Hilfe und ausschließlich unter Verwendung der aufgeführten Quellen und Hilfsmittel angefertigt habe.

10.08.2022

Berlin, den

Unterschrift

Zusammenfassung

Die Verarbeitung und Verbreitung von digitalen Bildern ist ein zentraler Bestandteil der heutigen Kommunikationsgesellschaft. Zur Gewährleistung der Bildqualität werden automatisierte Bildqualitätsalgorithmen eingesetzt. Der Referenzstandard dieser Algorithmen sind subjektive Bildqualitätsdatenbanken. Diese Datenbanken enthalten Bilder und deren Bewertungen, die aus Wahrnehmungsexperimenten mit Versuchspersonen gewonnen wurden. Ein bewährtes Maß für die Bildqualität dieser Wahrnehmungsexperimente ist der Mean-Opinion-Score (MOS). Die LIVE2-Datenbank der University of Texas ist eine der bekanntesten und am häufigsten referenzierten Bildqualitätsdatenbanken. Sie basiert ebenfalls auf dem MOS als subjektivem Qualitätsmaß.

In dieser Arbeit wird die Eignung von MOS als Maß für die Bildqualität am Beispiel der LIVE2-Datenbank untersucht. Für die Bewertung wurden vier grundlegende Anforderungen an eine Qualitätsmetrik formuliert. Nach einer detaillierten Untersuchung aller LIVE2-Bilder und ihrer Bewertungen habe ich bei den veröffentlichten MOS-Bewertungen der LIVE2-Datenbank zahlreiche Verstöße gegen diese Anforderungen festgestellt. Um die Ergebnisse weiter zu verifizieren, führte ich ein Experiment durch, bei dem die MOS-Bewertungen für eine Auswahl von LIVE2-Bildern gemessen wurden. Darüber hinaus wurde ein zweites Experiment mit einer alternativen psychophysikalischen Methode, dem Maximum Likelihood Difference Scaling (MLDS), entworfen und durchgeführt.

Die Ergebnisse aus beiden Experimenten bestätigen die Defizite von MOS und zeigen die Vorteile von MLDS als Qualitätsmaßstab auf. Die gewonnen Erkenntnisse stellen die Eignung der LIVE2-Datenbank MOS-Bewertungen für die Entwicklung und Validierung von Bildqualitätsalgorithmen in Frage. Diese Arbeit legt die weitere Untersuchung von MLDS als alternative Qualitätsmetrik nahe. Mögliche Erweiterungen und die Überprüfung dieser Schlussfolgerungen mit anderen Bildqualitätsdatenbanken werden diskutiert.

Abstract

The processing and distribution of digital images is a central part of today's communication society. Automated image quality algorithms are used to ensure image quality. The standard reference of these algorithms are subjective image quality databases. These databases contain images and their ratings obtained in experiments with human subjects. An established image quality measure from these experiments is the Mean-Opinion-Score (MOS).The LIVE2 database of the University of Texas is one of the most well-known and widely referenced image quality databases. It is also based on MOS as a subjective quality score.

This thesis investigates the suitability of MOS as a measure of image quality, taking the LIVE2 database as an example case. Four basic requirements for a quality metric were formulated for the evaluation. After a detailed examination of all LIVE2 images and their ratings, I found numerous violations of these requirements for the published MOS ratings of the LIVE2 database. To further verify the results, I ran an experiment measuring MOS ratings on a selection of LIVE2 images. In addition, a second experiment was designed and run using an alternative psychophysical method, Maximum Likelihood Difference Scaling (MLDS).

The results from both experiments confirm the deficits of MOS and show the advantages of MLDS as a quality metric. They also questions the suitability of MOS scores to the LIVE2 database for the development and validation of image quality algorithms. This work also suggest further investigation of MLDS as an alternative quality metric. Possible extensions and verification of these conclusions with other image quality databases are discussed.

Danksagung

Mein besonderer Dank gilt meinem Erstgutachter und Betreuer Dr. Guillermo Aguilar für seine motivierende Art und unterstützende Begleitung bei dieser Arbeit. Ich habe viel durch ihn lernen dürfen. Bei Prof. Dr. Marianne Maertens möchte ich mich dafür bedanken, dass sie mir ermöglicht hat, meine Bachelorarbeit an Ihrem Lehrstuhl Computational Psychology zu schreiben, und bei Prof. Dr. Marc Alexa dafür, dass er sich als Zweitgutachter engagiert hat. Darüber hinaus danke ich den zahlreichen Kommilitonen, mit denen ich die Freude hatte, dieses Studium gemeinsam absolvieren zu können, besonders meinem guten Freund Paul Darius. Abschließend möchte ich mich bei meiner Familie, meinem Freund und meinen Freunden bedanken, die mich immer begleitet und unterstützt haben.

Inhaltsverzeichnis

Liste der Abkürzungen

DL	Distortion Level
DMOS	Difference Mean Opinion Score
ITU	International Telecommunication Union
MLDS	Maximum Likelihood Difference Scaling
MOS	Mean Opinion Score
QoS	Quality of Service
QoE	Quality of Experience
sRGB	Standard-RGB-Farbraum

1 Einleitung

Digitale Bilder und Videos haben sich in den letzten Jahren zu einem zentralen Element sozialer Interaktion und Mediennutzung entwickelt. 2017 wurden allein mit Smartphones rund 1,2 Milliarden digitale Bilder erfasst (Zhai & Min, 2020). Inzwischen nutzen weltweit mehr als 1 Milliarden Menschen soziale Medien und Streamingdienste (Abb. 1) (Statista, 2021).

Bilder und Videos sind damit zu einem essenziellen Bestandteil der digitalen Kommunikation geworden. Erfassung und Speicherung, Bearbeitung und Komprimierung, Übertragung und Wiedergabe der Bilder führen dabei zu einer fortlaufenden Veränderung des Bildmaterials. Die einzelnen Veränderungen sind hierbei von sehr unterschiedlicher Art und Ausprägung. Sie haben wesentlichen Einfluss auf die vom Nutzer[1] wahrgenommene Bildqualität[2]. Um das Nutzererlebnis veränderter Bilder mit optimaler Qualität gestalten zu können, sind daher Qualitätskennzahlen zu den visuellen Inhalten erforderlich (Zhai & Min, 2020). Heutzutage werden Qualitätskennzahlen durch Algorithmen generiert, die Vorhersagen über die vom Nutzer wahrgenommene Qualität des visuellen Inhalts machen. Grundlage dieser Vorhersagen sind Datensätze aus Wahrnehmungsexperimenten, die visuelle Inhalte mit gemessenen Kennzahlen subjektiver Bildqualität verknüpfen. Mit Hilfe dieser Datensätze können die Algorithmen getestet und weiter-

[1] Aus Gründen der besseren Lesbarkeit wird auf die gleichzeitige Verwendung der Sprachformen männlich, weiblich und divers (m/w/d) verzichtet. Sämtliche Personenbezeichnungen gelten gleichermaßen für alle Geschlechter.

[2] In dieser Arbeit stehen die Begriffe Qualität, Bildqualität, Bildeindruck etc. synonym für den wahrgenommen Qualitätseindruck einer Person beim Betrachten eines Bildes.

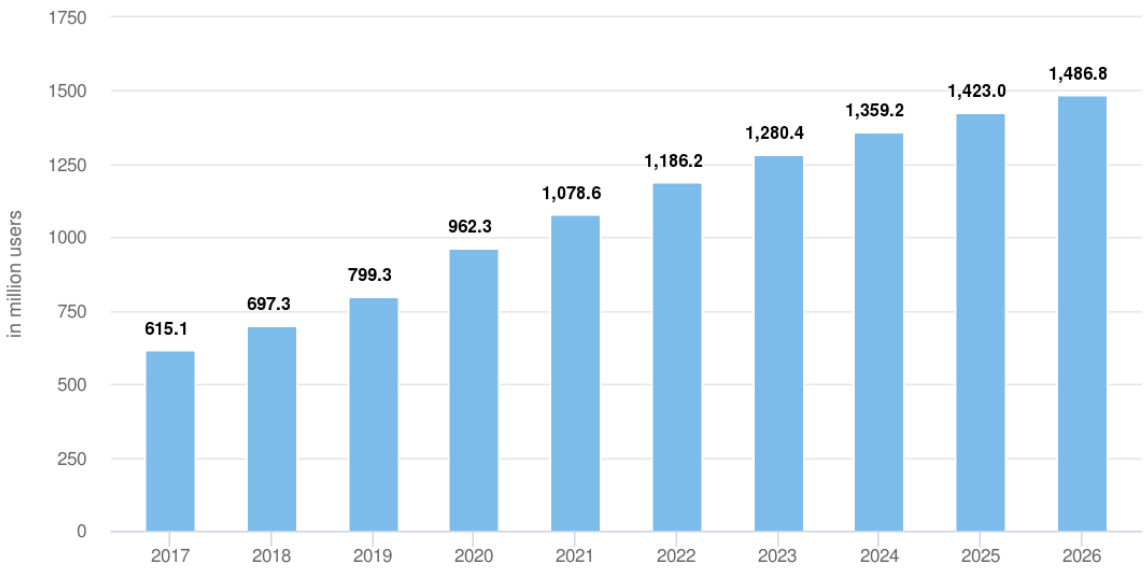


Abb. 1: Weltweite Anzahl an Nutzern von Video Streaming

entwickelt werden. Dies geschieht in der Annahme, dass die Datensätze repräsentativ für die vom Menschen wahrgenommene subjektive Bildqualität sind. Die vorliegende Arbeit untersucht diese Annahme.

Das folgende Kapitel führt in die für diese Arbeit relevanten Grundlagen der subjektiven Wahrnehmung visueller Inhalte ein. Die Bedeutung von Bildqualität, Bildqualitätsalgorithmen und Bildqualitätsdatenbanken wird erläutert. Die psychooptischen Bewertungsansätze des *Mean Opinion Score* (MOS) und des *Maximum Likelihood Difference Scaling* (MLDS) werden vorgestellt. Der Abschnitt Fragestellung geht auf die Motivation und das Untersuchungsziel dieser Arbeit ein.

1.1 Bildqualität

Der Goldstandard der Bildqualität ist die subjektive Wahrnehmung und Bewertung durch den Betrachter (Schade, 1975, zitiert nach Chandler, 2013). Die Herausforderung besteht darin, diese subjektive Wahrnehmung messbar zu machen. Erst durch die Messbarkeit der subjektiven Wahrnehmung können Bilder mit einer Qualitätsmetrik bewertet werden. Der Begriff Bildqualität wird also durch die subjektive visuelle Wahrnehmung bestimmt (Bosse, 2018).

Visuelle Wahrnehmung gehört wiederum zum Forschungsgebiet der Neurobiologie und experimentellen Psychologie. Visuelle Neurowissenschaft ist ein eigenes Forschungsgebiet, das sich mit der Verarbeitung visueller Stimuli durch das Auge und Gehirn beschäftigt (Kandel, Koester, Mack & Siegelbaum, 2021). Die Messung der visuellen Wahrnehmung ist dabei besonders komplex und nicht trivial (Haas, Hass, Spocter & de Sousa, 2020). Das Ziel ist es, die subjektiv visuell wahrgenommene Bildqualität eines Bildes zu erfassen und mit anderen Bilder zueinander in Relation zu setzen. Hierfür ist es wichtig, die Parameter zu kennen, die sich auf die Qualität eines Bildes auswirken können. Hierbei lassen sich einerseits physikalische, anderseits psychische Aspekte als übergeordnete Kategorien unterscheiden. Es gibt zahlreiche Studien, die sich besonders mit den physikalischen Parametern wie beispielsweise Luminanz, Kontrast und spektraler Zusammensetzung eines Bildes auseinander setzen. Daneben gibt es Untersuchungen zu psychischen Parametern wie der Erwartungshaltung, dem Bildinhalt oder dem Kontext, in dem ein Bild betrachtet wird (Fiedler, Hossfeld & Tran-Gia, 2010). Die Psychophysik erforscht sowohl die Auswirkung physikalischer als auch psychischer Parameter auf die Wahrnehmung. Die Wechselwirkung dieser Parameter und ihre Auswirkung auf die dabei gemessene Bildqualität werden untersucht (Goldstein & Cacciamani, 2021).

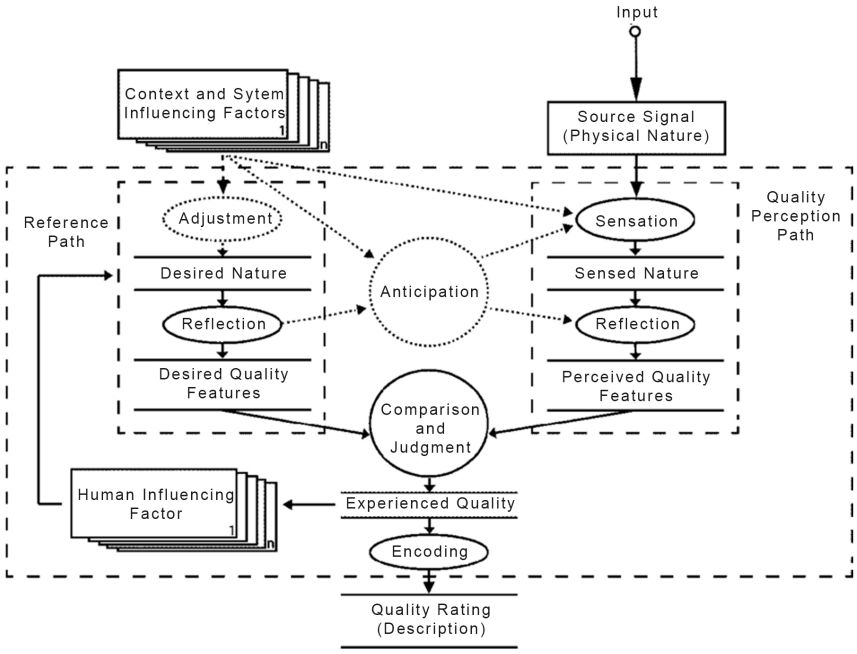


Abb. 2: Abbildung der einzelnen Schritte eines Qualitätsbewertungsprozesses (Quality-formation-process). In der rechten Hälfte des Bildes ist die Verarbeitung des zu bewertenden Inputs dargestellt. Die linke Hälfte repräsentiert beeinflussende Faktoren wie beispielsweise Kontext und Erwartungshaltung (Abb. aus Brunnström et al., 2013)

Diese beiden Aspekte von Bildqualität spiegeln sich auch in den Ansätzen Quality of Service (QoS) und Quality of Experience (QoE) zur Klassifizierung von Qualitätsmerkmalen wieder. Der ursprüngliche Ansatz des QoS betrachtet die physikalischen Eigenschaften eines Bildes wie Signal-Rausch-Abstand, Auflösung, Farbraum (Nahrstedt, 2012). Im Zuge der zunehmenden Digitalisierung wurde zusätzlich der Begriff QoE eingeführt. QoE versucht über die physikalischen Parameter hinaus, psychische Qualitätsmerkmale wie beispielsweise die Nutzerzufriedenheit zu erfassen (Brunnström et al., 2013; Fiedler et al., 2010). Mit Hilfe dieser Ansätze zu Qualitätsmerkmalen wird an Lösungen zur Bildqualitätsmessung gearbeitet.

Besonders anspruchsvoll ist es, Verfahren zu entwickeln, die gemessene Qualitätskennzahlen einzelner Bilder darüber hinaus zueinander in Relation setzen (Bosse, 2018). Abbildung 2 zeigt den Prozess und die komplexen Einflussfaktoren, die die wahrgenommene Bildqualität bestimmen.

Zur experimentellen Aufklärung dieser Faktoren werden psychovisuelle Wahrnehmungsexperimente durchgeführt. Hierbei werden Bilder beispielsweise mit verlustbehafteten

Kompressionsalgorithmen oder durch die Überlagerung von weißem Rauschen verändert. Der Grad der Veränderung wird von geringfügig bis extrem variiert. Gleichzeitig wird die jeweilige subjektive Bildqualität durch Versuchspersonen bewertet (Brunnström et al., 2013). Das Ziel dieser Experimente ist es, psychovisuelle Differenzskalen zu entwickeln, bei denen die Metrik der gemessenen Qualitätskennzahlen den subjektiven Abständen der Merkmalsausprägungen entspricht (Chandler, 2013).

1.2 Bildqualitätsalgorithmen

Im vorherigen Abschnitt ist deutlich geworden, wie stark die Bildqualität von der subjektiven visuellen Wahrnehmung des Betrachters bestimmt wird. Gleichzeitig kann die Qualitätsbewertung bildverarbeitender Systeme auf Grund des Zeit- und Kostenaufwands nicht durch einzelne Versuchspersonen erfolgen. Die Aufgabe der permanenten Qualitätskontrolle der über das Internet übertragenen Bildmengen ist viel zu umfangreich und zu komplex. Darüber hinaus nehmen die Anwendungsbereiche und die Datenmengen der Bildverarbeitung kontinuierlich immer weiter zu.

Auf Grund dieser Anforderungen werden sogenannte Bildqualitätsalgorithmen entwickelt, um eine automatisierte Qualitätsbewertung unabhängig von Versuchspersonen zu ermöglichen. Ihre Aufgabe besteht darin, eine Qualitätskennzahl zu einem Bild zu berechnen, die der subjektiven visuellen Wahrnehmung eines Betrachters möglichst nahe kommt. Die berechnete Qualitätskennzahl wird auch als die *objektive* Bildqualität bezeichnet. Die algorithmische, objektive Bildqualität versucht also, die subjektive Qualitätsbewertung des Bildes durch Versuchspersonen zu ersetzen (Chandler, 2013).

Bildqualitätsalgorithmen werden dabei danach unterschieden, ob zusätzlich zum evaluierten Bild weitere Informationen bereitgestellt werden. Vollreferenzalgorithmen erhalten zusätzlich neben dem zu evaluierenden Bild das nicht verarbeitete Referenzbild (Atidel, Bouridane, Viennet & Haddadi, 2013). Referenzlose Algorithmen versuchen eine Qualitätsbewertung nur anhand des evaluierten Bildes.

Bevor solche Algorithmen jedoch zur Evaluierung von Bildqualität eingesetzt werden können, müssen sie vorab normiert und validiert werden. Hierzu wird überprüft, wie genau die vorhergesagte objektive Bildqualität mit einer bekannten subjektiven visuellen Wahrnehmung durch Versuchspersonen übereinstimmt. Grundlage hierfür sind Bildqualitätsdatenbanken, die subjektive visuelle Bewertungen von Versuchspersonen zu den Bildern der Datenbanken beinhalten. Bildqualitätsalgorithmen werden anhand dieser Datenbanken trainiert, evaluiert und kontinuierlich angepasst. Stimmen die Vorhersagen eines Algorithmus mit den Qualitätsbewertungen der subjektiven Datenbanken überein, wird der Algorithmus als sehr gut eingestuft. Bei der Bewertung eines

Algorithmus wird besonders darauf geachtet wie gut seine Vorhersagekonsistenz, Vorhersagegenauigkeit und Vorhersagemonotonie in Bezug auf die Qualitätsdatenbanken ist (Lahoulou, Viennet, Bouridane & Haddadi, 2011).

1. Die Vorhersagekonsistenz gibt an, wie häufig die algorithmische objektive Bildqualität mit den subjektiven Qualitätsbewertungen der Versuchspersonen übereinstimmt. Man untersucht das sogenannte Ausreißerverhältnis (Outlier Ratio) (Gl. 1). Das Ausreißerverhältnis gibt das Verhältnis falscher objektiver Vorhersagen zur Anzahl aller Vorhersagen an. Eine algorithmische Qualitätskennzahl wird dabei als falsch bewertet, wenn sie einen festgelegten Akzeptanzbereich um die subjektive Qualitätskennzahl der Datenbank überschreitet. Befindet sich die vorhergesagte Qualitätsbewertung innerhalb dieses Bereichs, wird sie als übereinstimmend betrachtet. Zusätzlich berechnet man meistens den Abstand einer falschen Vorhersage zum Akzeptanzbereich (VQEG, 2000; Wang, Lu & Bovik, 2004).

$$R_{\text{out}} = \frac{N_{\text{false}}}{N_{\text{total}}} \tag{1}$$

2. Die Vorhersagegenauigkeit gibt Auskunft über das Ausmaß der Abweichung der algorithmischen Vorhersage im Vergleich zur subjektiven Qualitätskennzahl der Datenbank (VQEG, 2000).

3. Bei der Vorhersagemonotonie vergleicht man die Rangfolgen der subjektiven und objektiven Qualitätsbewertungen zu mehreren Bildern. Die Rangfolge ist die Reihenfolge der nach Qualitätskennzahl sortierten Bilder. Je nach Berechnungsverfahren, ist das Wissen über die originalen Bewertungen der Versuchspersonen der Datenbankbilder notwendig. Nicht alle Bildqualitätsdatenbanken stellen diese zur Verfügung (Streijl, Winkler & Hands, 2014).

Bei der Vorhersagekonsistenz, -genauigkeit und -monotonie werden jeweils die objektiven maschinellen Qualitätskennzahlen mit den subjektiven Qualitätsbewertungen der Datenbanken verglichen. *Ground truth* sind hierbei immer die subjektiven Qualitätsbewertungen aus den Datenbanken. Die Entwicklung, Evaluierung und Anpassung der Qualitätsalgorithmen basiert also auf den Daten der subjektiven Bildqualitätsdatenbank (Chandler, 2013).

Abb. 3: Beispiele von Verzerrungen aus einer subjektiven Bildqualitätsdatenbank. Links ein Ausschnitt des unverzerrten Referenzbildes *WomanHat* aus der LIVE2 Datenbank, rechts daneben verschiedene Verzerrungsarten und Verzerrungsgrade (Martinez-Garcia et al., 2018).

1.3 Subjektive Bildqualitätsdatenbanken

Subjektive Bildqualitätsdatenbanken beinhalten Bilder und ihre zugehörigen subjektiven Qualitätskennzahlen. Diese subjektiven Bewertungen sind Ergebnisse psychophysischer Wahrnehmungsexperimente. Unverzerrte, sogenannte Referenzbilder werden auf verschiedene Arten verzerrt und anschließend durch Versuchspersonen bewertet. Als Verzerrungsarten kommen Kompressionsalgorithmen, Glätten, Rauschüberlagerung und Ähnliches zu Anwendung. Um ein großes Qualitätsspektrum an Bildern von sehr schlechter bis sehr guter Bildqualität zu generieren, werden unterschiedliche Verzerrungsgrade zur selben Verzerrungsart verwendet. Darüber hinaus werden die Motive der Bilder vielfältig variiert. Abbildung 3 zeigt exemplarische Auswirkungen von Verzerrungen eines Referenzbildes (Martinez-Garcia, Bertalmío & Malo, 2018).

Die LIVE2 Datenbank ist ein bekanntes, weit referenziertes Beispiel einer der subjektiven Bildqualitätsdatenbanken. Zum Zeitpunkt der Veröffentlichung 2005 galt sie als die größte Bildqualitätsdatenbank in Bezug auf Anzahl der Bilder, der Verzerrungsarten und der Bewertungen durch Versuchspersonen pro Bild. LIVE2 umfasst insgesamt 29 Referenzbilder und 779 verzerrte Bilder mit den Verzerrungsarten *JPEG2000*, *JPEG*, *Gaussian Blur*, *White Noise*, *Fast Fading*. Aus den rund 25000 Bewertungen durch Versuchspersonen wurden für alle Bilder subjektive Qualitätskennzahlen berechnet, die damals wie heute zur Evaluierung von Bildqualitätsalgorithmen eingesetzt werden.

Die Qualitätsdatenbanken stehen häufig in der Kritik, dass die verwendeten Bildinhalte, Verzerrungsarten und Verzerrungsgrade ein zu geringes Spektrum an Bildqualitä-

ten untersuchen (Martinez-Garcia et al., 2018). Zusätzlich werden die meist im Labor durchgeführten Wahrnehmungsexperimente als zu anwendungsfern bewertet. Außerdem wird kritisiert, dass die subjektiven Qualitätsbewertungen zu stark vom Design des Experiments und der Art der verwenden Stimuluspräsentation beeinflusst zu werden (Streijl et al., 2014).

1.4 Mean Opinion Score (MOS) als Maß subjektiver Bildqualität

Die Quantifizierung und der Vergleich subjektiver Bildqualität erfordert eine einheitliche Kennzahl zu wahrgenommenen Qualität eines Bildes. Hierbei hat sich der MOS als Maß subjektiver Qualität durchgesetzt (Streijl et al., 2014).

Zur Messung des MOS wird ein Bild als einzelner Stimulus präsentiert und von Versuchspersonen bewertet. Eine der in diesen Experimenten am häufigsten verwendeten Bewertungsskalen ist eine 5-stufige Likert-Skala, bei der die Bildqualität durch die Begriffe *bad*, *poor*, *fair*, *good* und *excellent* gekennzeichnet wird (Bosse, 2018; Streijl et al., 2014). Anschließend wird der Mittelwert aller Bewertungen des Bildes über alle Versuchspersonen berechnet. Wurde das unveränderte Referenzbild ebenfalls in die Bewertung eingeschlossen kann zusätzlich der Difference Mean Opinion Score (DMOS) (Gl. 2) berechnet werden (Bosse, 2018). Dabei kennzeichnet *ref* das Referenzbild, i den Index des Verzerrungsgrades eines veränderten Bildes.

$$DMOS_i = MOS_{ref} - MOS_i \tag{2}$$

Damit erhält das unveränderte Referenzbild einen DMOS von null. Bei allen verzerrten Bildern gibt der DMOS den subjektiven Qualitätsabstand zum Referenzbild wieder.

Auch wenn MOS sich als Bildqualitätsmaß in unterschiedlichen Anwendungsbereichen breit etabliert hat, wird MOS stark kritisiert. MOS wird oft verwendet, ohne ausreichend zu berücksichtigen, wie das Design und die Durchführung eines Wahrnehmungsexperiments gestaltet wurden.

Auch der Bildungsgrad, das soziale Umfeld, das Geschlecht und das Alter einer Versuchsperson können zu deutlichen Abweichungen bei der Qualitätsbewertung führen (Streijl et al., 2014). Gleichzeitig prägt die verwendete Bewertungsskala die gemessenen Daten und beeinflusst Effekte wie die Tendenz-zur-Mitte. Darunter versteht man die

Neigung von Versuchspersonen, bei mehrstufigen, ordinalen Skalen vermehrt die mittleren Skalenwerte auszuwählen (Zerman, Hulusic, Valenzise, Mantiuk & Dufaux, 2018). Um solchen Einflussfaktoren auf die MOS-Werte entgegenzuwirken, werden Standards für das Design von Wahrnehmungsexperimenten formuliert. Die International Telecommunication Union (ITU) veröffentlicht hierzu jährlich Richtlinien zu den maßgeblichen Einflussfaktoren für Wahrnehmungsexperimente (ITU, 2019). Die Richtlinien beschreiben Aspekte wie zum Beispiel:

- Einzel-, Doppel- oder Mehrfachstimulus: Anzahl der Bilder, die gemeinsam präsentiert werden.
- Wiederholungen: Anzahl der Bewertungen des selben Stimulus durch die selbe Versuchsperson.
- Bereitstellen eines Referenzbildes: Die Versuchsperson wird darüber aufgeklärt, dass es sich um das Referenzbild handelt. Alternativ wird das Referenzbild ohne Kenntlichmachung gezeigt.
- Art der Bewertung: Zuweisung von numerischen Werten, Vergleiche mit oder ohne Referenzbild inbegriffen.
- Interaktivität des Abstimmungsprozesses: Individuelle oder parallele Bewertung durch Versuchspersonen.
- Zeitlicher Ablauf: Zeitdiskrete Bewertung (eine Bewertung pro Stimulus) oder kontinuierlich (eine Bewertung pro Zeitintervall).
- Verwendung von Skalenankern: Verfügbarkeit von hohen oder niedrigen Stimuli vorab oder permanent zur Kalibrierung der Bewertungsskala (Streijl et al., 2014).

Die in solchen Wahrnehmungsexperimenten gemessen MOS-Werte bilden die Grundlage der subjektiven Bildqualitätsdatenbanken (Kapitel 1.3). Der in solchen Datenbanken erfasste MOS wird auch subjektiver MOS genannt. Demgegenüber werden algorithmische MOS-Bewertungen als objektiver MOS bezeichnet (Wang, Bovik & Lu, 2002; Streijl et al., 2014).

1.5 Maximum-Likelihood-Difference-Scaling (MLDS)

Auf Grund der Kritik am MOS werden weiterführende Qualitätsmaße untersucht. Im Einzelstimulus-Verfahren zur Berechnung eines MOS-Wertes wird eine exklusive Bewertung jedes einzelnen Bildes abgegeben. Eine Herausforderung der Psychophysik

besteht aber darin, nicht nur die Bildqualitätsbewertungen eines einzelnen Bildes zu erfassen, sondern die Qualität mehrerer Bilder vergleichbar zu machen. Ein Ansatz dazu ist, Versuchspersonen aufzufordern, eine Reihe von Bildern von guter bis schlechter subjektiver Qualität zu sortieren. Das Ziel ist es, intervallskalierte Differenzskalen zu entwickeln, die nicht nur zeigen, dass sich zwei Bilder unterscheiden, sondern auch wie stark sie sich unterscheiden (Knoblauch & Maloney, 2008; Krantz, Luce, Suppes & Tversky, 2006). Ein solches Verfahren ist das Maximum-Likelihood-Difference-Scaling (MLDS).

Beim MLDS handelt es sich im Gegensatz zum MOS um ein Multistimulus-Verfahren. Der Versuchsperson werden drei Bilder (Triade) gleichzeitig präsentiert. Die Versuchsperson wird gebeten, zwei Bildpaare (a, b) und (b, c) miteinander zu vergleichen. Es soll bewertet werden, ob der wahrgenommene Qualitätsunterschied von a zu b als stärker empfunden wird als der von b zu c. Durch diesen Ansatz werden Qualitätsintervalle gemessen und die Qualitäten vergleichbar gemacht. Anwendungsbereiche, in denen MLDS erfolgreich eingesetzt wurde, sind Bildqualität (Charrier, Maloney, Cherifi & Knoblauch, 2007), Farbunterschiedscharakterisierung (Lindsey et al., 2010), Materialtransparenz (Fleming, Jäkel & Maloney, 2011) oder Helligkeit (Wiebel, Aguilar & Maertens, 2017).

1.6 Fragestellung

Die vorangegangenen Abschnitte machen deutlich, wie wichtig die algorithmische Bewertung von Bildqualität im Zeitalter der digitalen Kommunikation geworden ist. Dabei wird die Entwicklung und Normierung von Bildqualitätsalgorithmen von den Bildqualitätsdatenbanken geprägt. Die Eignung einzelner Bildqualitätsdatenbanken wird wiederum zunehmend kritisch hinterfragt. Eine Kritik ist die Verwendung des MOS als Maßstab der Bildqualität. Meistens wird bei der Verwendung von MOS-Werten nicht ausreichend berücksichtigt, wie die MOS-Daten gemessen wurden. Das experimentelle Design und die Ermittlung der subjektiven Qualitätsdaten kann sich dabei stark unterscheiden. Zusätzlich sind Likert-Skalen, auf denen MOS häufig basiert, ordinal und semantisch attribuiert. Sie geben daher keine Auskunft über die Qualitätsintervalle zwischen den Skalenabschnitten. Dennoch hat MOS sich als Bewertungsmaßstab für die Medienqualität weitreichend durchgesetzt (Bosse, 2018; Streijl et al., 2014). Tatsächlich wird subjektiver MOS meistens für das Testen und Evaluieren von Bildqualitätsalgorithmen eingesetzt (Streijl, Winkler & Hands, 2010). Es ist daher nicht überraschend, dass die Bewertung eines Bildqualitätsalgorithmus deutlich von der verwendeten Datenbank abhängt, anhand derer die objektiven MOS-Werte mit den subjektiven MOS-Werten normiert wurden (Lahoulou et al., 2011).

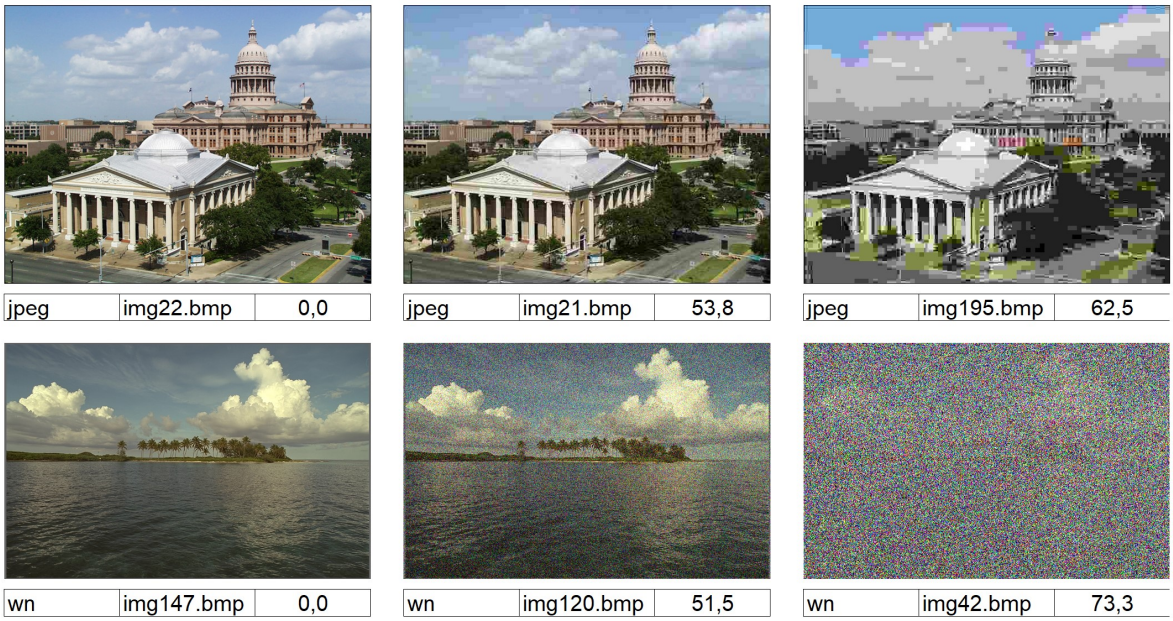


Abb. 4: Beispiele aus der LIVE2 Datenbank mit zunehmendem Grad der Verzerrung von links nach rechts. In der linken Spalte ist das Referenzbild ohne Verzerrung dargestellt. Die DMOS-Skala reicht von null bis 100. Zu jedem Bild sind von links nach rechts die Informationen Verzerrungsart (*JPEG2000*, *JPEG*, *White Noise*, *Gaussian Blur*, *Fast Fading*), originaler Dateiname in der LIVE2 Datenbank und DMOS-Wert angegeben.
Die Bilder zeigen Beispiele für DMOS-Intervalle, die nicht mit dem subjektiven Bildeindruck übereinstimmen. Konkret bedeutet das, dass der subjektiv wahrgenommene Bildunterschied vom linken zum mittleren Bild geringer ist, als der vom mittleren zum rechten Bild. Gleichzeitig ist der numerische Abstand der DMOS-Werte des linken zum mittleren Bild sehr viel größer als der Abstand der DMOS-Werte des mittleren zum rechten Bild.

Vor dem Hintergrund dieser Kritikpunkte untersucht die vorliegende Arbeit zunächst die Qualität und Konsistenz der LIVE2 Datenbank im Detail. Bei der Betrachtung mehrerer Bilder nebeneinander ist besonders auffällig, dass die Abstände der DMOS-Werte häufig nicht den Abständen der subjektiven Qualitätswahrnehmung der Bilder entsprechen (Abb. 4). Auf Grund dieser Beobachtungen wurde das Thema der vorliegenden Arbeit formuliert: *Untersuchung zu MOS als geeignetes Maß für die subjektive Bildqualität der LIVE2 Datenbank.* Auf Grund des Umfangs der Untersuchung betrachtet die vorliegende Arbeit ausschließlich die Daten der LIVE2 Datenbank. Erst durch den Vergleich der LIVE2 Bilder nebeneinander wurden die Unstimmigkeiten der DMOS-Werte erkennbar. Daraufhin wurde die LIVE2 Datenbank systematisch auf weitere Auffälligkeiten hin untersucht. Um ein besseres Verständnis der subjektiven Bildqualität zu bekommen, wurden im Weiteren zwei eigene Experimente mit einer Auswahl an LIVE2 Bildern konzipiert und durchgeführt. Hierbei kamen ein Einzelstimulus MOS-Verfahren und ein vergleichendes MLDS-Verfahren zu Anwendung.

2 Untersuchung des LIVE2 DMOS als Metrik der subjektiven Bildqualität

Die Frage nach dem MOS als geeignetem Maß der subjektiven Qualität für die Bilder der LIVE2 Datenbank führt zunächst zu einer Untersuchung der öffentlich verfügbaren DMOS-Bewertungen dieser Bilder. Im folgenden Abschnitt wird dazu zunächst das methodische Vorgehen bei dieser Untersuchung sämtlicher LIVE2 Bilder dargelegt. Daran anschließend werden die Ergebnisse der Untersuchung ausgeführt.

2.1 Methode

Die LIVE2 Datenbank wurde im Jahr 2005 vom Laboratory for Image & Video Engineering der University of Texas als Erweiterung ihrer LIVE Datenbank aus 2003 veröffentlicht. Sie kann als subjektive Bildqualitätsdatenbank unter `http://live.ece.utexas.edu/research/quality` abgerufen werden. LIVE2 stellt eine Kollektion von 29 Bildmotiven (Abb. 5) bereit. Die Auflösung der einzelnen Bilder bewegt sich zwischen 480 x 720 und 768 x 512 Pixeln. Die Bilder haben eine 24-Bit/Pixel Farbkodierung aus dem sRGB Farbraum. Jedes Bildmotiv wird in den Verzerrungsarten *JPEG2000*, *JPEG*, *White Noise*, *Gaussian Blur* und *Fast Fading* mit jeweils fünf bis acht Verzerrungsgraden bereitgestellt. Bei *JPEG2000* und *JPEG* reduziert die Verzerrung die Bitrate des Bildes. Bei *White Noise* wird weißes gaussverteiltes Rauschen auf allen Farbkanälen des Bildes überlagert, bei *Gaussian Blur* werden die Bilder mit einem kreissymmetrischen Gauss-Filter verändert. *Fast Fading* schließlich verzerrt die Bilder mit Bit-Fehlern unterschiedlicher Signal-Rausch-Abstände. Die Bildqualität der einzelnen Bilder wurde für die LIVE2 Erhebung in Einzelstimulus-Präsentationen von 20 – 29 Versuchspersonen auf einer Skala mit den fünf semantischen Kategorien *bad*, *poor*, *fair*, *good* und *excellent* bewertet. Diese Bewertungen wurden in Bezug zur Bewertung des nicht verzerrten Referenzbildes über alle Versuchspersonen zu einem DMOS zwischen null und 100 transformiert. Das Referenzbild erhält dabei einen DMOS von null, ein Bild mit der Bewertung *bad* durch alle Versuchspersonen einen DMOS von 100. Die LIVE2 Datenbank gibt zu jedem Tupel (Motiv, Verzerrungsart, Verzerrungsgrad) den zugehörigen DMOS an. Die Rohdaten der Einzelbewertungen durch die Versuchspersonen werden nicht zur Verfügung gestellt.

Für die systematische Untersuchung der Datenbank wurden alle verfügbaren Angaben der Bilder in einem Spreadsheet (`https://github.com/AnnalenaSchillen/LIVE2_MOS_BA`) zusammengefasst und ergänzt (Abb. 6). Zur Aufbereitung kamen sowohl LibreOffice Calc als auch Microsoft Excel zur Anwendung.

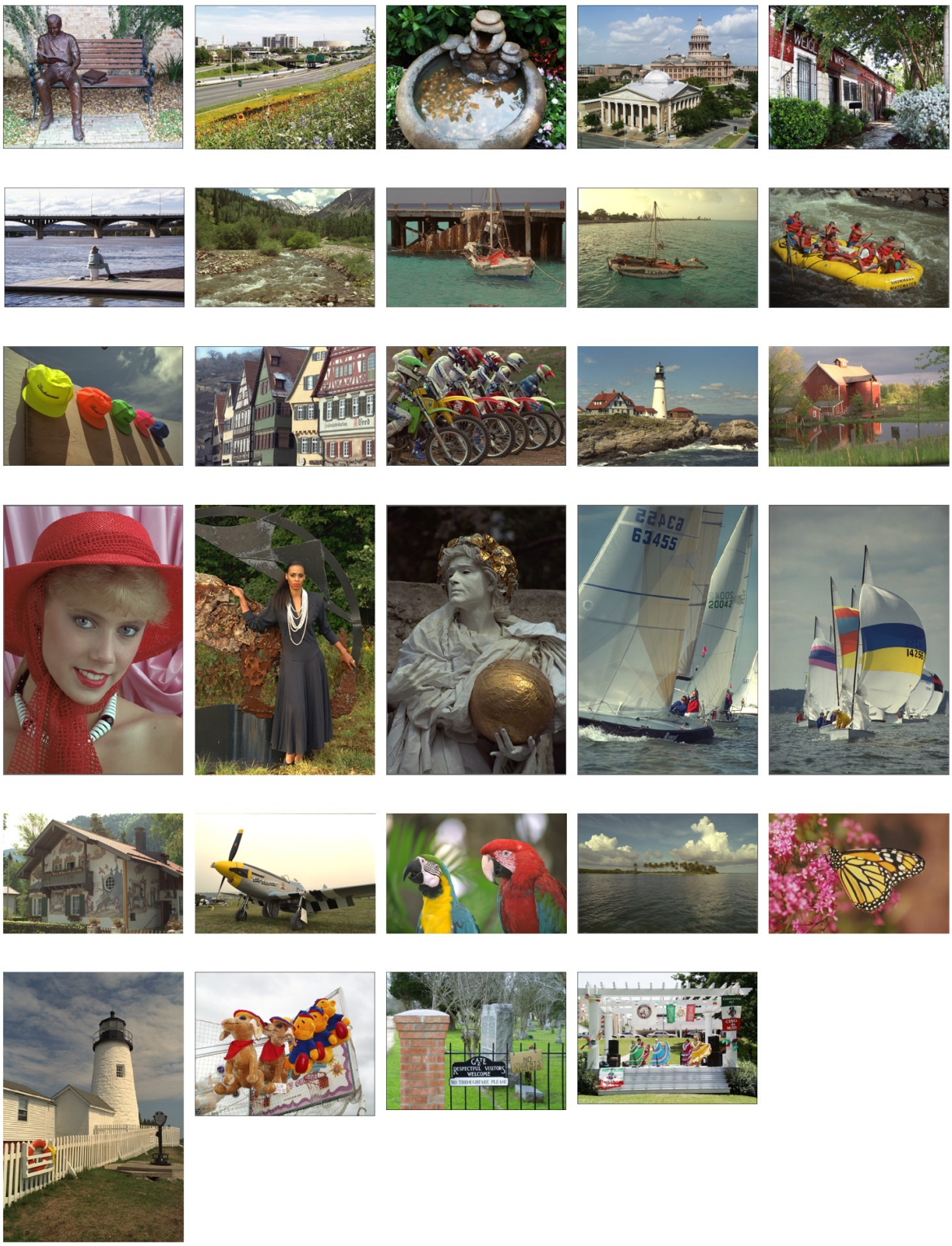


Abb. 5: Die 29 Referenzbilder der LIVE2 Datenbank

Key	motive	filename	distType	distortionLevel	filepath	width	height	DMOS.LIVE2	
bikes.jp.0,209	bikes	img183.bmp	jpeg	0,20876	D:\BA\LIVE2\databaserelease2\jpeg\img183.bmp	768	512	56,821	
churchandcapitol.wn.1,9	churchandcapit	img20.bmp	wn	1,99609	D:\BA\LIVE2\databaserelease2\wn\img20.bmp	634	505	75,678	
ocean.gb.0,792	ocean	img106.bmp	gblur	0,79163	D:\BA\LIVE2\databaserelease2\gblur\img106.bmp	768	512	28,227	
womanhat.ff.18,900	womanhat	img82.bmp	fastfading	18,90000	D:\BA\LIVE2\databaserelease2\fastfading\img82.bmp	480	720	40,450	

Abb. 6: Auszug Spreadsheet-Organisation: Integration von Bildparametern und Bildern der LIVE2 Datenbank als Datensätze in einem Spreadsheet. Die Aufbereitung als Spreadsheet unterstützt die systematische Analyse aller Bilder der Datenbank. Die Filterfunktion erlaubt die Gegenüberstellung aller Bilder ausgewählter Motive, Verzerrungsarten, Verzerrungsgrade und DMOS-Intervalle. Die Sortierfunktion ermöglicht die Untersuchung von Bildsequenzen sortiert nach Verzerrungsgrad, Verzerrungsrang oder DMOS-Bewertung.

Das Spreadsheet gibt in jeder Datenzeile Motivname, Dateiname, Verzerrungsart, Verzerrungsgrad, Dateipfad, Breite und Höhe des Bildes und den LIVE2 DMOS zusammen mit der Darstellung des Bildes wieder. Ein zusätzliches Datenfeld kennzeichnet alle Duplikatbilder der Datenbank.

Als ein weiteres Datenfeld wurde der Verzerrungsrang zur Darstellung der Sortierung der Verzerrung ergänzt. Dabei erhielten alle nicht verzerrten Referenzbilder den Rang eins. Bei *White Noise* und *Gaussian Blur* wurde der Verzerrungsrang mit aufsteigendem, bei den übrigen Verzerrungsarten mit absteigendem Wert des Verzerrungsgrades inkrementiert. Damit ist für jede Bildserie zu Motiv und Verzerrungsart eine einfache aufsteigende Sortierung der Bildverzerrung vom unveränderten Referenzbild mit Rang eins bis zum am stärksten verzerrten Bild mit dem numerisch höchsten Rang möglich.

Das gesamte Spreadsheet erlaubt damit eine einfache vergleichende Darstellung, Filterung und Sortierung aller Bilder auf Grundlage der genannten Bildparameter. Der Verzerrungsrang ermöglicht außerdem, Abweichungen der Monotonie des DMOS aller LIVE2 Datensätze einfach algorithmisch zu detektieren. Betrachten wir eine Reihe von Bildern i_1, i_2, i_3 mit ansteigendem Verzerrungsrang $r(i_1) < r(i_2) < r(i_3)$, dann bedeutet Monotonie des DMOS, dass dieser mit zunehmendem Verzerrungsrang ebenfalls zunimmt: $\mathrm{DMOS}(i_1) < \mathrm{DMOS}(i_2) < \mathrm{DMOS}(i_3)$. Bei fehlender Monotonie gilt dagegen $\mathrm{DMOS}(i_1) < \mathrm{DMOS}(i_2) > \mathrm{DMOS}(i_3)$.

Zusätzlich wurde eine vergleichende Darstellung der Bildserien zu Motiv, Verzerrungs-

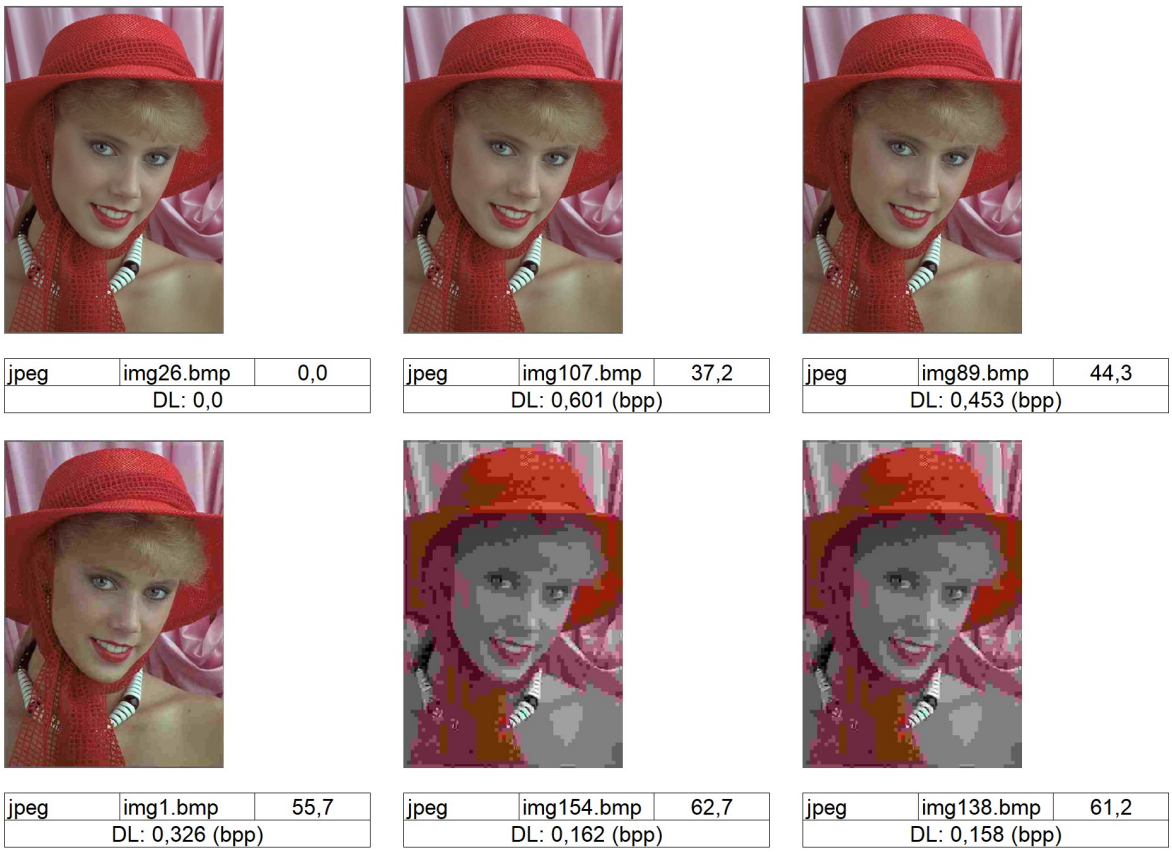


Abb. 7: Auszug Spreadsheet-Übersicht der Bildserie zum selben Motiv, zur selben Verzerrungsart und allen Verzerrungsgraden mit Angabe der verfügbaren Bildparameter aus der LIVE2 Datenbank: Verzerrungsart, Bilddatei, LIVE2 DMOS und Verzerrungsgrad (DL = distortion level). Die Sortierung der Bildserie erfolgt aufsteigend von Verzerrungsrang eins (Referenzbild) zu Verzerrungsrang sechs (stärkste Verzerrung). Das explizite Beispiel zeigt einen Verzerrungsgrad in Bits pro Pixel (bpp).

art und Verzerrungsgrad in einem weiteren Spreadsheet ergänzt (Abb. 7). Die Bilder werden jeweils mit den verfügbaren Informationen Verzerrungsart, originaler Dateiname, Verzerrungsgrad und LIVE2 DMOS dargestellt. Diese Zusammenstellung der vollständigen Bildserien zu Motiv und Verzerrungsart vermittelt für die Untersuchung einen schnellen Überblick zur subjektiven Bildqualität aller verfügbaren Bilder im Vergleich. Auch diese Darstellung kann nach dem Schlüssel (Motiv, Verzerrungsart) gefiltert werden.

Anhand dieser Aufbereitungen im Spreadsheet wurden anschließend alle Datensätze der LIVE2 Datenbank systematisch auf Auffälligkeiten analysiert. Die einzelnen Bilder und ihre zugehörigen DMOS-Werte wurden darauf untersucht, wie stark die eigene, subjektiv wahrgenommene Bildqualität des Untersuchers mit dem LIVE2 DMOS korreliert. Die Möglichkeiten zum Filtern der Datensätze des Spreadsheets wurden eingesetzt, um das Spektrum an Bildern zu verschiedenen DMOS Bereichen gemeinsam darzustellen. Darüber hinaus wurden die DMOS-Werte algorithmisch auf Monotonie des DMOS als

Funktion des Verzerrungsgrades ausgewertet.

2.2 Ergebnisse

Bei der Betrachtung einzelner Bildsequenzen aus der LIVE2 Datenbank und ihrer zugehörigen DMOS-Werte findet man schnell Auffälligkeiten, bei denen die subjektive Metrik der Qualitätsabstände zwischen den Bildern nicht durch den LIVE2 DMOS abgebildet wird. Die Aufbereitung der LIVE2 Bildserien im Rahmen dieser Arbeit (Abb. 6, 7) bestätigt diesen Eindruck. Als Bildserie werden dabei die Bilder aller verfügbaren Verzerrungsgrade zum selben Motiv und zur selben Verzerrungsart betrachtet. Die systematische Untersuchung aller Bildserien zeigt, dass es sich bei den Divergenzen von subjektiver Qualitätsmetrik und DMOS nicht nur um seltene Einzelfälle handelt.

2.2.1 DMOS $\geq$ 70

Nach einer ersten Sichtung aller 145 Bildserien mit der implementierten Seriendarstellung (Abb. 7) wurden im Spreadsheet zunächst alle Einzelbilder mit einem DMOS $\geq$ 70 genauer untersucht. Ein DMOS von null entspricht der Bewertung *excellent*, die alle Referenzbilder erhalten. Einen DMOS von 100 erhält ein Bild mit der Bewertung *bad*, das eine besonders schlechte Bildqualität aufweist. Bei DMOS-Werten zwischen 70 und 100 besteht die Erwartung, fast ausschließlich Bilder mit erheblich reduzierter Wahrnehmungsqualität zu finden. Dies ist jedoch nicht der Fall. Es finden sich insgesamt 44 Bilder mit einem DMOS $\geq$ 70. Davon haben 13 (29,5%) Bilder eine gute bis sehr gute subjektive Qualität (Abb. 8). In diesem Sinne auffällige Bilder finden sich bei den Verzerrungsarten *JPEG2000*, *JPEG* und *Fast Fading*, nicht jedoch bei *White Noise* oder *Gaussian Blur*.

jp2k	img32.bmp	70,2

jpeg	img4.bmp	79,6

fastfading	img31.bmp	74,0

Abb. 8: Bilder mit guter bis sehr guter subjektiver Bildqualität bei DMOS $\geq$ 70.

Demgegenüber zeigt Abbildung 9 die weitreichende Divergenz der subjektiven Qualität bei einem fast gleichen DMOS um 70. Während die Bilder in der oberen Reihe noch

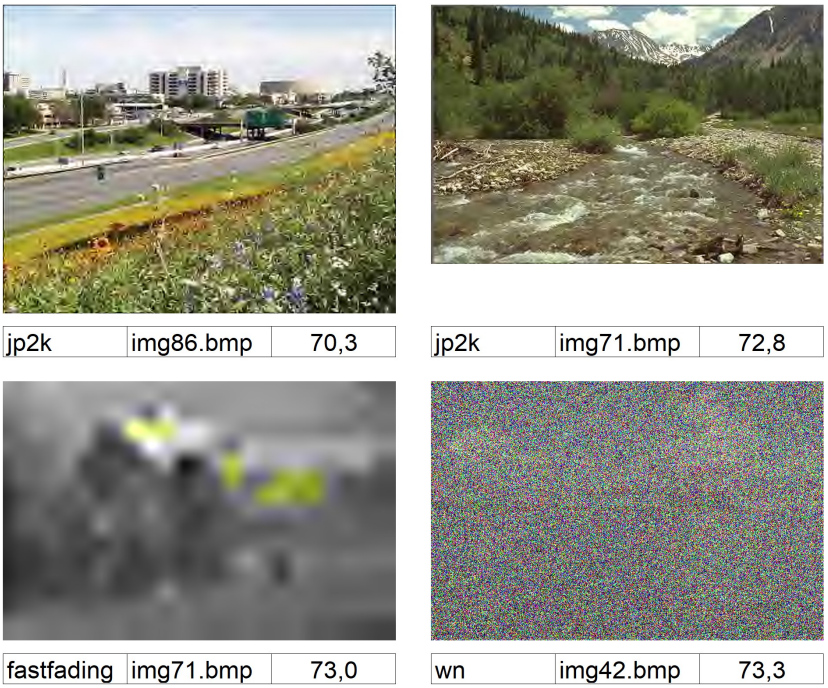


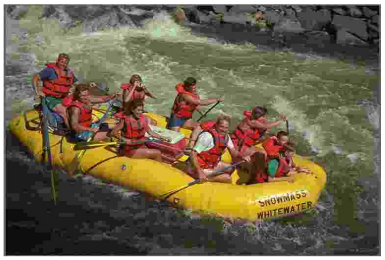
Abb. 9: Bilder mit vergleichbarem DMOS ≈ 70 reichen in der Bildqualität von gut erkennbar bis nicht mal ansatzweise erkennbar.

gut erkennbar sind, lässt sich bei den Bildern der unteren Reihe ohne Kenntnis der Motive nicht mal ansatzweise entnehmen, welcher Inhalt dort dargestellt ist. Dennoch beträgt die maximale DMOS-Differenz dieser Bildauswahl nur drei Punkte.

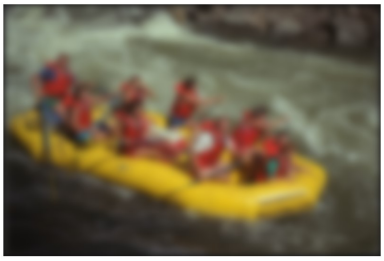
Um einen Eindruck vom Einfluss des Motivs auf diese Effekte zu erhalten, wurde als Nächstes die subjektive Bildqualität für ähnliche DMOS ≈ 70 beim gleichen Motiv untersucht. Abbildung 10 demonstriert dazu das Spektrum der Bildqualität für das Motiv *Rapids* in den Verzerrungsarten *JPEG2000*, *JPEG*, *Gaussian Blur* und *Fast Fading*. Während die obere Reihe mit schlechten DMOS-Werten um 70 eine sehr gute Bildqualität aufweist, trifft dies auf die untere Reihe nicht mehr zu. Dabei ist der dargestellte Inhalt auf dem Bild `img17.bmp` mit *Gaussian Blur* trotz einem DMOS von 80 für die meisten Betrachter noch erkennbar. Dem gegenüber ist dasselbe Bild mit *Fast Fading* ohne Kenntnis des Motivs nicht mehr zuzuordnen, obwohl der DMOS von 73 kaum von der oberen Reihe abweicht.

Die bisherige Untersuchung lässt die Interpretation zu, dass die Divergenz der subjektiven Bildqualität im Wesentlichen durch die Verzerrungsart bestimmt wird. Aus diesem Grund wurde die Datenbank nach Beispielen mit deutlicher Abweichung der Bildqualität beim selben Motiv, derselben Verzerrungsart und vergleichbarem DMOS durchsucht. Abbildung 11 zeigt hierzu zwei Beispiele. In der oberen Reihe zum Motiv *House* unterscheidet sich der DMOS nur um zwei Punkte, in der unteren mit dem Motiv *ManFishing* um 0,8 Punkte zwischen den Vergleichsbildern. Dennoch besteht ein

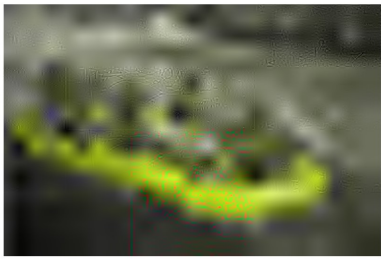
jp2k	img32.bmp	70,2



jpeg	img35.bmp	71,0



gblur	img17.bmp	80,1



fastfading	img76.bmp	73,0

Abb. 10: Bilder zum selben Motiv und verschiedenen Verzerrungsarten zeigen erhebliche Unterschiede in der subjektiven Qualität trotz ähnlichem DMOS ≥ 70.

jpeg	img50.bmp	58,7

jpeg	img160.bmp	61,7

jpeg	img93.bmp	59,8

jpeg	img148.bmp	60,6

Abb. 11: Bilder zum selben Motiv und zur selben Verzerrungsart zeigen erhebliche Unterschiede in der subjektiven Qualität trotz vergleichbarem DMOS.

sehr erheblicher Unterschied beim Vergleich der Bildqualität zwischen der linken und der rechten Spalte.

2.2.2 DMOS Intervalle

In einer weiteren Untersuchung wurde der Divergenz der subjektiven Qualitätsintervalle zu den DMOS-Intervallen anhand von Bildserien nachgegangen. Abbildung 12 zeigt hierzu eine Bildsequenz zu jeder der fünf Verzerrungsarten.

In der linken Spalte ist jeweils das nicht verzerrte Referenzbild dargestellt. In der mittleren Spalte finden sich Bilder mit mittlerem Verzerrungsgrad und DMOS-Werten um 50. Die Qualität dieser Bilder ist erkennbar geringer, aber immer noch sehr gut und sehr nah am Referenzbild. In der rechten Spalte sind die am stärksten verzerrten Bilder der jeweiligen Bildserie dargestellt. Die Qualität dieser Bilder ist deutlich bis erheblich schlechter und hat einen sehr hohen subjektiven Abstand zu den Bildern der mittleren Spalte. Gleichzeitig liegen ihre DMOS-Werte zwischen 60 und 75. Bei allen fünf Sequenzen ist also das subjektive Qualitätsintervall vom Referenzbild links zum Bild in der Mitte sehr gering, vom mittleren zum rechten Bild dagegen erheblich. Diese Qualitätsintervalle kontrastieren zu den DMOS-Abständen von etwa 50 zwischen mittlerem und Referenzbild und DMOS-Intervallen zwischen zehn und 25 vom mittleren zum rechten Bild.

2.2.3 DMOS Monotonie

Bei der Untersuchung der Bildserien wurde ebenfalls festgestellt, dass die DMOS-Werte nicht immer monoton mit dem Verzerrungsgrad der Bilder steigen. Dabei fanden sich Beispiele, bei denen die fehlende Monotonie mit der subjektiven Qualitätsbewertung der Untersucher übereinstimmte, und andere, bei denen dies nicht der Fall war. Abbildung 13 zeigt eine Bildsequenz zum Motiv *PaintedHouse* mit Verzerrungsart *JPEG* sowie eine weitere zum Motiv *Plane* mit *Fast Fading*. Bei beiden Sequenzen steigt der Verzerrungsgrad vom linken zum rechten Bild. Beim *PaintedHouse* fällt die Bit-Rate von 1,45 auf 0,18 bpp. Beim Motiv *Plane* reduziert sich die Signal-to-Noise-Ratio von 26,1 auf 15,5 dB. Bei beiden Sequenzen steigt der DMOS zunächst um etwa 50 vom linken zum mittleren Bild, um dann wieder zehn bzw. 25 Punkt vom mittleren zum rechten Bild abzufallen. Bei der Bildserie *Plane* korrespondiert die fehlende Monotonie des DMOS mit den subjektiven Qualitätsveränderungen in der Abfolge der Bilder, bei der Bildserie *PaintedHouse* dagegen nicht.

Die Verletzungen der Monotonie des DMOS als Funktion des Verzerrungsgrades wurde im Weiteren algorithmisch ausgewertet. Zu jedem Motiv und jeder Verzerrungsart wurden alle Bilder der zugehörigen Bildserie nach ihrem Verzerrungsrang sortiert und dann die Häufigkeit fallender DMOS-Werte bei steigendem Verzerrungsrang ausgewertet.

Abb. 12: Divergenz subjektiver Qualitätsintervalle zu LIVE2 DMOS Intervallen. Beispiele für alle fünf Verzerrungsarten der LIVE2 Datenbank. Der subjektive Qualitätsunterschied vom jeweils linken zum mittleren Bild ist deutlich kleiner als vom mittleren zum rechten. Diese Metrik der subjektiven Qualitätsabstände spiegelt sich jedoch nicht in den Intervallen der LIVE2 DMOS-Bewertungen wider.

Abb. 13: Bildsequenzen, bei denen sich der LIVE2 DMOS als nicht monotone Funktion des Verzerrungsgrades darstellt. Bei beiden Bildserien steigt der Verzerrungsgrad von links nach rechts. Der DMOS steigt vom linken zum mittleren Bild um etwa 50 und fällt vom mittleren zum rechten Bild wieder um zehn bzw. 25 Punkte ab. In der oberen Bildserie mit der Verzerrungsart *JPEG* entspricht dieser Verlauf des DMOS nicht dem subjektiven Qualitätseindruck. Das mittlere Bild (img23.bmp) hat eine schlechtere DMOS-Bewertung als das rechte Bild (img215.bmp), wobei das mittlere Bild (img23.bmp) eine bessere subjektive Bildqualität aufweist als das rechte (img215.bmp). In der unteren Reihe entspricht der Verlauf der DMOS-Bewertungen zwar dem subjektiven Qualitätseindruck, sodass schlechtere Qualitäten eine schlechtere DMOS-Bewertung aufweisen. Gleichzeitig entspricht die subjektiv wahrgenommenen Verzerrung nicht dem Verzerrungsgrad. Das mittlere Bild (img58.bmp) mit geringerem Verzerrungsgrad als das rechte (img56.bmp) hat einen schlechteren subjektiven Qualitätseindruck als das rechte.

Wie Tabelle 1 zeigt, ist die Funktion DMOS(Verzerrungsgrad) bei knapp 9% aller Bilder der LIVE2 Datenbank nicht monoton. Dabei zeigen sich relevante Unterschiede abhängig von der Verzerrungsart. Bei *Fast Fading* ist fast ein Viertel der Bilder betroffen, bei *JPEG2000* und *JPEG* sind es etwa 10%. Bei *White Noise* und *Gaussian Blur* tritt der Effekt praktisch gar nicht auf.

2.2.4 Fast Fading Qualitätsdiskontinuität

Die fehlende Monotonie der DMOS-Werte in der unteren Zeile der Abbildung 13 gibt den subjektiven Qualitätseindruck korrekt wieder. In diesem Fall entspricht die Verletzung der DMOS-Monotonie nicht einem Defizit der MOS-Bewertungen. Ursächlich ist vielmehr eine intermittierende Zunahme der Bildqualität trotz weiter steigendem

Tab. 1: Häufigkeit nicht monotoner DMOS-Werte als Funktion des Verzerrungsgrades aller Bildserien der LIVE2 Datenbank. Fehlende Monotonie bedeutet, dass der DMOS-Wert einer Bildserie nicht kontinuierlich mit zunehmender Verzerrung steigt.

Verzerrungsart	# DMOS Werte	# DMOS nicht monoton	% DMOS nicht monoton
jp2k	198	17	8,6%
jpeg	188	20	10,6%
wn	174	1	0,6%
gblur	174	1	0,6%
fastfading	174	41	23,6%
Total	**908**	**80**	**8,8%**

Verzerrungsgrad.

Abbildung 14 zeigt, dass der Effekt nicht nur vereinzelt auftritt. Er lässt sich vielmehr bei etwa 50% der Bildserien zur Verzerrungsart *Fast Fading* beobachten. Mit zunehmender Verzerrung und monoton abnehmendem Signal-Rausch-Verhältnis nimmt die Bildqualität zunächst ab und dann wieder deutlich zu. Mit weiterer Verzerrung verschlechtert sich die Qualität des Bildes dann wieder (ohne Abbildung). Der Bereich des Signal-Rausch-Verhältnisses, in dem die Besserung der Bildqualität auftritt, variiert zwischen den einzelnen Bildserien.

Bei 16 der 41 *Fast Fading*-Bilder mit einer Verletzung der DMOS-Monotonie (Tab. 1) entspricht die Abnahme des DMOS der Verbesserung der subjektiven Bildqualität durch den beschriebenen Effekt. Bei den anderen 25 Bildern ist dies nicht der Fall.

Bei anderen Verzerrungsarten als *Fast Fading* ist dieser Effekt nicht festzustellen.

2.2.5 DMOS Verteilung in der LIVE2 Datenbank

Bei der Untersuchung der LIVE2 Bilder fällt auf, dass selbst Bilder, die praktisch nicht erkennbar sind, dennoch keinen DMOS im Bereich zwischen 90 und 100 aufweisen. Um einen Eindruck von der Gesamtverteilung der LIVE2 DMOS-Werte zu erhalten, wurde daher die Verteilung der DMOS-Bewertungen ausgewertet (Abb. 15). Die 29 Referenzbilder ohne Verzerrung, mit DMOS = 0 sind bei allen fünf Verzerrungsarten identisch und werden hier ohne Duplikate gezählt. Auch bei den verzerrten Bildern wurden die Duplikatbilder ausgeschlossen. Bei den Bewertungen der verzerrten Bilder ist eine deutliche Tendenz zur Mitte festzustellen. Der überwiegende Teil der Bilder weist LIVE2 DMOS zwischen 25 und 75 auf. Der maximale LIVE2 DMOS beträgt 85. Häufungsgipfel finden sich bei DMOS-Werten um 25 und 50.

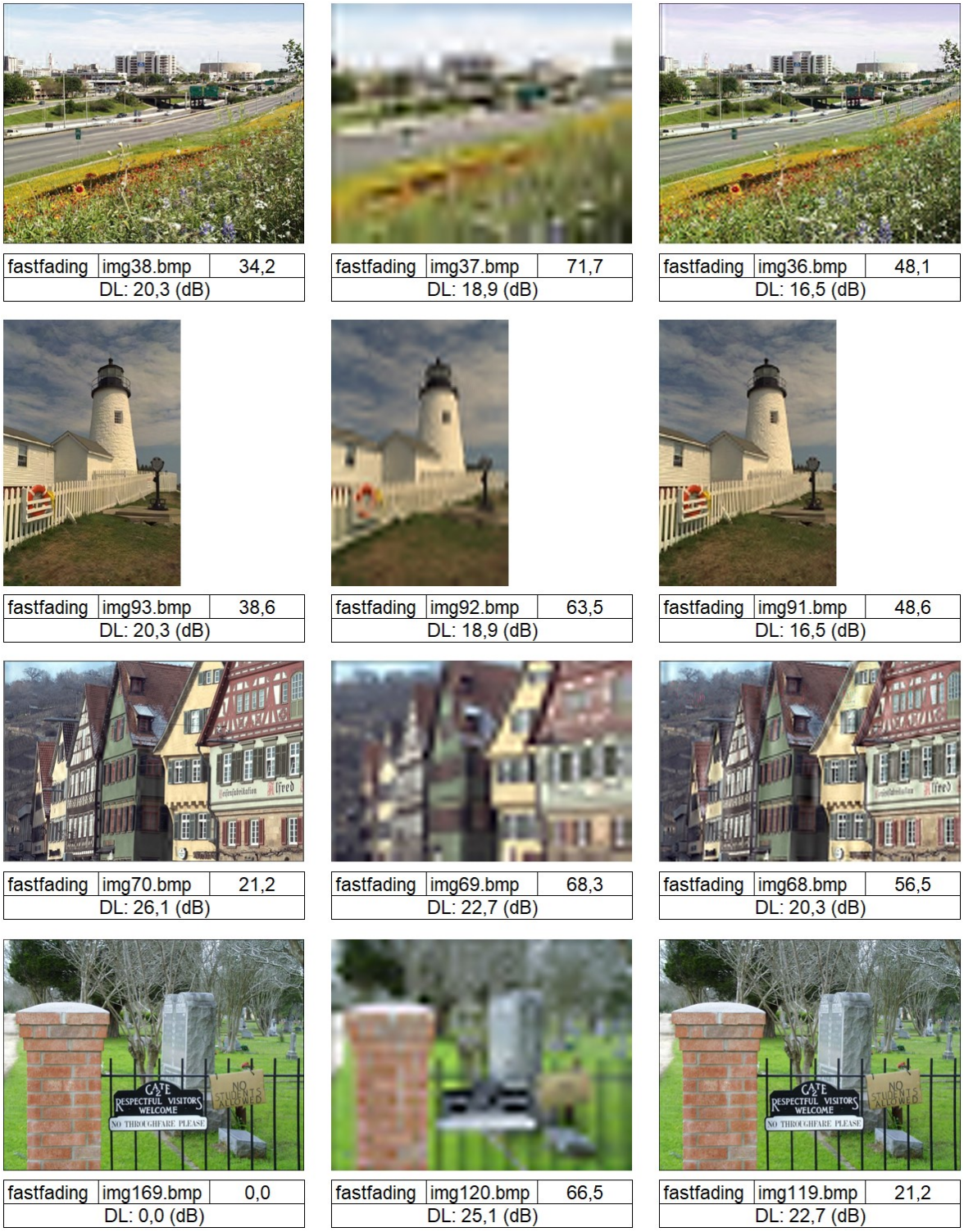


Abb. 14: Exemplarische Ausschnitte aus Bildserien zur Verzerrungsart *FastFading*, bei denen sich die subjektive Bildqualität mit zunehmendem Verzerrungsgrad zunächst verschlechtert und bei weiter zunehmender Verzerrung dann intermittierend wieder bessert. Verzerrungsgrad (DL) in Dezibel (dB).

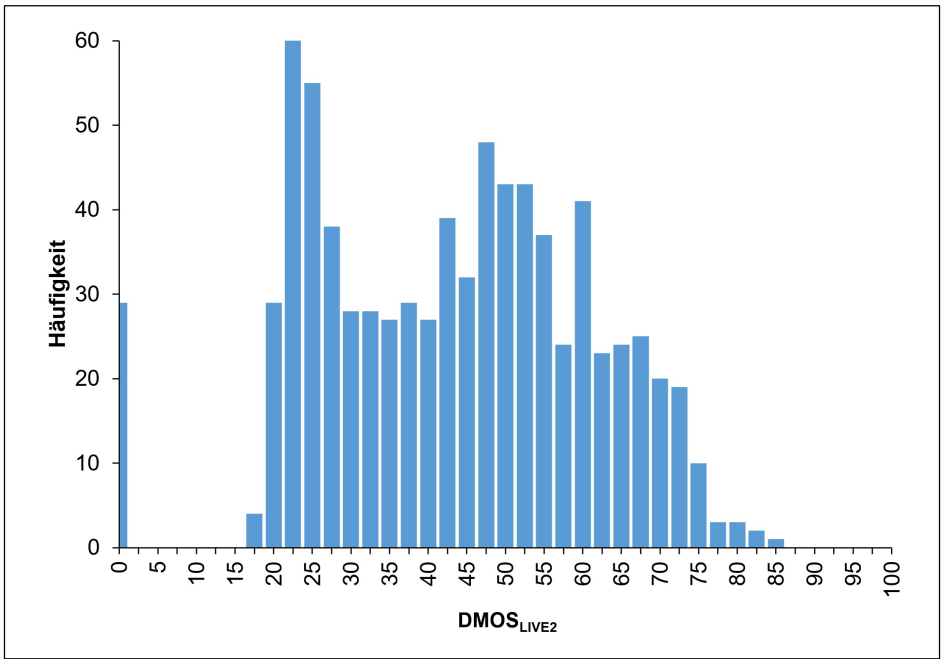


Abb. 15: Häufigkeitsverteilung der LIVE2 DMOS-Werte zu allen Bildern der LIVE2 Datenbank ohne Duplikate. Bining: Breite 2,5, zentriert ab null. Die 29 Bewertungen mit DMOS = 0 entsprechen den 29 Referenzbildern in den fünf Verzerrungsarten. Der minimale DMOS beträgt 17,9, der maximale DMOS 84,5. Die Häufigkeitsverteilung zeigt eine starke Tendenz zur Mitte. Auch unkenntliche Bilder erhalten keinen DMOS > 85.

2.3 Zwischenzusammenfassung

Im vorangegangenen Kapitel wurde die Untersuchung der LIVE2 Datenbank vorgestellt. Hierzu wurde zunächst im Abschnitt Methode das zur Untersuchung verwendete Spreadsheet erläutert. Dieses integriert sämtliche LIVE2 Daten und ermöglicht eine vergleichende Auswertung der Bilder an Hand dieser Parameter. Anschließend wurden die Untersuchungsergebnisse vorgestellt. Explizit wurde die Datenbank auf die folgenden Faktoren hin untersucht: DMOS ≥ 70, DMOS Intervalle, DMOS Monotonie, Fast Fading Qualitätsdiskontinuität, DMOS Verteilung in der LIVE2 Datenbank.

3 Vergleich experimenteller DMOS und MLDS als Metriken der subjektiven LIVE2 Bildqualität

Auf Grund der gewonnen Erkenntnisse der LIVE2 DMOS Untersuchung wurden anschließend zwei Experimente durchgeführt. Ziel war die vergleichende Differenzierung geeigneter Metriken zur subjektiven Bildqualität anhand einer Auswahl von LIVE2 Bildern. Hierfür wurden die Bewertungen exemplarischer Versuchspersonen in zwei Experimenten sowohl zum MOS, als auch nach MLDS erhoben. Nachfolgend wird zunächst die Methode der beiden Experimente dargestellt, daran anschließend die experimentellen Ergebnisse.

3.1 Methode

Die Abschnitte Stimuli, Versuchspersonen, Hardware & Software sowie Instruktionen erläutern zunächst die für beide Experimente identischen Faktoren. Daran anschließend werden Design, Durchführung und Auswertung jeweils für das MOS und das MLDS Experiment detailliert.

3.1.1 Stimuli

Für die beiden eigenen Wahrnehmungsexperimente dieser Arbeit wurden nach Analyse der LIVE2 Datenbank (Kapitel 2) die vier Motive *Bikes*, *ChurchAndCapitol*, *Ocean* und *WomanHat* mit den Verzerrungsarten *JPEG*, *White Noise*, *Gaussian Blur* und *Fast Fading* ausgewählt. Für jede dieser Bildserien wurden jeweils sechs Verzerrungsgrade zur experimentellen Untersuchung festgelegt. Abbildung 16 zeigt hierzu die nicht verzerrten Referenzbilder der vier Motive. Im Original hatten die Bilder folgende Pixelformate (Breite x Höhe): *Bikes*: 768 x 512, *ChurchAndCapitol*: 634 x 505, *Ocean*: 768 x 512, *WomanHat*: 480 x 720. Um bei der Präsentation auf dem Monitor eine vergleichbare Höhe für alle Bilder zu erreichen, wurde der Stimulus *WomanHat* auf die Größe 341 x 512 transformiert.

Abbildung 17 gibt eine vergleichende Übersicht über alle untersuchten Verzerrungsarten und Verzerrungsgrade am Beispiel des Motivs *WomanHat*. Eine entsprechende Übersicht zu allen vier Motiven der Experimente findet sich im Anhang (Abb. A1 - A4).

Wie Lévêque et al. (2020) gezeigt haben, beeinflussen auch die dargestellten Motive eines Bildes die Beurteilung der wahrgenommenen Bildqualität. Dem entsprechend

Abb. 16: Die vier Referenzbilder der eigenen Wahrnehmungsexperimente: *Bikes*, *ChurchAndCapitol*, *Ocean* und *WomanHat*

wurden die oben aufgeführten Motive aus den unterschiedlichen Inhaltskategorien Social (*Bikes*), Outdoor man-made (*ChurchAndCapitol*), Outdoor natural (*Ocean*) und Portrait (*WomanHat*) ausgewählt.

3.1.2 Versuchspersonen

Die Versuche wurden vom Erstbetreuer (GA) der Arbeit (38 Jahre, männlich, Postdoktorand, freiwillige Teilnahme, keine Vergütung) sowie der Autorin selbst (AS) (26 Jahre, weiblich, Studentin der Informatik, freiwillige Teilnahme, keine Vergütung) durchgeführt. Versuchsperson GA hat dabei fünf Wiederholungen der beiden Experimente durchgeführt, AS insgesamt zehn Wiederholungen.

3.1.3 Hardware und Software

Die eigenen Wahrnehmungsexperimente wurden alle im psychovisuellen Labor unter Standardbedingungen und für alle Versuchspersonen identisch durchgeführt. Der physische Aufbau des Experiments entsprach den ITU Richtlinien (ITU, 2019):

Raumumgebung: vollständig abgedunkelter Raum, Bildschirm vor einer weißen Wand, Reduktion von Streulicht durch Platzierung der Hardware auf einer schwarzen Oberfläche.

Betrachtungsabstand: 1 m

Bildschirm: Viewpixx 30 (Vpixx Technologies), 22.5 inch (diagonal)

Bildschirmauflösung: 1920(H) x 1080(V) Pixel

Um eine konstante Ausrichtung der Augen zum Bildschirm gewährleisten zu können, wurde für alle Durchläufe eine Kinnkopfstütze verwendet.

Abb. 17: Motiv *WomanHat*, in jeder Zeile eine Verzerrungsart, in jeder Spalte ein Verzerrungsgrad sortiert nach Verzerrungsrang. Unter jedem Bild ist der Verzerrungsgrad in den Einheiten der jeweiligen Verzerrungsart angegeben: Bits pro Pixel (bpp), Sigma (σ), Dezibel (dB). Die Titelzeile gibt den Verzerrungsrang von 1 bis 6 wieder. Die linke Spalte mit Verzerrungsgrad null bzw. Rang eins enthält das für alle Verzerrungsarten identische unverzerrte Referenzbild. Das am stärksten verzerrte Bild befindet sich in Spalte sechs. Der numerische Wert des Verzerrungsgrades ist dabei für *JPEG* und *FastFading* von links nach rechts absteigend, für *WhiteNoise* und *GaussianBlur* von links nach rechts aufsteigend.

3.1.4 Instruktionen

Beide Versuchspersonen haben sich mit der implementierten Nutzeroberfläche, der visuellen Präsentation der Stimuli und der Abgabe der Bewertungen für die beiden Experimente anhand von Demonstrationsbildern vertraut gemacht. Die Demonstrationsbilder waren nicht Teil der bewerteten Bilder des Experiments. Die einzelnen Wiederholungen der Experimente haben die beiden Versuchspersonen eigenständig und unabhängig voneinander durchgeführt. Die Wiederholungen erfolgten zu unterschiedlichen Tageszeiten und Arbeitstagen.

3.1.5 MOS Experiment

Das erste Wahrnehmungsexperiment dieser Arbeit zielte darauf ab, die MOS-Werte der LIVE2 Datenbank für die ausgewählten Bilder nachzuvollziehen. Die Bewertung der Bilder wurde dazu als Einzelstimulus-Präsentation (Abb. 18) mit PsychoPy implementiert. PsychoPy ist eine Open-Source Python Software, die für die experimentelle neurowissenschaftliche und psychologische Forschung entwickelt wurde (Peirce, 2007).

Design und Durchführung

Die Einzelstimuli wurden auf weißem Hintergrund präsentiert. Abweichend zur LIVE2 Datenerhebung wurde eine visuelle Analogskala zwischen null (schlechteste Bildqualität) und 100 (beste Bildqualität) ohne semantische Klassifizierung der Bewertung mittels eines Schiebereglers auf dem Bildschirm realisiert (Abb. 18). Die subjektive Bewertung eines Bildes wird durch die Positionierung des Bildschirmzeigers auf der Linie des Schiebereglers und Klicken der Maustaste abgegeben.

Ein Durchlauf des MOS-Experiments beinhaltete 96 Bilder mit vier Motiven, vier Verzerrungsarten und sechs Verzerrungsgraden. Jedes Bild wurde genau einmal präsentiert. Die Reihenfolge der Bilder innerhalb eines Durchlaufs wurde randomisiert.

Die Implementierung erlaubt beliebige Pausen, Unterbrechungen und Fortsetzungen eines Durchlaufs. Für die Bewertung eines Bildes sind ungefähr drei Sekunden vorgesehen. Die Durchführung von insgesamt zehn Wiederholungen mit insgesamt 960 Bildern benötigt damit etwa 45 Minuten.

Die Präsentationssoftware speichert die Identifikation des gezeigten Bildes und die abgegebene Bewertung in csv-Dateien. Die Daten der verschiedenen Wiederholungen wurden abschließend mit einer Python Implementierung zusammengeführt und ausgewertet.

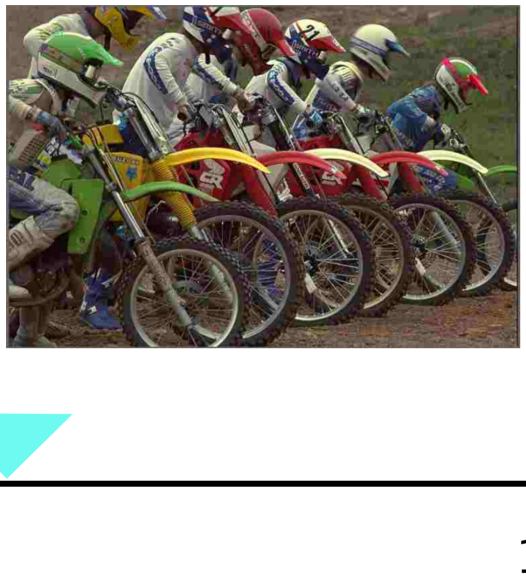


Abb. 18: Nutzeroberfläche des MOS-Experiments. Die Versuchsperson wurde gebeten, durch Klicken der Maustaste auf der Linie eine Bewertung der wahrgenommen Bildqualität abzugeben. 0: schlechteste Bildqualität, 100: beste Bildqualität

Einschätzung der MOS-Skala

Jede einzelne Bewertung v_i wurde gemäß

$$OS_i = 100 - v_i \tag{1}$$

auf einen Opinion Score $OS_i \in [0, 100]$ abgebildet, wobei null die beste und 100 die schlechteste Bildqualität repräsentiert. Der MOS und DMOS wurden gemäß Gleichungen (2) und (3) ermittelt:

$$MOS_{mkj} = \frac{1}{n} \sum_{i=1}^{n} OS_{mkji} \tag{2}$$

$$DMOS_{mkj} = MOS_{mkj} - MOS_{mk1} \tag{3}$$

mit $m =$ Index des Motivs, $k =$ Index der Verzerrungsart, $j =$ Index des Verzerrungsgrades und $i =$ Index der Präsentation. Es gilt: $m, k \in \{1, ..., 4\}$, $j \in \{1, ..., 6\}$, $i \in \{1, ..., n\}$ und $m, k, j, i \in \mathbb{N}$. Der Verzerrungsindex $j = 1$ kennzeichnet das nicht verzerrte Referenzbild. n entspricht der Gesamtzahl aller Wiederholungen des Stimulus

mkj, abhängig von der jeweiligen Auswertung über einzelne oder alle Versuchspersonen.

3.1.6 MLDS Experiment

Im zweiten Experiment dieser Arbeit wurde die subjektive Wahrnehmung derselben Bilderauswahl mit der Methode des MLDS untersucht. Hierzu wurde eine Multistimulus-Präsentation von Bildtriaden (Abb. 19) mittels der Python Bibliothek `pyglet` implementiert. `Pyglet` wird überwiegend im Multimediabereich zur Entwicklung von Spielen eingesetzt.

Design und Durchführung

Eine einzelne Triade besteht aus drei Bildern unterschiedlicher Verzerrungsgrade zum selben Motiv und zur selben Verzerrungsart. Die Bilder wurden auf weißem Hintergrund präsentiert. Die Versuchsperson bewertet, ob der subjektive Qualitätsunterschied zwischen dem mittleren und linken Bild größer ist als der zwischen mittlerem und rechten Bild. Die Bewertung wird mit den Pfeil-Tasten der Tastatur abgegeben. Empfindet die Versuchsperson den Unterschied zwischen linkem und mittlerem Bild größer als zwischen mittlerem und rechtem Bild, dann drückt sie die linke Pfeiltaste, andernfalls die rechte.

Die Verzerrungsgrade der drei Bilder eines Multistimulus sind nicht überlappend gewählt, also monoton aufsteigend ($j_{Links} < j_{Mitte} < j_{Rechts}$) oder absteigend ($j_{Links} > j_{Mitte} > j_{Rechts}$) sortiert, mit j = Index des Verzerrungsgrades. Zu p Verzerrungsgraden pro Motiv und Verzerrungsart ergeben sich damit n Kombinationen an nicht überlappenden Triaden gemäß Gleichung (4) (Knoblauch & Maloney, 2012).

$$n = \binom{p}{3} = \frac{p!}{(p-3)! \cdot 3!} \tag{4}$$

Bei sechs Verzerrungsgraden sind das 20 mögliche Triaden. Ein Durchlauf des MLDS-Experiments mit vier Motiven und vier Verzerrungsarten umfasst daher die Präsentation von insgesamt 320 Triaden. Jede Triade wird jeweils einmal präsentiert. Die Reihenfolge der Triaden innerhalb eines Durchlaufs ist randomisiert. In gleicher Weise wird jede einzelne Triade bei jeder Präsentation randomisiert aufsteigend oder absteigend präsentiert. Dadurch wird verhindert, dass sich die Ergebnisse einem systematischen Verlauf anpassen und Stimulus-Antwort-Strukturen entstehen.

Die Implementierung der Stimuluspräsentation erlaubt beliebige Pausen, Unterbrechungen und Fortsetzungen eines Durchlaufs. Für die Bewertung eines Bildes werden ungefähr drei Sekunden benötigt. Die Durchführung von zehn Wiederholungen ergibt insgesamt 3200 präsentierte Triaden, für die eine Bearbeitung von etwa zweieinhalb Stunden vorgesehen war.

Die implementierte Präsentationssoftware speichert zu jeder Präsentation eines Multistimulus die Identifikation von Motiv, Verzerrungsart, die drei Verzerrungsgrade der Triade sowie die Bewertung der Triade mit 0 (Pfeil links) und 1 (Pfeil rechts) in csv-Dateien. Die Daten der verschiedenen Wiederholungen wurden abschließend algorithmisch zusammengeführt und ausgewertet.

Einschätzung der MLDS-Skala

Zu jedem Motiv und jeder Verzerrungsart wurde die psychophysische Wahrnehmungsfunktion $\psi(x)$ mit dem generalisierten linearen Modell (GLM) aus Gleichung (5) bestimmt (Knoblauch & Maloney, 2012).

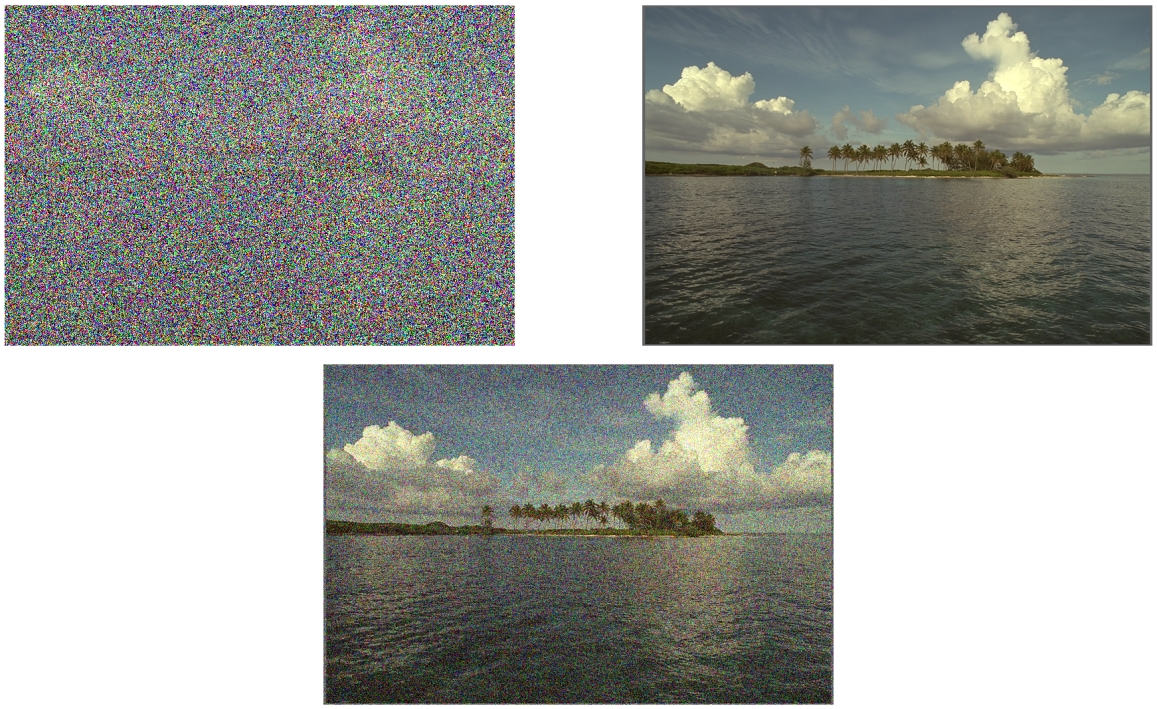


Abb. 19: Nutzeroberfläche des MLDS-Experiments. Die Versuchsperson wurde gebeten, eine Bewertung abzugeben, zwischen welchen Bildern sie einen größeren Qualitätsunterschied wahrnimmt. Wird der Qualitätsunterschied zwischen dem linken und mittleren Bild als größer wahrgenommen als zwischen dem mittleren und rechten Bild, wird auf der Tastatur die Pfeiltaste nach links gedrückt. Wird der Qualitätsunterschied zwischen dem mittleren und rechten Bild als größer wahrgenommen, wird die Pfeiltaste nach rechts betätigt.

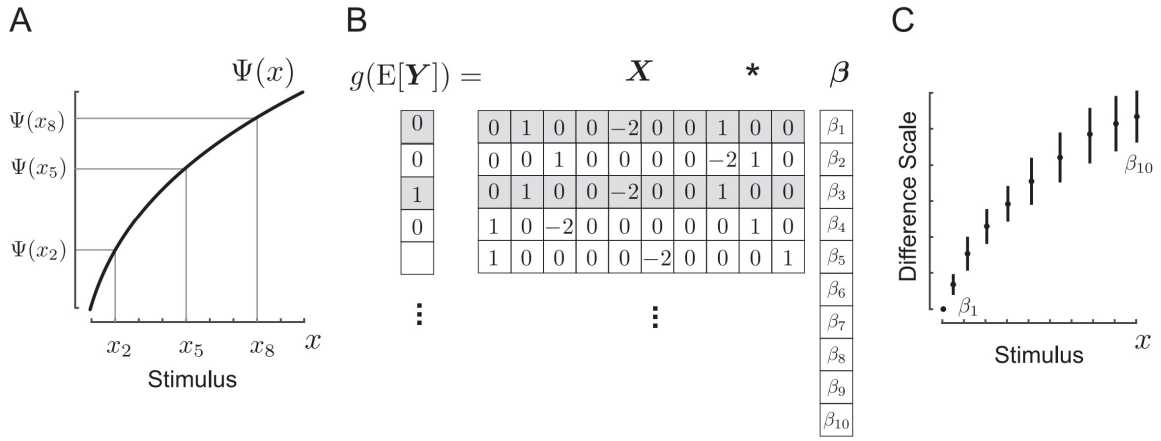


Abb. 20: Schätzung der psychophysikalischen Skalen $\psi(x)$ mittels MLDS für Stimulustriaden. (A) Hypothetische Skala der wahrgenommenen Bildqualität $\psi(x_j)$ zum Stimuluswert x_j. Für die Beispieltriade (x_2, x_5, x_8) wurden Versuchspersonen gebeten das Paar zu nennen, (x_2, x_5) oder (x_5, x_8), das einen größeren subjektiven Qualitätsunterschied aufweist. Das Entscheidungsmodell für die Beispieltriade ist dabei $\Delta = [\psi(x_8) - \psi(x_5)] - [\psi(x_5) - \psi(x_2)]$ bzw. $\Delta = \psi(x_8) - 2\psi(x_5) + \psi(x_2)$. (B) Jede Zeile der Designmatrix X repräsentiert die Elemente einer Triade. Die schraffierten Zeilen sind zwei Wiederholungen der gleichen Triade, d. h. der Beispieltriade in (A). Die Designmatrix enthält alle in einem Versuch gezeigten Triaden. (C) Empirische Skala, die sich aus der Lösung des generalisierten linearen Modells (GLM) und der Ermittlung des Parameters β ergibt. Die Fehlerbalken sind Schätzungen des Messfehlers, die mittels Bootstrap ermittelt wurden (Abb. aus Wiebel, Aguilar & Maertens, 2017).

$$E[Y] = g^{-1}(X\beta) \tag{5}$$

Dabei repräsentiert $Y = \{y_1, ..., y_n \in \{0, 1\}, n \in \mathbb{N}\}$ den Bewertungsvektor aller n präsentierten Triaden mit Dimension $n \times 1$, X die Design-Matrix aller Triaden mit Dimension $n \times p$ mit p der Anzahl an verwendeten Einzelstimuli, aus denen die Triaden gebildet werden. In der vorliegenden Arbeit gilt $p = 6$ für sechs Verzerrungsgrade je Motiv. $E[Y]$ ist der mittlere Erwartungswert des Bewertungsvektors und $g()$ die Kopplungsfunktion zwischen linearem Prädiktor $X\beta$ und Erwartungswert $E[Y]$ (Abb. 20) (Wiebel et al., 2017). Die Schätzwerte der unbekannten Parameter β_j werden durch lineare Regression mit Maximum-Likelihood bestimmt. Die so ermittelten $\beta_j(x_j)$ approximieren die psychophysische Differenzskala $\psi(x)$, wobei j wieder den Verzerrungsgrad der Stimuli indiziert.

Das GLM wurde mit der Statistik Software *R* unter Verwendung der R-Bibliothek `mlds` gelöst. Dazu wurden die Rohdaten des Experiments zu jedem Motiv und jeder Verzerrungsart mit Python in einzelne csv-Dateien im Eingabeformat der `mlds`-Funktion

konvertiert. Dabei wird jede Bewertung einer präsentierten Triade durch einen Datensatz (y, x_1, x_2, x_3) wiedergeben, wo $y \in \{0, 1\}$ die Antwort der Versuchsperson und $x_{j,\ j=1,2,3} \in \{1, \ldots p\}$ die drei Indizes der Triade aus den sechs möglichen Verzerrungsgraden repräsentieren. Die Funktion `mlds` gibt nach linearer Regression des GLM im Vektor $\psi(x)$ die geschätzte psychophysische Differenzskala zurück.

Zur vergleichenden grafischen Darstellung von MOS und MLDS Bewertungen wurden die $\psi(x)$-Werte über alle Motive und Verzerrungsarten gemeinsam linear auf den Wertebereich [0, 100] abgebildet. Ebenfalls zur grafischen Darstellung wurden die Verzerrungsgrade null der nicht verzerrten Bilder für die Verzerrungsart *JPEG* auf acht bpp und für die Verzerrungsart *Fast Fading* auf 30 dB gesetzt. Damit werden die nicht verzerrten Bilder in der grafischen Darstellung auch für *JPEG* und *Fast Fading* in der Abfolge fallender Verzerrungsgrade korrekt entsprechend dem Verzerrungsrang eins neben dem Bild mit dem geringsten Verzerrungsgrad platziert. Der Wert von acht bpp für *JPEG* Bilder mit Verzerrungsgrad null orientiert sich am RGB-Farbraum mit maximal acht Bit pro Pixel. Der Wert von 30 für *Fast Fading* Bilder mit Verzerrungsgrad null ist frei gewählt. Grundsätzlich gilt für das Signal-Rausch-Verhältnis gemessen in Dezibel (dB): SNR $\rightarrow \infty$. Die Grafiken wurden mit `jupyter notebooks` und der Python Bibliothek `seaborn` untersucht. Die `jupyter notebooks` sind unter `https://github.com/AnnalenaSchillen/LIVE2_MOS_BA` abrufbar.

3.2 Ergebnisse

Ausgehend von der Analyse der LIVE2 Datenbank wurden ausgewählte Bildserien in eigenen Wahrnehmungsexperimenten weiter untersucht. Die Bilderauswahl orientierte sich dabei an den Auffälligkeiten der LIVE2 Untersuchung (Kapitel 2). Zusätzlich wurde darauf geachtet, Motive aus diversen Inhaltskategorien zu verwenden (Kapitel 3.1.1). Diese Überlegungen bestimmten die Auswahl der vier Motive *Bikes*, *ChurchAndCapitol*, *Ocean* und *WomanHat* (Abb. 16). Als Verzerrungsarten wurden *JPEG*, *White Noise*, *Gaussian Blur* und *Fast Fading* in die Untersuchung eingeschlossen. Zu jedem dieser Motive und Verzerrungsarten wurden jeweils sechs Verzerrungsgrade (Abb. 17) für die Experimente ausgewählt. Zwei Versuchspersonen bewerteten jede dieser Bildserien sowohl in einem MOS Einzelstimulus-Experiment als auch in einem MLDS Multistimulus-Design (Kapitel 3.1).

Für die Fragestellung dieser Arbeit ist die subjektive Qualität der Bilder einzeln und im Vergleich die maßgebliche Grundlage. Die Bewertung der experimentellen Ergebnisse erfordert deshalb, die gemessenen MOS und MLDS-Verläufe den Bildserien direkt gegenüberzustellen (vgl. Abb. 22). Im Folgenden wird hier jeweils eine solche Auswertung

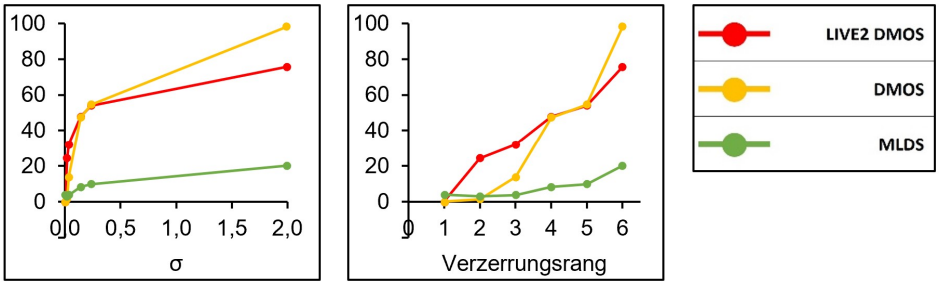


Abb. 21: Exemplarische Darstellung der Auswertung anhand von Motiv *ChurchAndCapitol*, Verzerrungsart *WhiteNoise*, (Links) Abszisse: Verzerrungsgrad 0.000, 0.02, 0.035, 0.142, 0.234, 1.996 σ, rot: DMOS LIVE2, gelb: DMOS beider Versuchspersonen, grün: MLDS beider Versuchspersonen der eigenen Experimente. Die enge Verteilung der einzelnen Bilder bei geringen Verzerrungsgraden überlagert viele Messpunkte. (Rechts) Abszisse: Verzerrungsrang. Rang eins: nicht verzerrtes Referenzbild. Rang sechs: am stärksten verzerrtes Bild. Diese Darstellung gibt die Messwerte ohne Überlagerung gut unterscheidbar wieder.

zu jedem der vier Motive und jeder der vier Verzerrungsarten exemplarisch untersucht. Eine Zusammenstellung aller 16 Auswertungen findet sich im Anhang (Abb. A5 - A20).

Bei *JPEG* und *Fast Fading* entspricht der zunehmende Verzerrungsgrad der Bilder einer Abnahme der Bit-Rate (bpp) bzw. des Signal-Rausch-Verhältnisses (dB), bei *White Noise* und *Gaussian Blur* einer Zunahme der Standardabweichung σ der addierten Verzerrung. Zur besseren Vergleichbarkeit zwischen den verschiedenen Verzerrungsarten sind die Messergebnisse so dargestellt, dass der Wert des nicht verzerrten, besten Bildes immer durch den ersten Datenpunkt ganz links in den Grafiken wiedergegeben wird. Auf Grund der häufig engen Verteilung der ersten Verzerrungsgrade bei vielen Bildern der LIVE2 Datenbank werden die Grafiken zusätzlich nach Verzerrungsrang dargestellt. Die Indizes von eins bis sechs geben dabei die zunehmende Verzerrung wieder. Rang eins ist immer das nicht verzerrte Referenzbild, Rang sechs das Bild mit der stärksten Verzerrung.

Abbildung 21 demonstriert exemplarisch für das Motiv *ChurchAndCapitol* mit Verzerrungsart *WhiteNoise*, wie die enge Verteilung von Verzerrungsgraden der LIVE2 Datenbank die Messwerte schlecht differenzierbar überlagert. Die Darstellung nach Verzerrungsrang vermeidet dieses Problem. Im Anhang finden sich sämtliche Ergebnisse zu allen Motiven und Verzerrungsarten sowohl in einer Darstellung nach Verzerrungsgrad (Abb. A5 - A12) als auch nach Verzerrungsrang (Abb. A13 - A20). Zur besseren Lesbarkeit der Diagramme wird die gemeinsame Legende einmalig in Abbildung 22 dargestellt. Für alle folgenden Beispiele ist diese identisch.

In jeder dieser Darstellungen (Abb. 22 - 25) befindet sich das nicht verzerrte Referenzbild in der linken oberen Kachel. Daneben folgen von links nach rechts die weiteren fünf

Bilder der Serie mit aufsteigendem Verzerrungsrang, so dass in der rechten mittleren Kachel das Bild mit der stärksten Verzerrung wiedergegeben ist. Die unteren Kacheln zeigen links die Ergebnisse der beiden Versuchspersonen für das MOS-Experiment (a) und in der Mitte die Ergebnisse beider Versuchspersonen für das MLDS-Experiment (b). Diese beiden Darstellungen ermöglichen eine Kontrolle der Messergebnisse zwischen den beiden Versuchspersonen. Die Kachel unten rechts (c) fasst alle Ergebnisse zusammen: DMOS der LIVE2 Datenbank (rot), DMOS beider Versuchspersonen (gelb), MLDS beider Versuchspersonen (grün).

Zu den Ergebnissen im Detail: Die Auswertung zum Motiv *ChurchAndCapitol* mit der Verzerrungsart *JPEG* (Abb. 22) zeigt eine sehr gute interne Kontrolle der beiden Versuchspersonen. Gut bedeutet dabei, dass die Kurven einen sehr ähnlichen Kurvenverlauf ohne besondere Abweichungen aufweisen. Sowohl beim Einzelstimulus-Experiment zum DMOS (a), als auch beim Multistimulus-Design zum MLDS (b) stimmen die Kurven der beiden Versuchspersonen sehr gut überein.

Beim subjektiven Qualitätsvergleich der gesamten Bildserie (Abb. 22 obere und mittlere Zeile) zeigen die ersten vier Bilder sehr geringe Unterschiede. Details und Farben der ersten drei Verzerrungsgrade verlieren wenig gegenüber dem nicht verzerrten Referenzbild. Erst die letzten beiden Bilder zeigen deutliche Qualitätseinbußen.

Der Verlauf des DMOS der LIVE2 Datenbank (c: rot) gibt diese subjektive Metrik nicht wieder. Bereits die Bilder zwei und drei werden mit DMOS-Werten von 20 und 40 deutlich schlechter bewertet als das Referenzbild. Für die Bilder vier und fünf nimmt der DMOS weiter zu. Das letzte Bild wird nach LIVE2 DMOS dann mit 62,5 wieder deutlich besser bewertet als Bild fünf. Nach dieser Bewertung verbessert sich die Bildqualität von Bild fünf nach sechs vergleichbar zur Verbesserung von Bild drei nach zwei und zwei nach eins. Der LIVE2 DMOS steht hier in Widerspruch zur subjektiven Qualitätsmetrik bei gemeinsamer Betrachtung aller sechs Bilder der Serie. Demgegenüber bewertet der DMOS der eigenen Messung (c: gelb) die ersten beiden Bilder als gleichwertig und Bild drei mit DMOS 13 nur wenig schlechter. Für die folgenden Bilder nimmt der DMOS kontinuierlich bis auf 90 zu. Beachtenswert ist, dass der DMOS-Abstand zwischen Bild vier und fünf bei 50 liegt, während von Bild fünf nach sechs ein Abstand von neun Punkten gemessen wurde. Diese Abstände der eigenen DMOS-Messung entsprechen ebenfalls nicht den subjektiven Qualitätsdifferenzen bei Betrachtung der gesamten Bildserie. In Kontrast dazu bewertet die MLDS-Messung die ersten vier Bilder als praktisch gleichwertig. Der MLDS-Abstand von Bild vier nach fünf beträgt dann etwa 25, der zwischen fünf und sechs vergleichbare 28 Punkte. Der gemessene MLDS-Verlauf zeigt damit eine sehr gute Übereinstimmung mit der oben skizzierten Metrik der subjektiven Qualitätsabstände aller sechs Bilder dieser Serie.

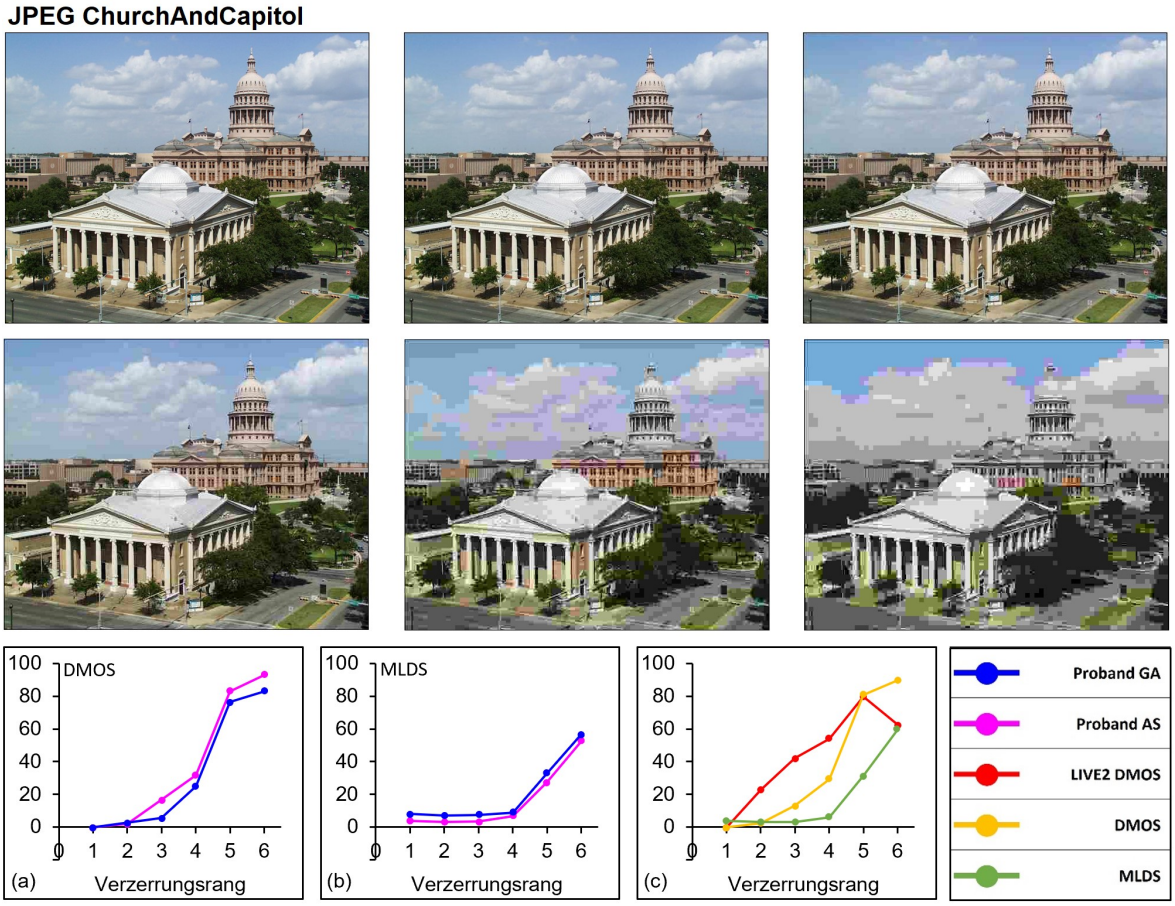


Abb. 22: Motiv *ChurchAndCapitol*, Verzerrungsart *JPEG*, Verzerrung: Rang 1 bis 6, Bilder: img22, img132, img2, img21, img4, img195. (a) DMOS, (b) MLDS, (a, b): blau: Versuchsperson GA, pink: Versuchsperson AS, (c) rot: DMOS LIVE2, gelb: DMOS beider Versuchspersonen, grün: MLDS beider Versuchspersonen. Details: siehe Text

Beim nächsten Beispiel mit dem Motiv *WomanHat* und der Verzerrungsart *WhiteNoise* (Abb. 23) zeigen die ersten drei Bilder keinen erkennbaren Qualitätsunterschied. Bild vier lässt bei Darstellung in Originalgröße einen Qualitätsverlust durch ein leichtes Rauschen erkennen. Der Qualitätsverlust durch Rauschen nimmt bei Bild fünf und sechs weiter zu. Der Qualitätsverlust von Bild drei zu vier wirkt stärker als von Bild vier zu fünf. Bei Bild sechs ist die Bildqualität dann erheblich reduziert, das Motiv bleibt aber auch bei Betrachtung als Einzelbild noch erkennbar.

Die Kontrolle auf konsistente Messungen (a, b) zeigt wieder eine gute Übereinstimmung beider Versuchspersonen. Die Auswertung (c) zeigt beim LIVE2 DMOS (rot) eine Bewertung von Bild zwei und drei mit DMOS-Werten von 25 und 37, die deutlich schlechter ausfällt als das Referenzbild. Der Abstand von Bild zwei zu drei wird mit acht Punkten als genauso groß bewertet, wie zwischen Bild drei und vier, während der Unterschied von Bild vier zu fünf mit vier Punkten sehr gering ausfällt. Von Bild fünf zu sechs ist die Differenz mit 19 Punkten ähnlich bewertet wie der Abstand von

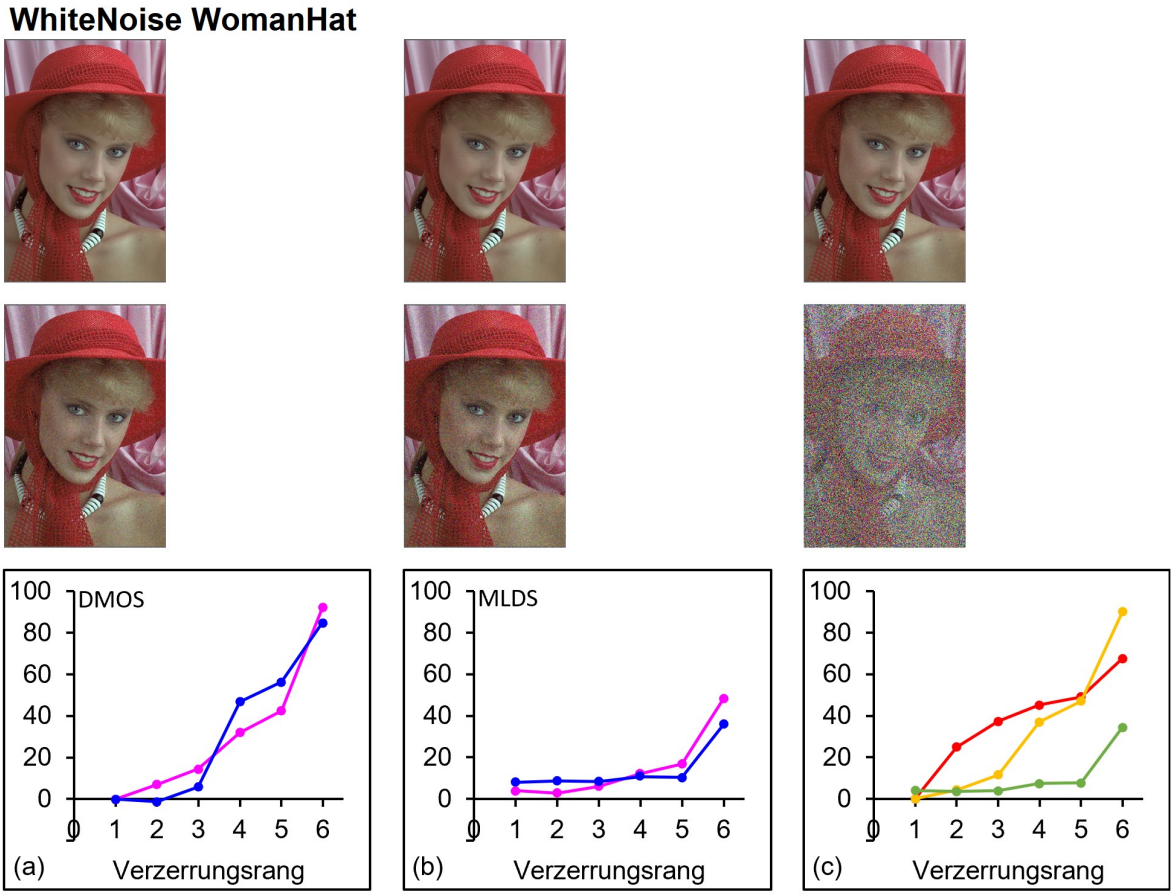


Abb. 23: Motiv *WomanHat*, Verzerrungsart *White Noise*, Verzerrung: Rang 1 bis 6, Bilder: img162, img11, img3, img28, img119, img60. (a) DMOS, (b) MLDS, (a, b): blau: Versuchsperson GA, pink: Versuchsperson AS, (c) rot: DMOS LIVE2, gelb: DMOS beider Versuchspersonen, grün: MLDS beider Versuchspersonen. Legende: siehe Abb. 22. Details: siehe Text.

Bild eins zu zwei. Der LIVE2 DMOS entspricht damit auch hier nicht der subjektiven Metrik bei Betrachtung der gesamten Bildserie. Bei der eigenen DMOS-Messung (c: gelb) wird Bild zwei mit vier Punkten als gleichwertig zum Referenzbild bewertet und auch Bild drei erhält mit DMOS 14 eine ähnlich gute Bewertung. Der Abstand von Bild drei zu vier wird mit 25 Punkten als 2,5-fach so groß wie von Bild vier zu fünf bewertet. Die DMOS-Differenz fünf nach sechs ist mit 42 Punkten dann besonders groß. Damit entspricht die eigene DMOS-Messung in einzelnen Aspekten besser der subjektiven Vergleichsbewertung als der LIVE2 DMOS, hat aber auch bei den ersten Bildern und Abstandsbewertungen ihre Abweichungen. Die Daten des MLDS (c: grün) zeigen für die ersten drei Bilder praktisch identische Werte um vier Punkte. Danach kommt eine leichte Qualitätseinbuße von Bild drei nach vier. Bild vier und fünf werden wieder ohne wesentlichen Qualitätsunterschied bewertet. Bei Bild sechs bildet der MLDS dann den deutlichen Qualitätsverlust ab. Insgesamt entspricht der MLDS damit auch bei dieser Bildserie sehr gut dem subjektiven Qualitätseindruck bei der gemeinsamen Betrachtung aller sechs Bilder.

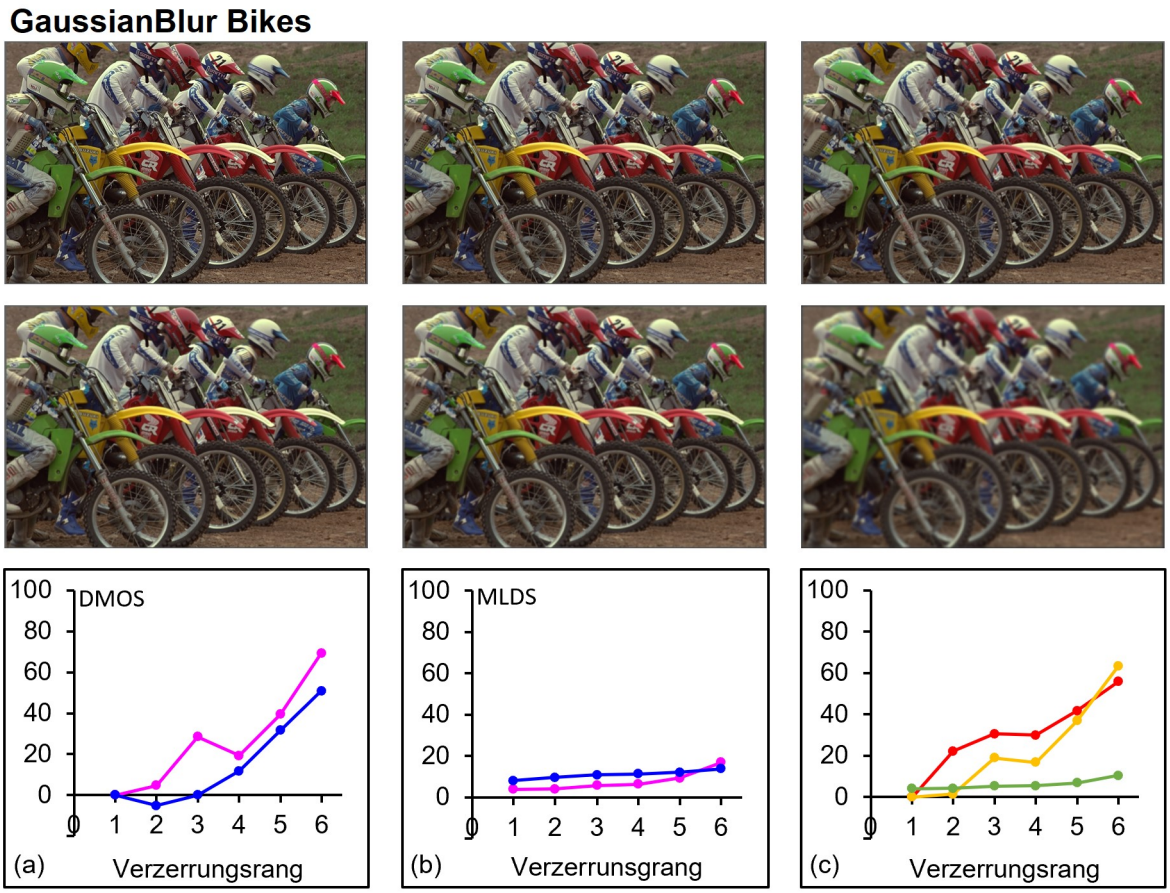


Abb. 24: Motiv *Bikes*, Verzerrungsart *Gaussian Blur*, Verzerrung: Rang 1 bis 6, Bilder: img152, img60, img112, img129, img26, img40. (a) DMOS, (b) MLDS, (a, b): blau: Versuchsperson GA, pink: Versuchsperson AS, (c) rot: DMOS LIVE2, gelb: DMOS beider Versuchspersonen, grün: MLDS beider Versuchspersonen. Legende: siehe Abb. 22. Details: siehe Text.

In der Reihe *GaussianBlur Bikes* (Abb. 24) weisen alle Bilder mit Ausnahme des am stärksten verzerrten Bildes minimale Unterschiede auf. Auch die Unterschiede von Details und Farben der Bilder sind untereinander sehr gering. Ohne eine Betrachtung der Bilder in Originalgröße sind die subjektiven Unterschiede fast unkenntlich. Bild sechs mit dem stärkstem Verzerrungsgrad ist weiterhin gut erkennbar und von brauchbarer subjektiver Bildqualität.

Die DMOS-Bewertungen der Versuchspersonen (a) weichen im Gegensatz zu den beiden vorhergehenden Beispielen erstmalig von einander ab. Die DMOS-Werte für Versuchsperson GA (a: blau) nehmen von Bild zwei bis sechs kontinuierlich zu. Die Werte für Versuchsperson AS (a: pink) hingegen nicht. Für AS hat Bild drei einen DMOS Wert von 28 während Bild vier mit einem DMOS Wert von 19 als besser bewertet wurde. Der größere Qualitätsunterschied von Bild fünf zu sechs ist für beide Versuchspersonen konsistent und durch einen deutlichen Anstieg erkennbar. Die Zunahme der DMOS-Werte insgesamt entspricht jedoch nicht den Abständen der wahrgenommen Bildqualität bei

Betrachtung der vollständigen Bildserie. Insgesamt ist ein DMOS-Wert für Bild sechs mit rund 70 Punkten (AS) sehr hoch. Die Kontrolle auf Konsistenz zeigt beim MLDS (b) wiederum sehr ähnliche Wahrnehmungen der Versuchspersonen. Der Verlauf der Kurven stimmt überein. Die DMOS-Werte (c: rot) der LIVE2 Datenbank zeigen Übereinstimmungen mit den DMOS-Werten des Experiments. Zwischen Bild drei und vier gibt es für LIVE2 DMOS und die eigenen DMOS-Messung eine Qualitätsverbesserung, wenn auch nur minimal. Der LIVE2 DMOS (c: rot) hat als einziger einen starken Anstieg und große Qualitätssprung zwischen Bild eins und zwei. Beide Kurven haben ebenfalls einen Qualitätssprung von Bild vier bis sechs. MLDS (c: grün) repräsentiert dem gegenüber die subjektiv wahrgenommenen Unterschiede sehr gut. Bilder eins bis fünf der Serie haben vergleichbare Qualität, was sich in einer flachen MLDS Kurve widerspiegelt. MLDS detektiert auch den Qualitätsverlust von Bild sechs.

Die Qualitätsunterschiede der Bildserie *FastFading Ocean* (Abb. 25) sind in Originalgröße erkennbar. Hierbei gibt es einen erkennbaren Qualitätsverlust von Bild zwei bis vier. Eine Verbesserung von Bild vier zu fünf und wieder eine Verschlechterung von Bild fünf zu sechs.

Die DMOS Kurven beider Versuchspersonen (a) machen deutlich, dass keine kontinuierliche Verschlechterung der Bilder wahrgenommen wird. Bild fünf wird sowohl von Versuchsperson GA (a: blau) als auch Versuchsperson AS (a: pink) als deutlich besser bewertet, als die benachbarten Bilder vier und sechs. Die Versuchspersonen untereinander haben eine ähnliche Wahrnehmung des Qualitätsverlaufs der Bildserie, wobei sie sich jedoch im Grad der Qualitätsbewertung unterscheiden. Der maximale DMOS der Bildserie liegt für GA nur bei 36 für Bild sechs im Gegensatz dazu für AS bei 75 für Bild vier. Insgesamt (c: gelb) haben die Bilder eins und zwei sehr ähnlich und gute DMOS-Werte. Von Bild zwei zu drei ist in Originalgröße ein deutlicher Qualitätsverlust erkennbar, der sich in den DMOS-Werten mit einem Anstieg von rund 60 Punkten widerspiegelt. Der Abstand von Bild drei und vier mit weniger als zwei Punkten bewertet diese Bilder als gleichwertig. Interessant ist eine hohe Qualitätsverbesserung von vier zu fünf mit DMOS-Werten von 61 für Bild vier und 28 für Bild fünf. Eine Verbesserung der Qualität bei zunehmendem Verzerrungsgrad tritt mit solchem Ausmaß ausschließlich für die Verzerrungsart FastFading auf. Bei der Betrachtung der MLDS Kurven beider Versuchspersonen (b) ist die Konsistenz groß. Beide Kurven sind durchgehend flach, was für eine Homogenität der Bilder sprechen würde. Die Auffälligkeiten, die durch die DMOS-Werte repräsentiert werden, werden bei der Messung mit MLDS nicht erfasst. Die Gegenüberstellung (c) von DMOS (c: gelb), DMOS LIVE2 (c: rot) und MLDS (c: grün) zeigt, dass MLDS an dieser Stelle am wenigsten geeignet ist, die Besonderheiten der Bildserie wiederzugeben. DMOS und DMOS LIVE2 zeigen beide den Effekt einer Qualitätsverbesserung von Bild vier zu fünf. Der maximale DMOS-Wert für LIVE2

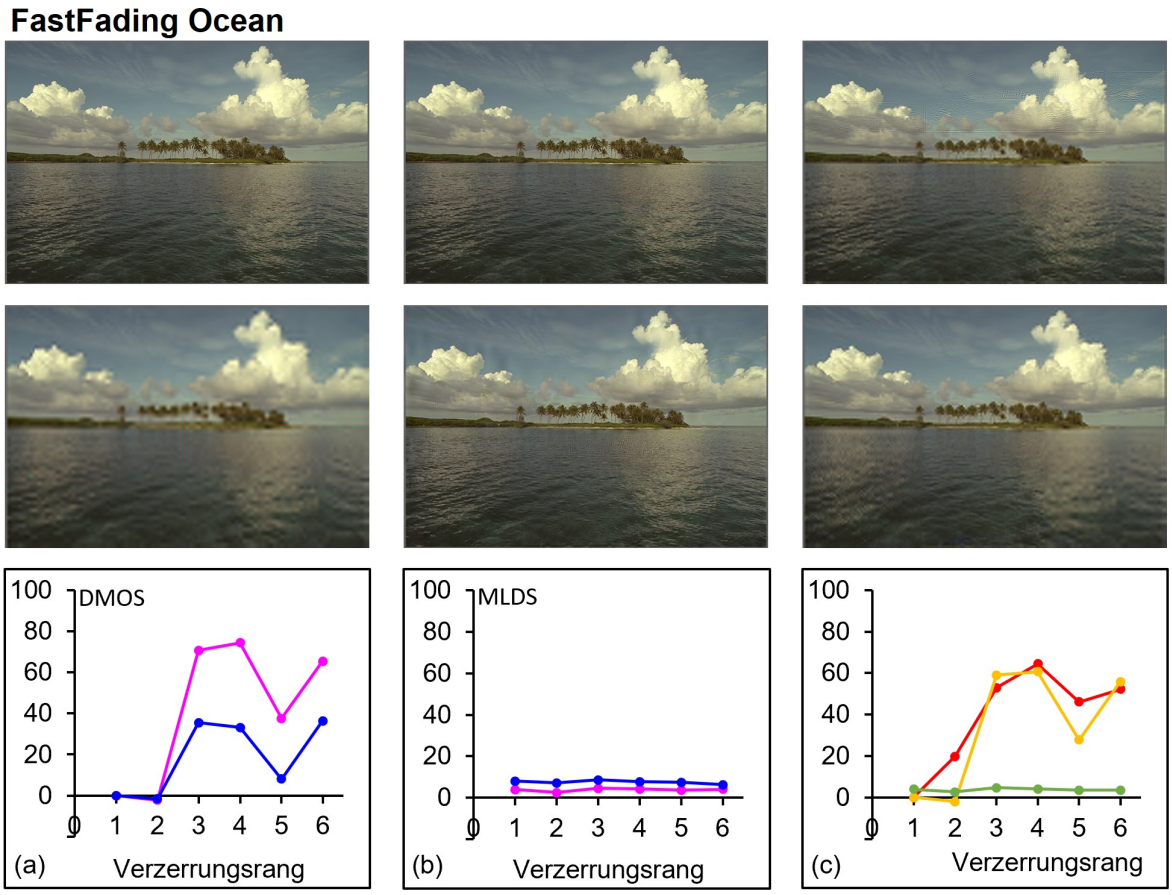


Abb. 25: Motiv *Ocean*, Verzerrungsart *Fast Fading*, Verzerrung: Rang 1 bis 6, Bilder: img147, img10, img9, img8, img7, img6. (a) DMOS, (b) MLDS, (a, b): blau: Versuchsperson GA, pink: Versuchsperson AS, (c) rot: DMOS LIVE2, gelb: DMOS beider Versuchspersonen, grün: MLDS beider Versuchspersonen. Legende: siehe Abb. 22. Details: siehe Text.

liegt bei 64 und ähnelt damit dem DMOS-Wert des Experiments von 61. Beide, LIVE2 DMOS als auch der DMOS des Experiments, bewerten Bild vier mit der schlechtesten Bildqualität.

3.3 Zwischenzusammenfassung

Im vorherigen Kapitel wurden die zwei eigenen Experimente zum MOS und MLDS vorgestellt. Die Methodik beider Experimente wurde erklärt und vier exemplarische Ergebnisse vergleichend dargestellt. Die individuellen Bewertungen der Versuchspersonen wurden erläutert und einander gegenübergestellt. Es erfolgte eine Verknüpfung der subjektiven Bildqualitätseindrücke mit den Kurvenverläufen zu LIVE2 DMOS sowie DMOS und MLDS der eigenen Experimente. Hierbei wurde besonders auf die Konsistenz und den Verlauf der Kurven in Bezug auf die Qualitätsmetrik der gesamten Bildserien eingegangen.

4 Diskussion

Die subjektiven Bewertungen der Bildqualitätsdatenbanken (Kapitel 1.3) sind die zentrale Grundlage für die Entwicklung von Bildqualitätsalgorithmen. Für hochwertige Bildqualitätsalgorithmen ist es daher erforderlich, dass die subjektiven Bewertungen der Bildqualitätsdatenbanken selber eine hohe Qualität aufweisen. Aufgrund dieser Schlüsselfunktion wurde die weit referenzierte LIVE2 Bildqualitätsdatenbank in dieser Arbeit systematisch untersucht. Das Qualitätsmaß der LIVE2 Datenbank ist dabei der DMOS. Aufgrund der in dieser Arbeit festgestellten Defizite der publizierten DMOS-Bewertungen der LIVE2 Datenbank wurde der Fragestellung dieser Arbeit in einer eigenen experimentellen MOS-Untersuchung exemplarischer LIVE2-Bilder weiter nachgegangen. Aufgrund der bekannten psychophysischen Beobachtungen zur vergleichenden Bewertung von Stimulusabfolgen wurde darüber hinaus eine weitere experimentelle Untersuchung nach dem MLDS Verfahren durchgeführt. Hierbei ging es darum, ein alternatives Qualitätsmaß zum MOS zu untersuchen.

Die Diskussion der Ergebnisse (Kapitel 2, 3) stützt sich auf die folgenden Anforderungen an ein Qualitätsmaß, die anhand der eigenen Auseinandersetzung mit dem LIVE2 DMOS entwickelt wurden.

(M1) Das Maß soll surjektiv sein: Bilder von subjektiv exzellenter bis sehr schlechter Bildqualität sollen auf den gesamten Zielraum abgebildet werden.

(M2) Das Maß soll bijektiv sein: Bilder mit subjektiv ähnlicher Qualität sollen ein ähnliches Qualitätsmaß erhalten. Umgekehrt soll ein ähnliches Qualitätsmaß Bilder mit ähnlicher Qualität kennzeichnen.

(M3) Das Maß soll transitiv sein: Eine Abfolge von Bildern mit subjektiv abnehmender Bildqualität soll eine korrespondierende Abfolge des Qualitätsmaßes aufweisen.

(M4) Die subjektiven Qualitätsabstände einer Abfolge von Bildern soll sich in der Abstandsmetrik des Qualitätsmaßes wiederfinden.

4.1 Zur Untersuchung des LIVE2 DMOS als Metrik der subjektiven Bildqualität

Zum Zeitpunkt ihrer Veröffentlichung in 2005 galt die LIVE2 Datenbank als eine der größten verfügbaren subjektiven Bildqualitätsdatenbanken. Ziel der Autoren war es, *Ground Truth* Daten zu Bildern und ihren Qualitätsbewertungen durch Versuchspersonen für die Entwicklung und Evaluierung von Bildqualitätsalgorithmen bereitzustellen

(Sheikh, Sabir & Bovik, 2006). Die LIVE2 Datenbank umfasst dazu 982 Bilder, davon 779 verzerrte Bilder, und ihre jeweiligen Qualitätsbewertungen. Die gewählten Verzerrungsarten sollen typische Bildveränderungen in alltäglichen Anwendungsszenarien repräsentieren: Kompressionsalgorithmen (*JPEG2000*, *JPEG*), Weichzeichner (*Gaussian Blur*), Rauschen (*White Noise*) und Datenverluste im WLAN (*Fast Fading*). Mit hohem Aufwand wurden insgesamt etwa 25.000 Bewertungen zur subjektiver Qualität der Bilder von Versuchspersonen erhoben. Über die Jahre hat sich die LIVE2 Datenbank als ein Referenzstandard etabliert, der auch bei der Entwicklung aktueller Bildqualitätsalgorithmen unverändert zum Einsatz kommt (Geng, Dong & Huang, 2022; Ma et al., 2021; Rehman, Nizami & Majid, 2022; Sun, Yu, Xu, Zhou & Chen, 2022; Zhang, Ma, Zhai & Yang, 2021). Das hierbei verwendete Qualitätsmaß ist der DMOS. Der MOS eines Bildes wurde als mittlere subjektive Bewertung über jeweils eine Gruppe von 20 - 29 Versuchspersonen ermittelt. Der DMOS wurde als MOS Differenz eines verzerrten Bildes zum MOS seines Referenzbildes ermittelt. Das unverzerrte Referenzbild erhält demnach einen DMOS von null. Der Zielraum der DMOS-Werte umfasst Werte zwischen null (exzellent) und 100 (schlecht).

Wie im Kapitel 2.2 zu den LIVE2 DMOS Ergebnissen dargestellt, findet die systematische Untersuchung aller LIVE2 Bilder zahlreiche Defizite des DMOS als Qualitätsmaß, die im Folgenden zusammengefasst und diskutiert werden.

4.1.1 LIVE2: Überprüfung der Qualitätsanforderungen M1 - M4

Die DMOS-Werte (Abb. 15) geben das Qualitätsspektrum $[0, 100]$ nicht wieder. Alle Referenzbilder haben per Definition einen DMOS von null. Der nächstkleinere DMOS liegt bereits bei 17,9 Punkten, obwohl die LIVE2 Datenbank zahlreiche geringfügig verzerrte Bilder enthält, deren Qualität sich kaum vom Referenzbild unterscheidet. Umgekehrt beträgt der maximale DMOS 84,5 Punkte, obwohl die Datenbank etwa 20 Bilder enthält, die aufgrund starker Verzerrungen praktisch nicht mehr erkennbar sind. Für diese Bilder würde man DMOS-Werte in der Nähe von 100 erwarten. Insgesamt zeigt Abbildung 15 den von mehrstufigen Likert-Skalen bekannten Effekt einer Tendenz zur Mitte, bei der die Versuchspersonen Bewertungen mit den extremen Skalenwerten typischerweise vermeiden (Schnell, Hill & Esser, 2018). Die Orientierung der Versuchspersonen an den zentralen semantischen Skalen-Bezeichnungen der LIVE2 Messungen (*bad*, *poor*, *fair*, *good* und *excellent*) (Sheikh et al., 2006) lässt sich auch anhand der Häufungspunkte bei den DMOS-Werten um 25 und 50 vermuten. Kriterium M1 ist damit verletzt.

Auch Kriterium M2 wird vom LIVE2 DMOS nicht erfüllt. Knapp 30% der Bilder mit einem DMOS ≥ 70 haben eine gute bis sehr gute subjektive Bildqualität (Abb. 8). Bei

den anderen Bildern in diesem DMOS-Bereich ist das nicht der Fall. Auch die Abbildungen 9, 10 und 11 belegen exemplarisch, welche geringe Aussagekraft zur Bildqualität dem LIVE2 DMOS zugemessen werden kann. Bei fast identischem DMOS-Wert reicht die Bildqualität von gut bis gar nicht erkennbar.

Abbildung 13 gibt in der oberen Zeile eine Bildfolge wieder, die Kriterium M3 verletzt. Die Bilder sind in der Reihenfolge abnehmender Bildqualität sortiert, der DMOS-Wert des rechten Bildes ist dennoch besser als der des mittleren. Tabelle 1 zeigt, dass es sich hierbei nicht um einen Einzelfall handelt.

Die weitreichende Verletzung des Kriteriums M4 wird wiederum in Abbildung 12 deutlich. Dargestellt ist eine exemplarische Bildserie für jede der fünf Verzerrungsarten der LIVE2 Datenbank. Für jede dieser fünf Bildserien ist der subjektive Qualitätsunterschied zwischen dem Referenzbild in der linken Spalte und dem verzerrten Bild in der mittleren Spalte gering. Demgegenüber besteht ein erheblich größerer subjektiver Qualitätsabstand vom mittleren zum rechten Bild, das jeweils eine deutlich schlechtere Qualität aufweist oder auch gar nicht mehr erkennbar ist. Diese Metrik der subjektiven Qualitätsabstände wird durch die DMOS-Bewertungen nicht wiedergegeben. Der DMOS-Abstand vom linken zum mittleren Bild liegt für alle fünf Verzerrungsarten bei um die 50 Punkte. Dagegen nimmt der DMOS vom mittleren zum rechten Bild lediglich um zehn bis 25 Punkte zu, obwohl die subjektive Bildqualität erheblich stärker abnimmt als vom linken zum mittleren Bild. Die DMOS-Abstände bilden die subjektive Qualitätsmetrik also in keiner Weise ab.

Zusammenfassend zeigt die systematische Analyse erhebliche Einschränkungen des LIVE2 DMOS als Qualitätsmaß für die Bilder der Datenbank. Bei allen oben formulierten Anforderungen an ein Qualitätsmaß finden sich teils erhebliche Defizite. Besonders bemerkenswert ist dabei die erhebliche Verletzung der subjektiven Abstandsmetrik bei der Bewertung von Bildfolgen.

Die in dieser Arbeit festgestellten Defizite der LIVE2 DMOS-Bewertungen stehen damit in Widerspruch zum Anspruch der Autoren, mit LIVE2 eine hochwertige Datenbasis für Bildqualitätsalgorithmen und die automatisierte Bewertung von Bildern in Übereinstimmung mit der menschlichen Qualitätsbewertung bereitzustellen.

4.1.2 LIVE2: Zusammenfassung und Ausblick

Im folgenden wird auf mögliche Ursachen eingegangen, die den Defiziten der LIVE2 DMOS-Bewertungen zugrunde liegen sein könnten. Der Publikation (Sheikh et al., 2006) ist zu entnehmen, dass jede einzelne Versuchsperson jeweils nur eine einzige

Verzerrungsart bewertete. Dieses Vorgehen könnte zusätzlich zur Divergenz von DMOS und Bildqualität zwischen unterschiedlichen Verzerrungsarten beitragen.

Auch die Bildung eines Mittelwertes ordinal skalierter Messwerte wird in der Literatur kritisiert (Bosse, 2018). Sheikh et al. (2006) versuchen dieses Problem zu umgehen, indem sie Messwerte auf einer kontinuierlichen Skala erheben, obwohl diese mit den semantischen Kategorien einer 5-stufigen Likert-Skala belegt ist. Die Tendenz zur Mitte und die Häufungsgipfel der DMOS-Werte bei den Skalenwerten *good* und *fair* könnten Ausdruck einer semantischen Interferenz bei der subjektiven Qualitätsbewertung sein.

Ein weiterer testpsychologischer Effekt beschreibt, dass Versuchspersonen auch bei Einzelstimulus-Verfahren ihre aktuelle Bewertung nach einem - bewussten oder unbewussten - Vergleich mit den bereits zuvor abgegeben Bewertungen vornehmen (Serial Dependencies) (Schnell et al., 2018). Die Bestrebung, unterschiedlich wahrgenommenen Bildern auch unterschiedliche Bewertungen zu geben, kann in Verbindung mit einer 5-stufigen Likert-Skala dazu führen, dass sehr ähnliche Bilder trotzdem mit unterschiedlichen Skalenwerten und damit Unterschieden von 25 bis 50 Punkten beurteilt werden (Schnell et al., 2018).

Zusätzlich können Lerneffekte die Bewertung der Bilder beeinflussen. Nachdem die Versuchsperson mit den Motiven der Referenzbilder vertraut ist, wird sie auch bei stärker verzerrten Bildern noch Details erkennen, die sie ohne diesen Lernvorgang nicht wahrnehmen würde. Das kann dazu beitragen, verzerrte Bilder besser zu bewerten.

Sheikh et al. (2006) berichten ferner, dass die einzelnen Bewertungsdurchgänge auf 30 min begrenzt wurden. Systematischen Abweichungen der psychometrischen Skala der Versuchspersonen zwischen den einzelnen Sitzungen sollte durch ein abschließendes *Realignment*-Experiment abgeholfen werden. Zusätzlich wurden die Rohwerte der Bewertungen jeder Versuchsperson individuell mit Mittelwert und Standardabweichung z-normiert und anschließend wieder linear auf DMOS-Werte transformiert (Sheikh et al., 2006). Die Auswirkungen dieser Transformationen der subjektiven MOS-Bewertungen zum DMOS-Wert der LIVE2 Datenbank sind nicht unmittelbar offensichtlich. Es wäre von Interesse, diese Auswirkungen anhand der Rohdaten zu untersuchen. Die auf der Download-Seite der LIVE2 Datenbank angekündigte Veröffentlichung der Rohdaten ist seit 2005 allerdings nicht erfolgt.

Eine Untersuchung der Rohdaten wäre auch deshalb von besonderem Interesse, da die in der LIVE2 Datenbank bereitgestellten Daten offenbar gravierend von den in der Publikation (Sheikh et al., 2006) berichteten Bewertungen abweichen. Die in dieser Arbeit festgestellte Tendenz der LIVE2 Bewertungen zur Mitte (Abb. 15, 26a) ist in Fig. 5 der Publikation nicht festzustellen (Abb. 26b). Auch der Verlauf der Häufigkeitsverteilung

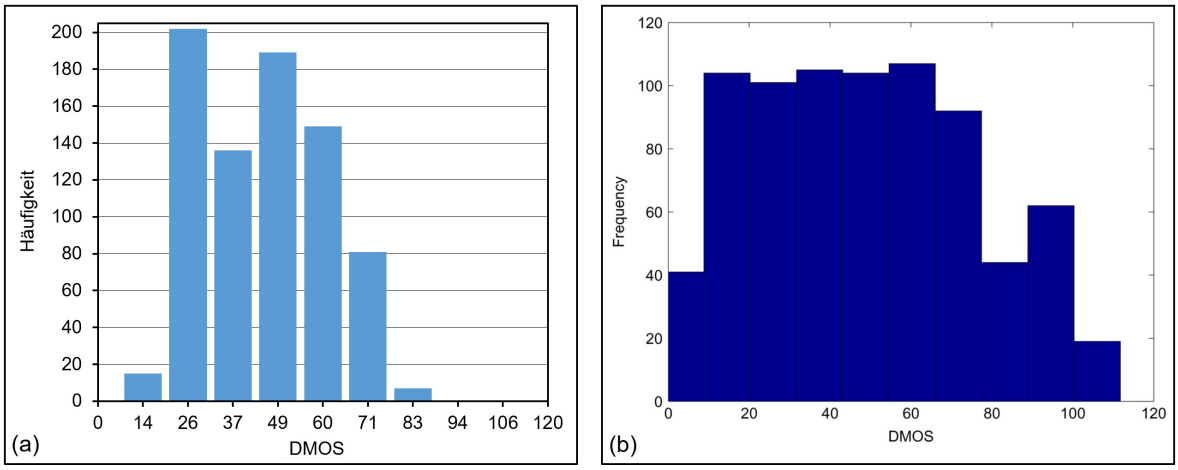


Abb. 26: Histogramm aller DMOS-Bewertungen der LIVE2 Datenbank: (a) Häufigkeit aller DMOS-Bewertungen aus `http://live.ece.utexas.edu/research/quality`, Bining wie in Sheikh et al. (2006) Fig. 5, (b) Häufigkeit der DMOS-Bewertungen nach Sheikh et al. (2006) Fig. 5. Die in Sheikh et al. (2006) präsentierten Daten scheinen nicht mit den Daten der LIVE2 Datenbank übereinzustimmen. Details: siehe Text.

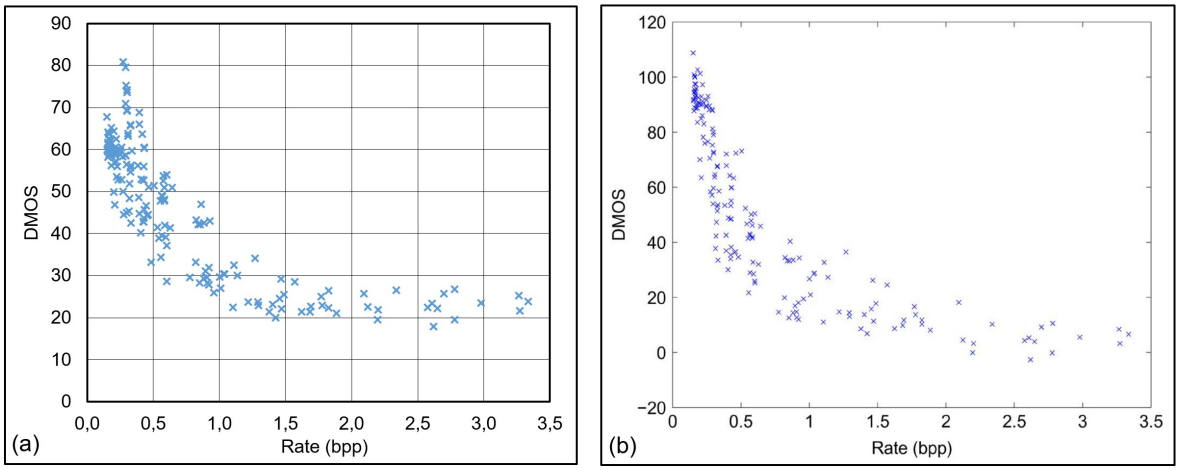


Abb. 27: Verteilung aller DMOS-Bewertungen zur Verzerrungsart JPEG der LIVE2 Datenbank: (a) Verteilung der DMOS-Bewertungen aus `http://live.ece.utexas.edu/research/quality`, (b) Verteilung der DMOS-Bewertungen nach Sheikh et al. (2006) Fig. 4b. Die in Sheikh et al. (2006) präsentierten Daten scheinen nicht mit den Daten der LIVE2 Datenbank übereinzustimmen. Details: siehe Text.

der LIVE2 DMOS in Abbildung 26a entspricht nicht der Gleichverteilung in Sheikh et al. (2006), Fig. 5 (Abb. 26b).

Abbildung 27 zeigt ergänzend eine exemplarische Gegenüberstellung der *JPEG*-Bewertungen der LIVE2 Daten mit Fig. 4b der Publikation, die diese Diskrepanz bestätigt. Die DMOS-Bewertungen der LIVE2 Bilder liegen im Intervall $[18, 81]$ (Abb. 27a), die Bewertungen der Publikation im Intervall $[-2, 110]$ (Abb. 27b). Nicht nur der Wertebereich, sondern auch die Verteilung der *JPEG*-Bewertungen zeigt deutliche Abweichungen. Bei den LIVE2 Daten finden sich oberhalb der Häufung um 60 noch etwa

14 Datenpunkte > 65. In Fig. 4b der Publikation sind oberhalb der Häufung um 90 nur noch vier bis fünf Datenpunkte > 95 zu erkennen. Gleichzeitig besteht eine relativ gute Übereinstimmung der LIVE2 Daten mit der Publikation im Bereich niedriger Verzerrungsgrade mit bpp $> 1,5$. Die Abweichungen entsprechen daher offenbar keiner einfachen, linearen Transformation.

Für künftige Forschung wäre es wünschenswert, die Diskrepanz zwischen der LIVE2 Datenbank und der zugehörigen Publikation (Sheikh et al., 2006) aufzuklären.

Ein weiterer Ansatz zur Fortführung der Fragestellung dieser Arbeit wäre eine Untersuchung der TID2013 Bildqualitätsdatenbank (Ponomarenko et al., 2013). Da TID2013 ebenfalls auf MOS als Qualitätsmaß der subjektiven Bewertungen basiert, wäre es interessant zu überprüfen, ob die gefunden Defizite der DMOS-Bewertungen der LIVE2 Datenbank sich dort wiederfinden lassen.

Unabhängig von den Defiziten der DMOS-Bewertung konnte die Analyse der LIVE2 Datenbank einen interessanten Effekt bei den Bildserien zur Verzerrungsart *Fast Fading* aufzeigen (Abb. 13 unten, 14). Bei etwa 50% dieser Bildserien mit *Fast Fading* findet sich ein Bereich der Verzerrung, in dem sich die Bildqualität trotz zunehmender Verzerrung für einzelne Bilder wieder bessert. Erst bei weiter zunehmender Verzerrung fällt die Bildqualität dann wieder ab. Dabei variiert der genaue Bereich des Signal-Rausch-Verhältnisses, in dem dieser Effekt zu beobachten ist, zwischen den verschiedenen Bildserien. Die Beobachtung dieses Effektes basiert auf der eigenen subjektiven Qualitätsbewertung der verzerrten Bilder und wird von Defiziten der DMOS-Bewertungen nicht berührt. Bei anderen Verzerrungsarten als *Fast Fading* ist diese Diskontinuität der Bildqualität nicht festzustellen.

Im folgenden Abschnitt werden die Ergebnisse der eigenen experimentellen Untersuchungen in Zusammenhang mit den oben formulierten Anforderungen (M1 - M4) an ein Qualitätsmaß gesetzt. Anschließend werden diese differenziert diskutiert.

4.2 Zum Vergleich experimenteller DMOS und MLDS als Metriken der subjektiven LIVE2 Bildqualität

Für das eigene MOS und MLDS Experiment wurden aus den LIVE2 Bildern die vier Motive *Bikes*, *ChurchAndCapitol*, *Ocean* und *WomanHat* mit den Verzerrungsarten *JPEG*, *White Noise*, *Gaussian Blur* und *Fast Fading* und jeweils sechs Verzerrungsgraden ausgewählt (Abb. 17, A1 - A4). Bei der Auswahl der Motive wurde außer auf die analysierten LIVE2 Auffälligkeiten auch auf unterschiedliche Inhaltskategorien Social (*Bikes*), Outdoor man-made (*ChurchAndCapitol*), Outdoor natural (*Ocean*) und

Portrait (*WomanHat*) (Lévêque et al., 2020) geachtet. Die gewählte Aufbereitung der Ergebnisse als Gegenüberstellung von Bildserie und Messwerten (Abb. 22 - 25) erlaubt eine unmittelbare Beurteilung der diskutierten Anforderungen (M1 - M4) an das jeweilige Qualitätsmaß. Dies gilt für die Qualitätsbewertungen sowohl der einzelnen Bilder als auch der Abstandsmetrik innerhalb der Bildserie im Vergleich.

4.2.1 DMOS und MLDS: Überprüfung der Qualitätsanforderungen M1 - M4

Bei den eigenen Experimenten zeigen die DMOS-Messungen mit Werten im Intervall $[-2, 98]$ eine besser Überdeckung des Zielraums als die Werte des DMOS LIVE2 (Abb. 15). Kriterium (M1) wäre damit, ähnlich wie bei Sheikh et al. (2006) Fig. 5, noch erfüllt. Ansonsten finden sich die gleichen Defizite wie bereits für DMOS LIVE2 diskutiert.

Bei den meisten Bildserien besteht eine gute bis sehr gute innere Konsistenz zwischen den beiden Versuchspersonen in dem Sinne, dass die Kurven mit zunehmendem Verzerrungsgrad konsistent steigen oder fallen. Dennoch findet sich bei elf der 16 Kurven mindestens ein Messwert, bei dem die Bewertung desselben Bildes durch die beiden Versuchspersonen um mehr als 20 DMOS-Punkte abweicht (Abb. 24 - 25, A13 - A20). Im Einzelfall beträgt die Divergenz sogar bis 50 DMOS-Punkte (Abb. A14 Ocean). Dieselbe Bildqualität führt also zu sehr unterschiedlichen DMOS-Bewertungen. Umgekehrt findet sich ein DMOS > 90 für noch sehr gut erkennbare (Abb. A13 Bikes) bis fast gar nicht mehr differenzierbare Bilder (Abb. A16 Ocean). Derselbe DMOS ist also sehr unterschiedlichen Bildqualitäten zugeordnet. Anforderung (M2) an das Qualitätsmaß ist damit auch bei den eigenen Messungen verletzt. Die starke Divergenz einzelner Bewertungen macht noch einmal die Problematik der DMOS-Mittelwerte über mehrere Versuchspersonen zusätzlich zu den bereits oben diskutierten Aspekten deutlich.

In gleicher Weise wie bei DMOS LIVE2 zeigen auch die eigenen Experimente Verletzungen der Transitivität der DMOS-Bewertungen (M3) (Abb. 24 - 25, Abb. A14 WomanHat, Abb. A17 ChurchAndCapitol). Auch die Abstandsmetrik der Qualitätsbewertungen (M4) nach DMOS stimmt wie bei DMOS LIVE2 nicht mit den subjektiven Qualitätsabständen innerhalb der Bildserie überein.

Die bisherige Diskussion macht immer wieder die Defizite der MOS-Bewertungen der einzelnen Bilder im Vergleich zur jeweiligen Bildserie deutlich. Demgegenüber ergibt die vergleichende Betrachtung aller Bilder der gesamten Bildserie einen Bezugsrahmen, der die formulierten Anforderungen (M1 - M4) an ein Qualitätsmaß unmittelbar erfüllt.

Ausgehend von diesen Überlegungen ist daher ein Qualitätsmaß vergleichender Bildpräsentationen in Gegenüberstellung zur Einzelstimulus-Präsentation des DMOS von Interesse. Da solche Daten für die LIVE2 Bilder bisher nicht publiziert sind, wurde dieser Aspekt im experimentellen Teil dieser Arbeit aufgegriffen. Als Multistimulus-Design für die vergleichende Bewertung von Bildern wurde dabei die Messung der psychophysischen Funktion mittels MLDS eingesetzt.

Bei den MLDS-Experimenten zeigen fast alle Bildserien eine sehr hohe Übereinstimmung der Bewertungen beider Versuchspersonen. Nur bei zwei der 16 Kurven findet sich ein Messwert, bei dem die Bewertung desselben Bildes durch die beiden Versuchspersonen um mehr als 20 MLDS-Punkte abweicht (Abb. A14 Ocean, A15 ChurchAndCapitol).

Ebenso findet sich bei fast allen Bildserien der Abbildungen 22 - 25, A13 - A20 eine sehr hohe Übereinstimmung der MLDS-Bewertungen mit den subjektiven Abstandsmetriken der Bildserien. Die Qualitätssprünge werden typisch bildgenau in den MLDS-Verläufen wiedergegeben. Eine Ausnahme stellt nur Abbildung A20 Ocean dar, bei der die MLDS-Bewertungen die leichte Qualitätsverbesserung beim Verzerrungsrang fünf nicht widerspiegeln. Insgesamt werden die Anforderungen Transitivität (M3) und Abstandsmetrik (M4) von den MLDS-Bewertungen für die einzelnen Bildserien erheblich genauer erfüllt, als dies bei den MOS-Bewertungen der Einzelstimulus-Präsentationen der Fall ist. Dieses Ergebnis erscheint auch intuitiv plausibel, da das experimentelle Design der MLDS-Messung keine absolute Qualitätsbewertung, sondern einen binären Qualitätsvergleich der präsentierten Triaden aus den einzelnen Bildserien ermittelt. Dabei ist die Qualitätsverbesserung bei Abbildung A20 Ocean Rang fünf so gering, dass diese im Vergleich zu den übrigen Bildern der Serie nicht als signifikant bewertet wird.

Um eine Vergleichbarkeit von DMOS- und MLDS-Verläufen in den gemeinsamen Grafiken zu ermöglichen, wurden die Gesamtheit aller MLDS-Werte über alle Motive und Verzerrungsarten auf das Intervall [0, 100] abgebildet. Anforderung (M1) ist daher trivial erfüllt. Gleichzeitig wird Anforderung (M2) verletzt. Die am stärksten verzerrten Bilder der Verzerrungsart *WhiteNoise* erhalten MLDS-Werte von 20 bis 100 (Abb. A15, A16). Dabei wird das noch leidlich erkennbare Bild WomanHat (Abb. 23) mit 35 MLDS-Punkten bewertet, während das kaum erkennbare ChurchAndCapitol (Abb. A15) auf 20 Punkte kommt. Auch MLDS zeigt sich bei der vorliegenden Untersuchung als defizitär, Bilder mit ähnlicher subjektiver Qualität auf ähnliche Bewertungen abzubilden (M2). Als Konsequenz folgt, dass die guten Ergebnisse von MLDS bei Transitivität (M3) und Abstandsmetrik (M4) jeweils nur innerhalb der einzelnen Bildserie gültig sind. Damit (M3) und (M4) auch beim Vergleich der Bildqualität zwischen

verschiedenen Motiven und Verzerrungsarten Bestand haben, wäre die Erfüllung von (M2) die Voraussetzung.

4.2.2 DMOS und MLDS: Zusammenfassung und Ausblick

Trotz der dargestellten Defizite stimmt die DMOS Abstandsmetrik aus den eigenen Experimenten in vielen Bereichen grundsätzlich besser mit den subjektiven Qualitätsabständen der Bildserien überein als die der DMOS LIVE2 Bewertungen (Abb. 22 - 25, A13 - A20: vgl. Bildserien mit DMOS LIVE2, c: rot, und DMOS Experiment, c: gelb). Hierfür sind mehrere Erklärungsansätze denkbar: Bei der eigenen Messung bestand keine semantische Konnotation der MOS-Bewertung. Alle Motive und Verzerrungsarten wurden von allen Versuchspersonen bewertet. Beide Versuchspersonen sind vermutlich als Experten im Vergleich zu den meisten Teilnehmern der LIVE2 Untersuchung einzustufen. Für eine detaillierte Analyse der eigenen Messungen im Vergleich zur LIVE2 Datenbank wären erneut die LIVE2 Rohdaten von großem Interesse.

Die vergleichende Bewertung von Stimulusabfolgen ist aus der Psychophysik für verschiedene sensorische Qualitäten bekannt. Das menschliche Gehirn bewertet Stimulusdifferenzen im Vergleich verschiedener Ausprägungen eines Stimulus. Die psychophysische Forschung findet dazu typische Differenzskalen der menschlichen Wahrnehmung (Bosse, 2018; Kandel et al., 2021; Kingdom & Prins, 2016; Nutter & Esker, 2006; Wiebel et al., 2017).

Auch für die Qualitätsbewertung von Bildern in alltäglichen Szenarien stehen typisch Referenzinformationen zur Verfügung. Insbesondere die Betrachtung von Videoinhalten beinhaltet natürlicherweise eine vergleichende Abfolge von Bildern.

Insgesamt zeigt sich, dass das MLDS-Design aufgrund der vergleichenden Bewertung von Bild-Triaden deutlich intuitiver alltäglichen Bewertungsszenarien subjektiver Bildqualität entspricht. Bei MLDS bestimmt der Betrachter kein absolutes Qualitätsmaß des einzelnen Bildes, sondern gibt nur eine binäre Auskunft, welches Bilderpaar der Triade den subjektiv größeren Qualitätsabstand aufweist. Aus diesem Grund ist MLDS besser als MOS geeignet, die Transitivität und Abstandsmetrik der subjektiven Bildqualität innerhalb der Bildserien abzubilden. Was dagegen fehlt, ist ein absoluter Qualitätsvergleich zwischen den verschiedenen Bildserien, der dazu führt, das Bilder ähnlicher Qualität auch invariant von Motiv und Verzerrungsart vergleichbare Qualitätsmaße erhalten.

MLDS ist per Design darauf ausgelegt, psychophysische Differenzskalen zu bestim-

men. Wie bereits diskutiert, sind die Verzerrungsgrade der LIVE2 Bilder nicht homogen verteilt, sondern bei geringen Verzerrungsgraden konzentriert. Es erscheint daher nicht verwunderlich, dass die damit bestimmbaren Differenzskalen nicht gut invariant von Motiv und Verzerrungsart skalieren. Ein Ansatz für weitere Untersuchungen wäre daher, das Spektrum der Verzerrungsgrade zu allen Motiven und Verzerrungsarten deutlich auszuweiten. Damit sollte es möglich sein, die jeweilige psychophysische Differenzskala genauer abzubilden. In Verbindung mit dem jeweiligen Referenzbild als invariantem Referenzpunkt der Qualitätsbewertung könnte damit ein vergleichendes Qualitätsmaß über Bildserien hinweg realisierbar sein.

Sollte dieses Vorgehen alleine nicht zielführend sein, wäre als weiterer Ansatz einer Ausweitung des MLDS-Designs zu prüfen, bei dem die Triaden nicht alleine aus einer einzelnen Bildserie, sondern aus allen Motiven, Verzerrungsarten und Verzerrungsgraden gezogen werden. Für diese Ausweitung des Designs wäre zu vermuten, dass die gemessenen Qualitätsmaße dann Transitivität (M3) und Abstandsmetrik (M4) über alle Bildserien hinweg etablieren und damit dann auch (M2) über die Bildserien skaliert.

Sollte auch diese Erweiterung nicht genügen, wäre einer Erweiterung des experimentellen Designs als MLDS+ (Forced Choice + Magnitude Estimation) denkbar, bei dem die Versuchsperson aufgefordert wird, zusätzlich zum binären (Forced Choice) Qualitätsvergleich der Bild-Triaden einen absoluten Abstand (Magnitude Estimation) der Bildpaare zu erfassen. Dabei würden die zentralen Probleme der absoluten MOS-Bewertungen durch das Multistimulus-Design vermieden. Gleichzeitig könnte auch dieser Ansatz geeignet sein, die psychophysische Differenzskalen invariant von Motiv und Verzerrungsart so zu skalieren, dass subjektiv vergleichbare Bildqualitäten auch ähnliche Qualitätsmaße erhalten.

4.3 Fazit

Bilder und Videos haben sich zu einem zentralen Bestandteil unserer digitalen Gesellschaften entwickelt. Smartphones sind heute weniger Telefone als vielmehr Bilddatenbanken. Durch neue Technologien hat die Qualität und das Datenvolumen dieser Bilder erheblich zugenommen. Der jährliche Austausch von Bildern und Videos in sozialen Medien und Streamingdiensten geht in die Milliarden (Statista, 2021; Zhai & Min, 2020). Dabei ist beim stetigen Wachstum des Datenvolumens bisher keine Begrenzung erkennbar (Statista, 2021). Wachsende Datenmengen kontrastieren mit begrenzten Übertragungskapazitäten. Vor diesem Hintergrund sind Kompromisse bei der Übertragungsqualität unvermeidbar. Kompressionsalgorithmen reduzieren die Auflö-

sung der Bildinhalte, Bildraten eines Videos werden begrenzt, Übertragungsfehler werden nicht vollständig kompensiert. Gleichzeitig hat mit den neuen Technologien zum einen die Qualität von visuellen Inhalten, zum anderen korrespondierend auch der Anspruch der Nutzer zugenommen. Eine Internetseite soll zügig aufgebaut werden, Bilder und Videos sollen detailgenau und nicht verpixelt dargestellt werden, Filme sollen nicht von Übertragungsverzögerungen unterbrochen werden. Der Zugang zu Bild und Videoinhalten soll außerdem bei hohen Bewegungsgeschwindigkeiten in der Bahn genauso gut funktionieren wie stationär in der eigenen Wohnung.

Die Umsetzung dieser Anforderungen innerhalb der Grenzen der verfügbaren Technologien erfordert eine umfassende adaptive Qualitätssteuerung der übertragenen Bildinhalte. Aufgrund der gewaltigen Datenmengen und hohen Prozessgeschwindigkeit können eine hohe Bildqualität und Nutzerzufriedenheit dabei nur mit einer algorithmischen Bewertung der Bildqualität gewährleistet werden. Das entscheidende Kriterium (*Ground truth*) für Bildqualitätsalgorithmen ist hierbei immer die subjektive Qualitätsbewertung durch die Betrachter. Gute Bildqualitätsalgorithmen müssen daher darauf abzielen, die subjektive menschliche Bewertung der Bildqualität möglichst gut wiederzugeben. Referenzstandards für die Entwicklung verlässlicher Bildqualitätsalgorithmen sind dabei Bildqualitätsdatenbanken mit von Versuchspersonen erhobenen subjektiven Qualitätsbewertungen. Validität und Reliabilität der subjektiven Bildqualitätsdatenbanken bestimmen daher die Vorhersagequalität der Bildqualitätsalgorithmen. Ein Qualitätsalgorithmus kann nicht besser werden als die Qualitätsdatenbanken, auf denen er basiert. Dies gilt heutzutage um so mehr, nachdem auch Bildqualitätsalgorithmen weniger regelbasiert, sondern vermehrt als neuronale Netze implementiert und mit Qualitätsdatenbanken trainiert werden (Geng et al., 2022; Ma et al., 2021; Zhang et al., 2021).

Zur Fragestellung dieser Arbeit lässt sich zusammenfassend festhalten: MOS ist kein geeignetes Maß für die subjektive Bildqualität der LIVE2 Datenbank. Bilder vergleichbarer Qualität erhalten deutlich divergente Qualitätsmaße. Transitivität und Abstandsmetrik der MOS-Bewertungen sind erheblich defizitär. Die LIVE2 Datenbank erscheint daher für die Entwicklung und das Training objektiver Bildqualitätsalgorithmen schlecht geeignet. Von besonderem Interesse wäre es, in weitergehenden Untersuchungen die Diskrepanz zwischen der veröffentlichten LIVE2 Bildqualitätsdatenbank und der publizierten Auswertung (Sheikh et al., 2006) anhand der Rohdaten der Versuchspersonen aufzuklären.

Das in dieser Arbeit zusätzlich experimentell untersuchte MLDS-Design erweist sich mit seiner Multistimulus-Präsentation der MOS-Bewertung deutlich überlegen. Transitivität und Abstandsmetrik der subjektiven Bewertungen entsprechen weitgehend

der subjektiven Metrik der einzelnen Bildserien. Die schwache Invarianz der absoluten Skalierung der MLDS-Bewertungen über Motive und Verzerrungsarten hinweg wäre weiter zu untersuchen. Hierbei könnten zum einen eine größeres Spektrum an Verzerrungsgraden als auch eine Erweiterung des MLDS-Designs (MLDS+) geeignet sein, die psychophysische Differenzskala relativ zum unverzerrten Referenzbild geeignet zu skalieren.

Weitere Ansätze zur Fortsetzung dieser Arbeit wären die Untersuchung mit einer größeren Anzahl an Versuchspersonen sowie die Übertragung der vorliegenden Ergebnis auf andere subjektive Bilddatenbanken.

Literatur

Atidel, L., Bouridane, A., Viennet, E. & Haddadi, M. (2013, 09). Full-reference image quality metrics performance evaluation over image quality databases. *Arabian Journal for Science and Engineering*, *38*, 2327-2356. doi: 10.1007/s13369-012-0509-6

Bosse, S. (2018). *Data-driven estimation and neurophysiological assessment of perceived visual quality* (Doctoral Thesis, Technische Universität Berlin, Berlin). doi: 10.14279/depositonce-7233

Brunnström, K., De Moor, K., Dooms, A., Egger-Lampl, S., Garcia, M.-N., Hossfeld, T., ... Zgank, A. (2013). *Qualinet white paper on definitions of quality of experience.*

Chandler, D. (2013, 01). Seven challenges in image quality assessment: Past, present, and future research. *ISRN Signal Processing*, *2013*. doi: 10.1155/2013/905685

Charrier, C., Maloney, L., Cherifi, H. & Knoblauch, K. (2007, 12). Maximum likelihood difference scaling of image quality in compression-degraded images. *Journal of the Optical Society of America. A, Optics, image science, and vision*, *24*, 3418-26. doi: 10.1364/JOSAA.24.003418

Fiedler, M., Hossfeld, T. & Tran-Gia, P. (2010). A generic quantitative relationship between quality of experience and quality of service. *IEEE Network*, *24* (2), 36-41. doi: 10.1109/MNET.2010.5430142

Fleming, R., Jäkel, F. & Maloney, L. (2011, 06). Visual perception of thick transparent materials. *Psychological science*, *22*, 812-20. doi: 10.1177/0956797611408734

Geng, K., Dong, G. & Huang, W. (2022, 02). Robust dual-modal image quality assessment aware deep learning network for traffic targets detection of autonomous vehicles. *Multimedia Tools and Applications*, *81*. doi: 10.1007/s11042-022-11924-1

Goldstein, E. B. & Cacciamani, L. (2021). *Sensation and perception (mindtap course list)*. Cengage Learning.

Haas, J., Hass, R., Spocter, M. A. & de Sousa, A. A. (2020). Human visual neurobiology. In T. K. Shackelford & V. A. Weekes-Shackelford (Hrsg.), *Encyclopedia of evolutionary psychological science* (S. 1–10). Cham: Springer International Publishing. Zugriff auf `https://doi.org/10.1007/978-3-319-16999-6_2768-1` doi: 10.1007/978-3-319-16999-6_2768-1

ITU, R. (2019). Methodology for the subjective assessment of the quality of television images. *Recommendation ITU-R BT.500-14*. Zugriff auf `https://www.itu.int/dms_pubrec/itu-r/rec/bt/R-REC-BT.500-14-201910-I!!PDF-E.pdf`

Kandel, E. R., Koester, J. D., Mack, S. H. & Siegelbaum, S. A. (2021). *Principles of neural science, 6e* (6e Aufl.). McGraw Hill Professional.

Kingdom, F. A. & Prins, N. (2016). Varieties of psychophysical procedures. In

F. A. Kingdom & N. Prins (Hrsg.), *Psychophysics: A practival introduction: Second edition* (S. 37-54). San Diego: Academic Press. Zugriff auf https://www.sciencedirect.com/science/article/pii/B9780124071568000037 doi: https://doi.org/10.1016/B978-0-12-407156-8.00003-7

Knoblauch, K. & Maloney, L. (2012). *Modeling psychophysical data in r*. doi: 10.1007/978-1-4614-4475-6

Knoblauch, K. & Maloney, L. T. (2008). Mlds: Maximum likelihood difference scaling in r. *Journal of Statistical Software*, *25* (2), 1–26. Zugriff auf https://www.jstatsoft.org/index.php/jss/article/view/v025i02 doi: 10.18637/jss.v025.i02

Krantz, D. H., Luce, R. D., Suppes, P. & Tversky, A. (2006). *Foundations of measurement: Additive and polynomial representations* (Bd. 1).

Lahoulou, A., Viennet, E., Bouridane, A. & Haddadi, M. (2011). A complete statistical evaluation of state-of-the-art image quality measures. In *International workshop on systems, signal processing and their applications, wosspa* (S. 219-222). doi: 10.1109/WOSSPA.2011.5931456

Lévêque, L., Yang, J., Yang, X., Guo, P., Dasalla, K., Li, L., ... Liu, H. (2020). Cuid: A new study of perceived image quality and its subjective assessment. *CoRR*, *abs/2009.13304*. Zugriff auf https://arxiv.org/abs/2009.13304

Lindsey, D., Brown, A., Reijnen, E., Rich, A., Kuzmova, Y. & Wolfe, J. (2010, 09). Color channels, not color appearance or color categories, guide visual search for desaturated color targets. *Psychological science*, *21*, 1208-14. doi: 10.1177/0956797610379861

Ma, J., Wu, J., Li, L., Dong, W., Xie, X., Shi, G. & Lin, W. (2021). Blind image quality assessment with active inference. *IEEE Transactions on Image Processing*, *30*, 3650-3663. doi: 10.1109/TIP.2021.3064195

Martinez-Garcia, M., Bertalmío, M. & Malo, J. (2018, 01). In praise of artifice reloaded: Caution with subjective image quality databases.

Nahrstedt, K. (2012). Basics of quality of service in wireless networks. In *Quality of service in wireless networks over unlicensed spectrum* (S. 1–15). Cham: Springer International Publishing. Zugriff auf https://doi.org/10.1007/978-3-031-02482-5_1 doi: 10.1007/978-3-031-02482-5_1

Nutter, F., Jr & Esker, P. (2006, 02). The role of psychophysics in phytopathology: The weber–fechner law revisited. *European Journal of Plant Pathology*, *114*, 199-213. doi: 10.1007/s10658-005-4732-9

Peirce, J. W. (2007). Psychopy—psychophysics software in python. *Journal of Neuroscience Methods*, *162* (1), 8-13. Zugriff auf https://www.sciencedirect.com/science/article/pii/S0165027006005772 doi: https://doi.org/10.1016/j.jneumeth.2006.11.017

Ponomarenko, N., Ieremeiev, O., Lukin, V., Egiazarian, K., Jin, L., Astola, J., ... Kuo, C.-C. J. (2013, 06). Color image database tid2013: Peculiarities and preliminary results. In *European workshop on visual information processing (euvip).*

Rehman, M. U., Nizami, I. F. & Majid, M. (2022). Deeprpn-biqa: Deep architectures with region proposal network for natural-scene and screen-content blind image quality assessment. *Displays*, *71*, 102101. Zugriff auf `https://doi.org/10.1016/j.displa.2021.102101` doi: 10.1016/j.displa.2021.102101

Schnell, R., Hill, P. B. & Esser, E. (2018). *Methoden der empirischen sozialforschung.* De Gruyter Oldenbourg.

Sheikh, H., Sabir, M. & Bovik, A. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, *15* (11), 3440-3451. doi: 10.1109/TIP.2006.881959

Statista. (2021, 09). Video streaming (svod). , Online-Ressource. Zugriff auf `https://www.statista.com/outlook/dmo/digital-media/video-on-demand/video-streaming-svod/worldwide?currency=EUR` (06.03.2022)

Streijl, R., Winkler, S. & Hands, D. (2010, 08). Perceptual quality measurement-towards a more efficient process for validating objective models. *Signal Processing Magazine, IEEE*, *27*, 136 - 140. doi: 10.1109/MSP.2010.936776

Streijl, R., Winkler, S. & Hands, D. (2014). Mean opinion score (mos) revisited: methods and applications, limitations and alternatives. *Multimedia Systems*, *22*, 213-227.

Sun, S., Yu, T., Xu, J., Zhou, W. & Chen, Z. (2022). Graphiqa: Learning distortion graph representations for blind image quality assessment. *IEEE Transactions on Multimedia*, 1-1. doi: 10.1109/TMM.2022.3152942

VQEG. (2000). *Final report from the video quality experts group on the validation of objective models of video quality assessment.* (Video Quality Experts Group)

Wang, Z., Bovik, A. C. & Lu, L. (2002). Why is image quality assessment so difficult? In *2002 ieee international conference on acoustics, speech, and signal processing* (Bd. 4, S. IV-3313-IV-3316). doi: 10.1109/ICASSP.2002.5745362

Wang, Z., Lu, L. & Bovik, A. (2004, 02). Video quality assessment based on structural distortion measurement. *Signal Processing: Image Communication*, *19*, 121-132. doi: 10.1016/S0923-5965(03)00076-6

Wiebel, C. B., Aguilar, G. & Maertens, M. (2017, 04). Maximum likelihood difference scales represent perceptual magnitudes and predict appearance matches. *Journal of Vision*, *17* (4), 1-1. Zugriff auf `https://doi.org/10.1167/17.4.1` doi: 10.1167/17.4.1

Zerman, E., Hulusic, V., Valenzise, G., Mantiuk, R. & Dufaux, F. (2018, Januar). The relation between MOS and pairwise comparisons and the importance of cross-content comparisons. In *Human Vision and Electronic Imaging Conference, IS*

T International Symposium on Electronic Imaging (EI 2018). Burlingame, United States. Zugriff auf `https://hal.archives-ouvertes.fr/hal-01654133`

Zhai, G. & Min, X. (2020, 11). Perceptual image quality assessment: a survey. *Science China Information Sciences*, *63*. doi: 10.1007/s11432-019-2757-1

Zhang, W., Ma, K., Zhai, G. & Yang, X. (2021). Uncertainty-aware blind image quality assessment in the laboratory and wild. *IEEE Transactions on Image Processing*, *30*, 3474-3486. doi: 10.1109/TIP.2021.3061932

5 Anhang

5.1 Stimuli

Tabelle A1 gibt eine Übersicht der Bilddateien aus der LIVE2 Datenbank für alle Stimuli der eigenen Wahrnehmungsexperimente nach Motiv, Verzerrungsart, Verzerrungsgrad bzw. Verzerrungsrang. Die zugehörigen Bilder sind nach Motiv zusammengefasst in den Abbildungen auf den nächsten Seiten wiedergegeben: *Bikes* (Abb. A1), *ChurchAndCapitol* (Abb. A2), *Ocean* (Abb. A3), *WomanHat* (Abb. A4).

Tab. A1: LIVE2 Bilddateien zu jedem Stimulus der eigenen Wahrnehmungsexperimente nach Motiv, Verzerrungsart, Verzerrungsgrad bzw. Verzerrungsrang

	Verzerrungsrang	1	2	3	4	5	6
Motiv	**Verzerrungsart**	**Bilddatei**					
		Verzerrungsgrad					
Bikes	JPEG	img42.bmp	img203.bmp	img81.bmp	img29.bmp	img188.bmp	img183.bmp
		0,000	1,032	0,564	0,307	0,229	0,209
	White Noise	img152.bmp	img95.bmp	img68.bmp	img97.bmp	img29.bmp	img87.bmp
		0,000	0,035	0,090	0,258	0,320	1,996
	Gaussian Blur	img152.bmp	img60.bmp	img112.bmp	img129.bmp	img26.bmp	img40.bmp
		0,000	0,620	0,906	0,964	1,307	2,625
	Fast Fading	img152.bmp	img35.bmp	img34.bmp	img33.bmp	img32.bmp	img31.bmp
		0,000	25,100	22,700	20,300	18,900	16,500
ChurchAndCapitol	JPEG	img22.bmp	img132.bmp	img2.bmp	img21.bmp	img4.bmp	img195.bmp
		0,000	1,692	0,839	0,575	0,291	0,219
	White Noise	img172.bmp	img13.bmp	img18.bmp	img62.bmp	img93.bmp	img20.bmp
		0,000	0,020	0,035	0,141	0,234	1,996
	Gaussian Blur	img172.bmp	img142.bmp	img2.bmp	img91.bmp	img66.bmp	img104.bmp
		0,000	0,734	0,906	1,250	1,565	7,667
	Fast Fading	img172.bmp	img135.bmp	img134.bmp	img133.bmp	img132.bmp	img131.bmp
		0,000	26,100	22,700	21,300	18,900	15,500
Ocean	JPEG	img28.bmp	img143.bmp	img8.bmp	img98.bmp	img156.bmp	img164.bmp
		0,000	1,423	0,881	0,819	0,298	0,150
	White Noise	img147.bmp	img99.bmp	img118.bmp	img21.bmp	img120.bmp	img42.bmp
		0,000	0,012	0,035	0,109	0,180	1,996
	Gaussian Blur	img147.bmp	img92.bmp	img106.bmp	img119.bmp	img57.bmp	img117.bmp
		0,000	0,562	0,792	1,078	1,479	5,833
	Fast Fading	img147.bmp	img10.bmp	img9.bmp	img8.bmp	img7.bmp	img6.bmp
		0,000	25,100	23,700	20,300	18,900	16,500
WomanHat	JPEG	img26.bmp	img107.bmp	img89.bmp	img1.bmp	img154.bmp	img138.bmp
		0,000	0,601	0,453	0,326	0,162	0,158
	White Noise	img162.bmp	img11.bmp	img3.bmp	img28.bmp	img119.bmp	img60.bmp
		0,000	0,020	0,039	0,094	0,137	1,000
	Gaussian Blur	img162.bmp	img82.bmp	img61.bmp	img42.bmp	img36.bmp	img132.bmp
		0,000	0,677	1,021	1,479	1,937	3,542
	Fast Fading	img162.bmp	img85.bmp	img84.bmp	img83.bmp	img82.bmp	img81.bmp
		0,000	25,100	23,700	21,300	18,900	16,500

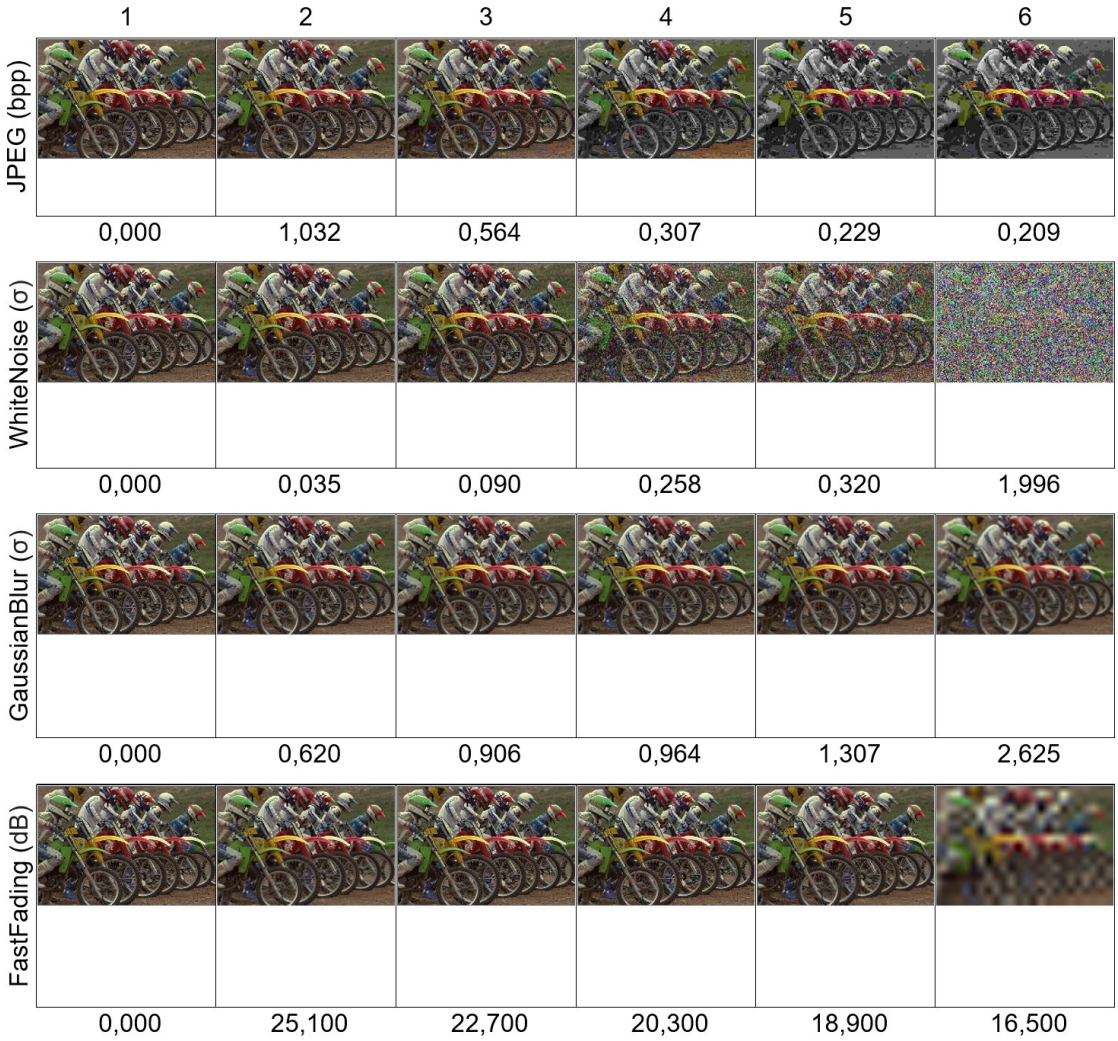


Abb. A1: Motiv: *Bikes*, Verzerrungsarten (Zeilen): *JPEG*, *WhiteNoise*, *GaussianBlur*, *FastFading*, Verzerrungsgrad (Spalten) sortiert nach Verzerrungsrang. Unter jedem Bild ist der Verzerrungsgrad in den Einheiten der jeweiligen Verzerrungsart angegeben: bpp, σ, dB. Die Titelzeile gibt den Verzerrungsrang von 1 bis 6 wieder. Die linke Spalte mit Verzerrungsgrad null bzw. Rang eins enthält das für alle Verzerrungsarten identische unverzerrte Referenzbild.

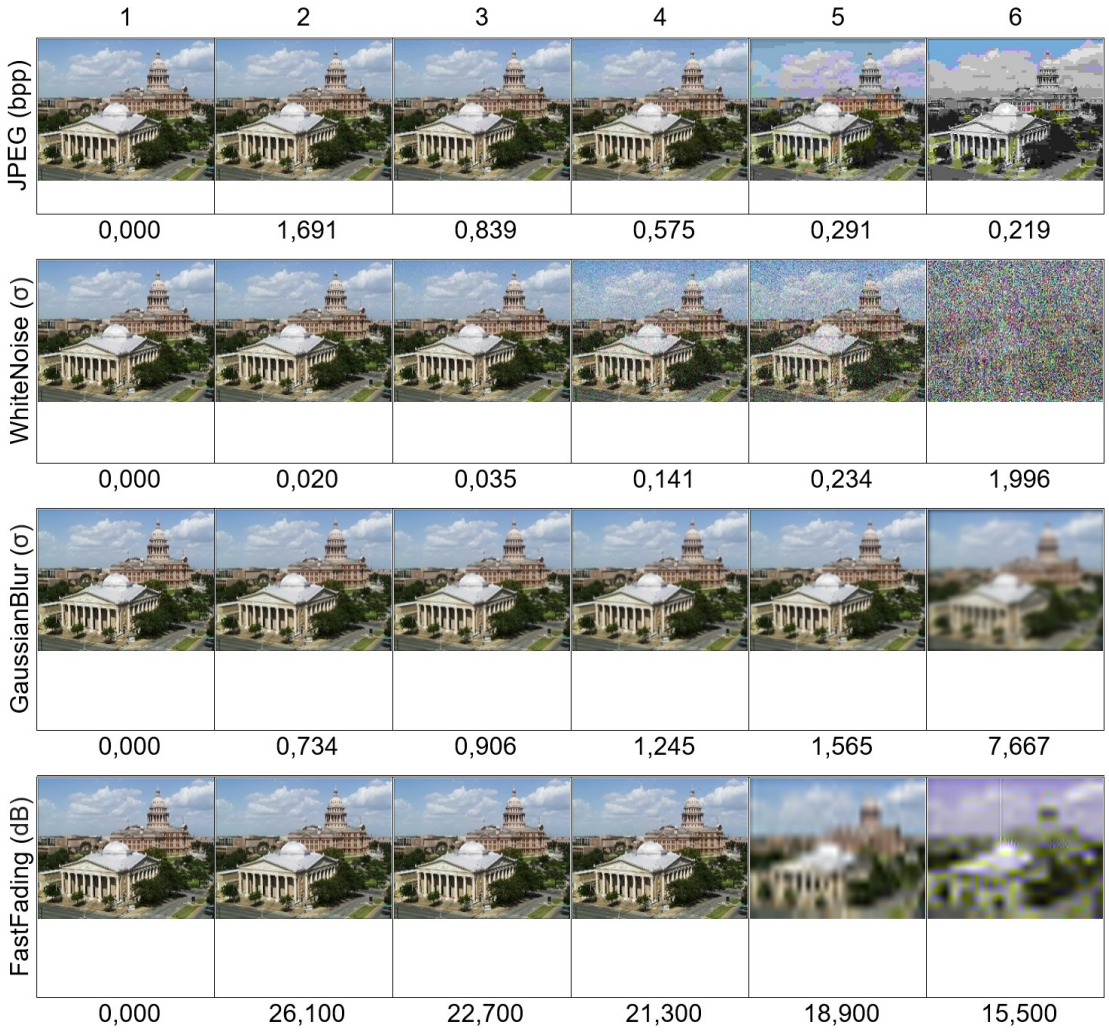


Abb. A2: Motiv: *ChurchAndCapitol*, Verzerrungsarten (Zeilen): *JPEG*, *WhiteNoise*, *GaussianBlur*, *FastFading*, Verzerrungsgrad (Spalten) sortiert nach Verzerrungsrang. Unter jedem Bild ist der Verzerrungsgrad in den Einheiten der jeweiligen Verzerrungsart angegeben: bpp, σ, dB. Die Titelzeile gibt den Verzerrungsrang von null bis sechs wieder. Die linke Spalte mit Verzerrungsgrad null bzw. Rang eins enthält das für alle Verzerrungsarten identische unverzerrte Referenzbild.

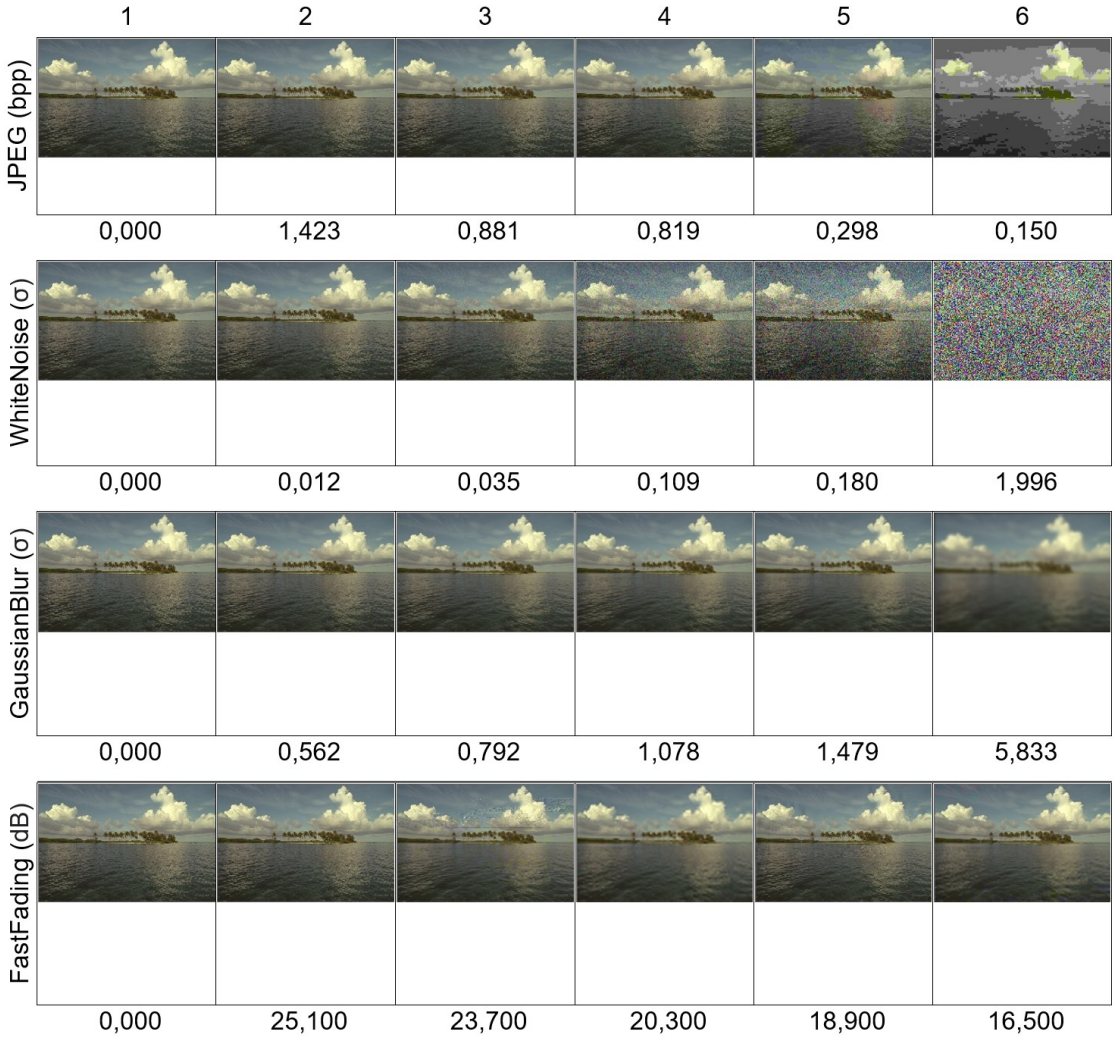


Abb. A3: Motiv: *Ocean*, Verzerrungsarten (Zeilen): *JPEG*, *WhiteNoise*, *GaussianBlur*, *FastFading*, Verzerrungsgrad (Spalten) sortiert nach Verzerrungsrang. Unter jedem Bild ist der Verzerrungsgrad in den Einheiten der jeweiligen Verzerrungsart angegeben: bpp, σ, dB. Die Titelzeile gibt den Verzerrungsrang von 1 bis 6 wieder. Die linke Spalte mit Verzerrungsgrad null bzw. Rang eins enthält das für alle Verzerrungsarten identische unverzerrte Referenzbild.

Abb. A4: Motiv: *WomanHat*, Verzerrungsarten (Zeilen): *JPEG*, *WhiteNoise*, *GaussianBlur*, *FastFading*, Verzerrungsgrad (Spalten) sortiert nach Verzerrungsrang. Unter jedem Bild ist der Verzerrungsgrad in den Einheiten der jeweiligen Verzerrungsart angegeben: bpp, σ, dB. Die Titelzeile gibt den Verzerrungsrang von 1 bis 6 wieder. Die linke Spalte mit Verzerrungsgrad null bzw. Rang eins enthält das für alle Verzerrungsarten identische unverzerrte Referenzbild.

5.2 Experimentelle Ergebnisse nach Verzerrungsgrad

Gegenüberstellung der Bildserien und Ergebnisse der eigenen Wahrnehmungsexperimente gruppiert nach Verzerrungsart und Motiv. Die Abszissen der Grafiken geben den Verzerrungsgrad der Bilder wieder.
In jeder der Darstellungen (Verzerrungsart, Motiv) befindet sich das nicht verzerrte Referenzbild in der linken oberen Kachel. Daneben folgen von links nach rechts die weiteren fünf Bilder der Serie mit aufsteigendem Verzerrungsgrad, so dass in der rechten mittleren Kachel das Bild mit der stärksten Verzerrung wiedergegeben ist. Die unteren Kacheln zeigen links die Ergebnisse der beiden Versuchspersonen für das MOS-Experiment (a), in der Mitte die Ergebnisse beider Versuchspersonen für das MLDS-Experiment (b). Diese beiden Darstellungen ermöglichen eine Kontrolle der Messergebnisse zwischen den beiden Versuchspersonen. Die Kachel (c) unten rechts fasst alle Ergebnisse zusammen: DMOS der LIVE2 Datenbank (rot), DMOS beider Versuchspersonen (gelb), MLDS beider Versuchspersonen (grün).
Zur besseren Vergleichbarkeit zwischen den verschiedenen Verzerrungsarten sind die Messergebnisse so dargestellt, dass der Wert des nicht verzerrten, besten Bildes immer durch den ersten Datenpunkt ganz links auf der Abszisse wiedergegeben wird. Für die Verzerrungsarten *JPEG* und *Fast Fading* ist es dafür erforderlich, dass die Abszisse von links nach rechts numerisch absteigend verläuft. Um die Datenpunkte der Referenzbilder grafisch korrekt eingeordnet darstellen zu können, wurde der Verzerrungsgrad des Referenzbildes für *JPEG* auf acht bpp und für *Fast Fading* auf 30 dB festgelegt.

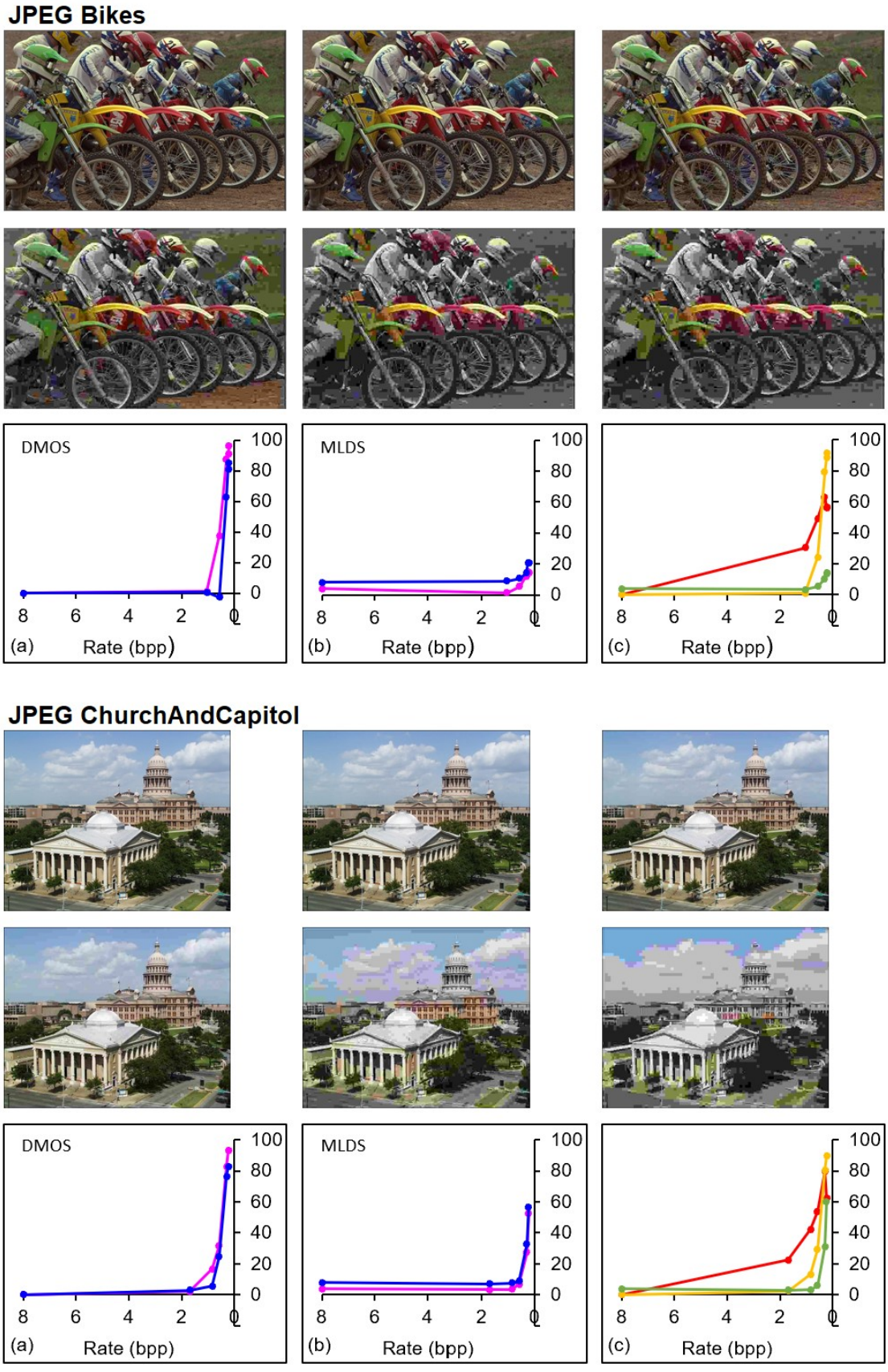


Abb. A5: JPEG: *Bikes*, *ChurchAndCapitol*

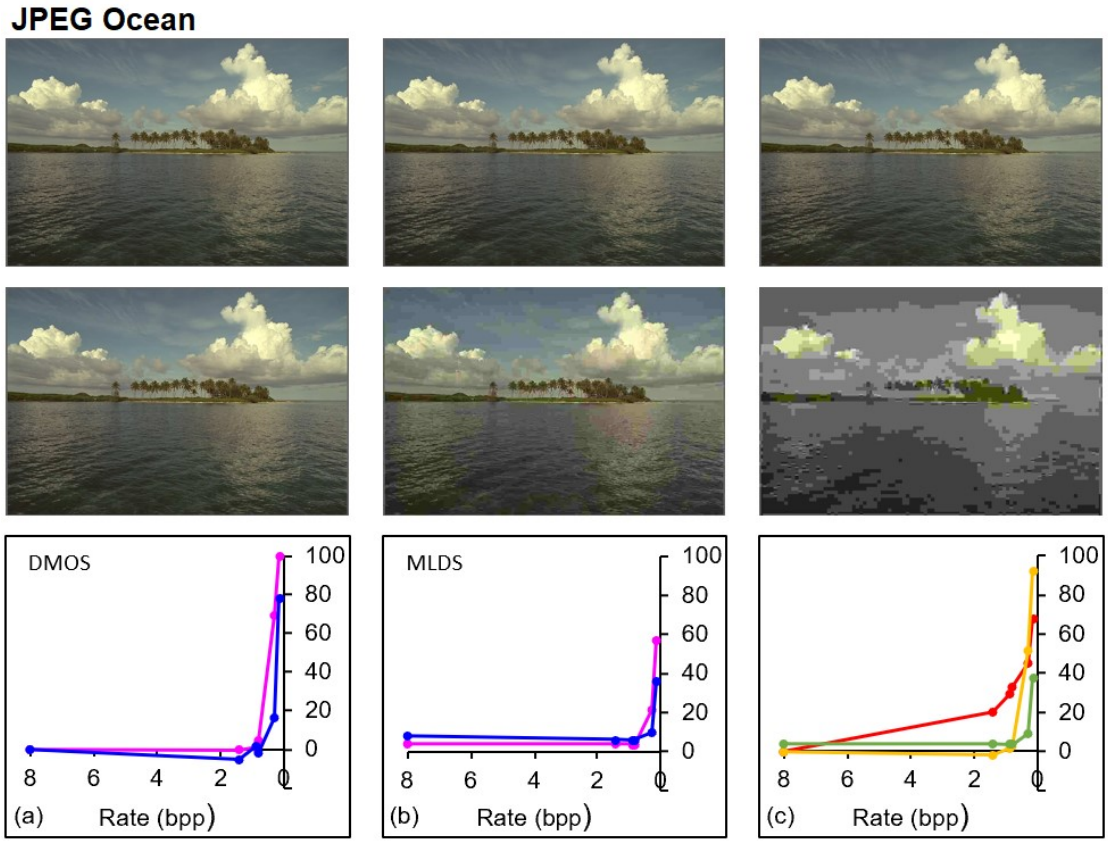


Abb. A6: JPEG: *Ocean*, *WomanHat*

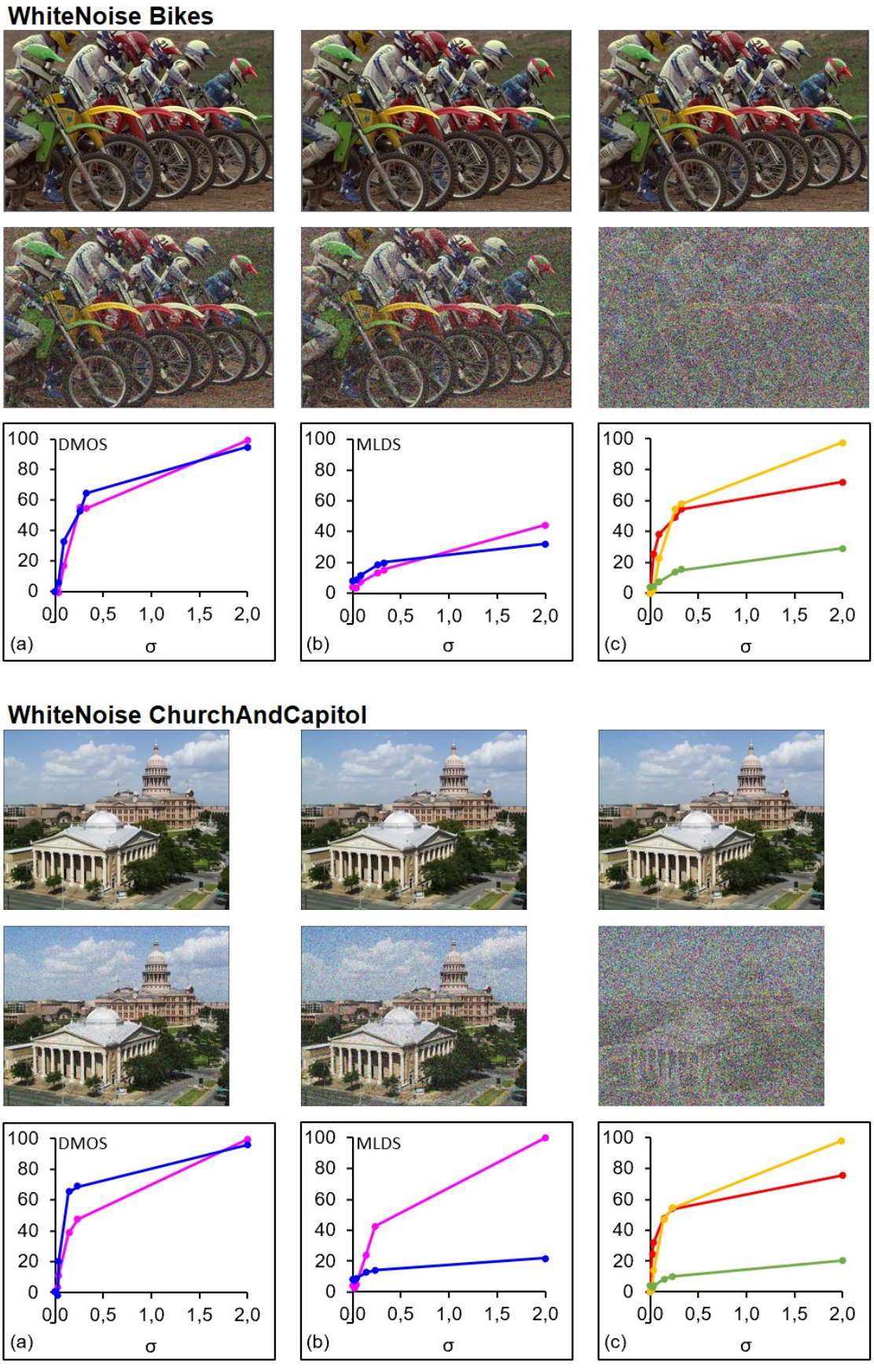


Abb. A7: WhiteNoise: *Bikes*, *ChurchAndCapitol*

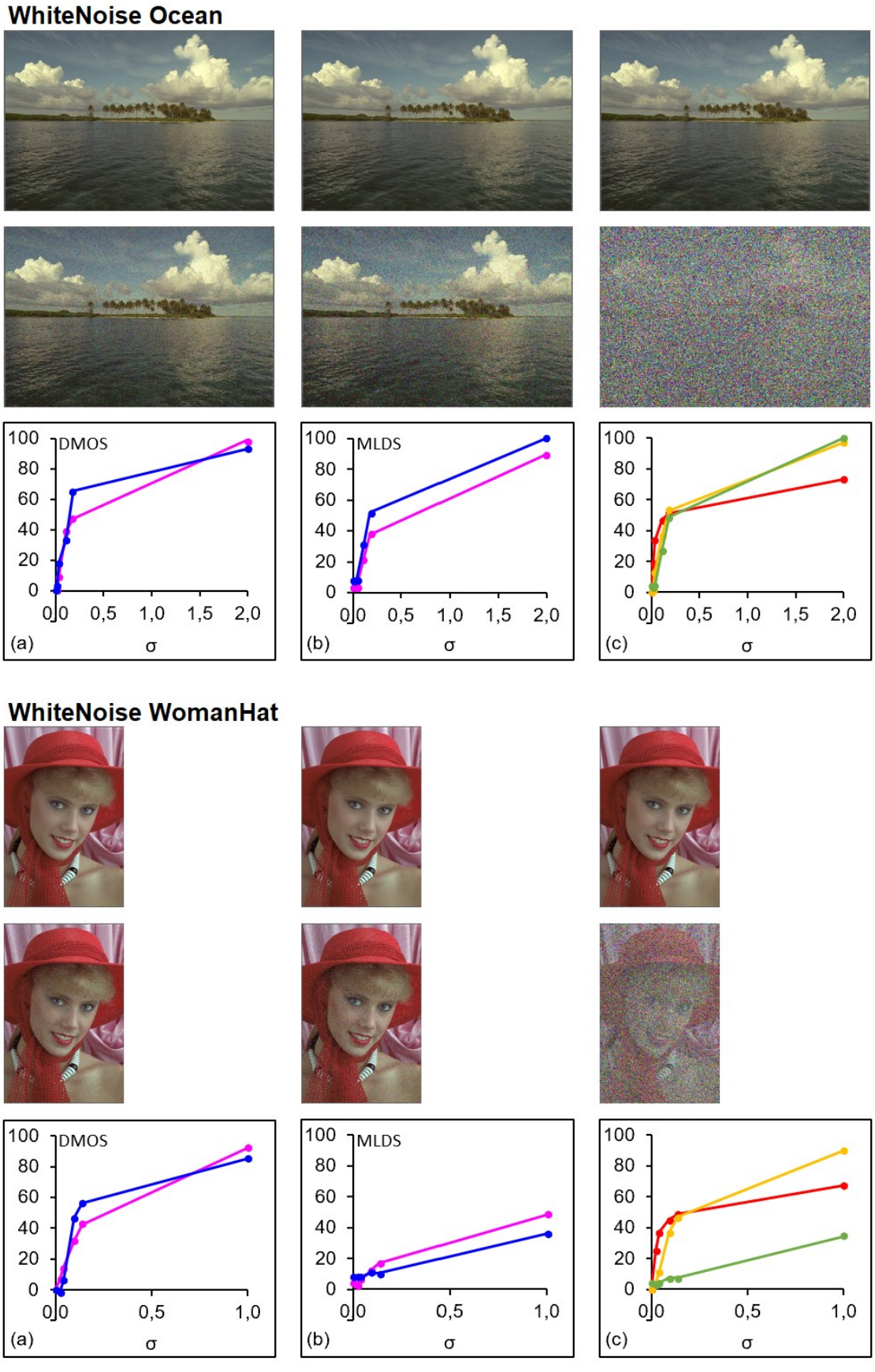


Abb. A8: WhiteNoise: *Ocean*, *WomanHat*

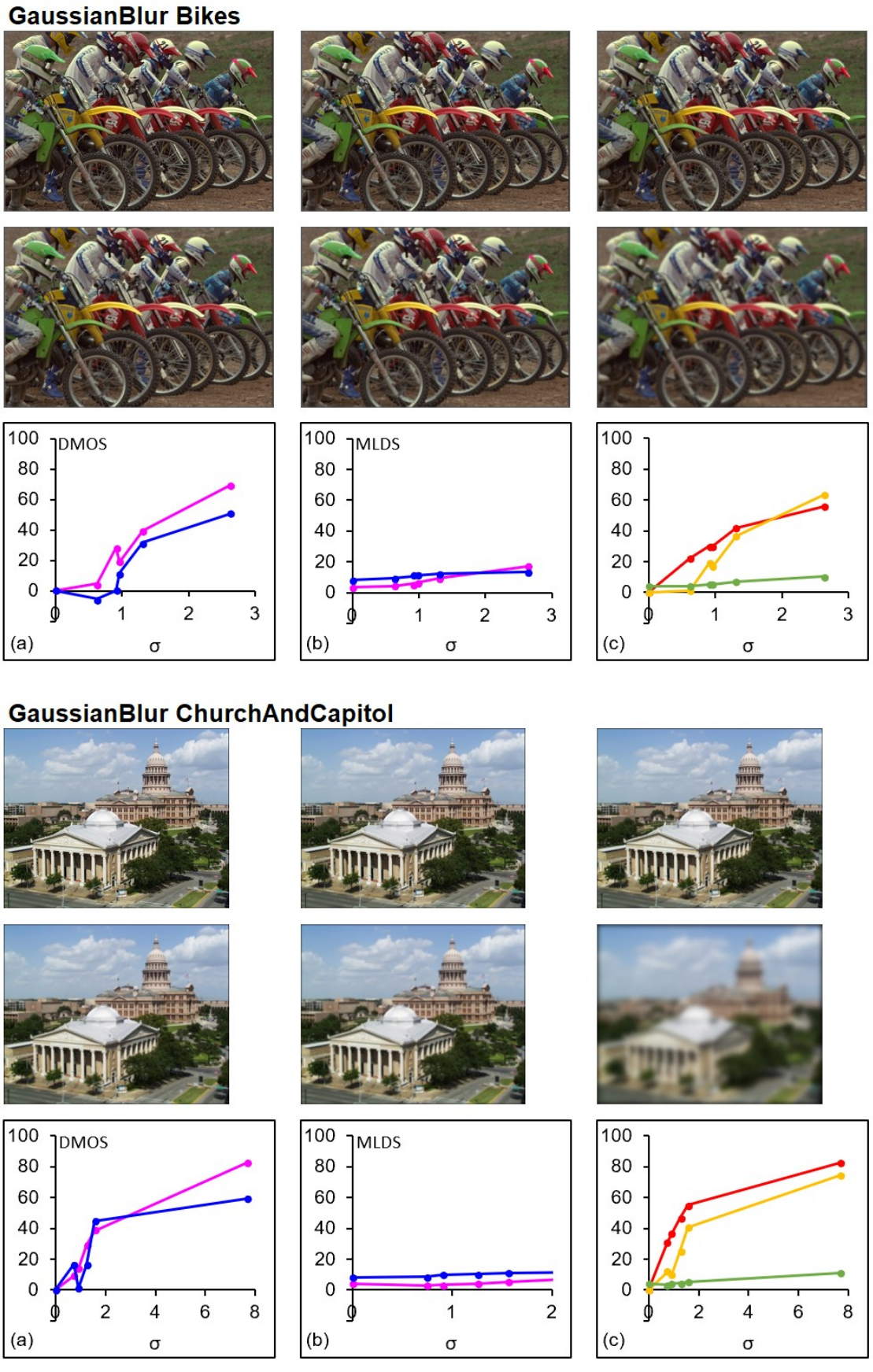


Abb. A9: GaussianBlur: *Bikes*, *ChurchAndCapitol*

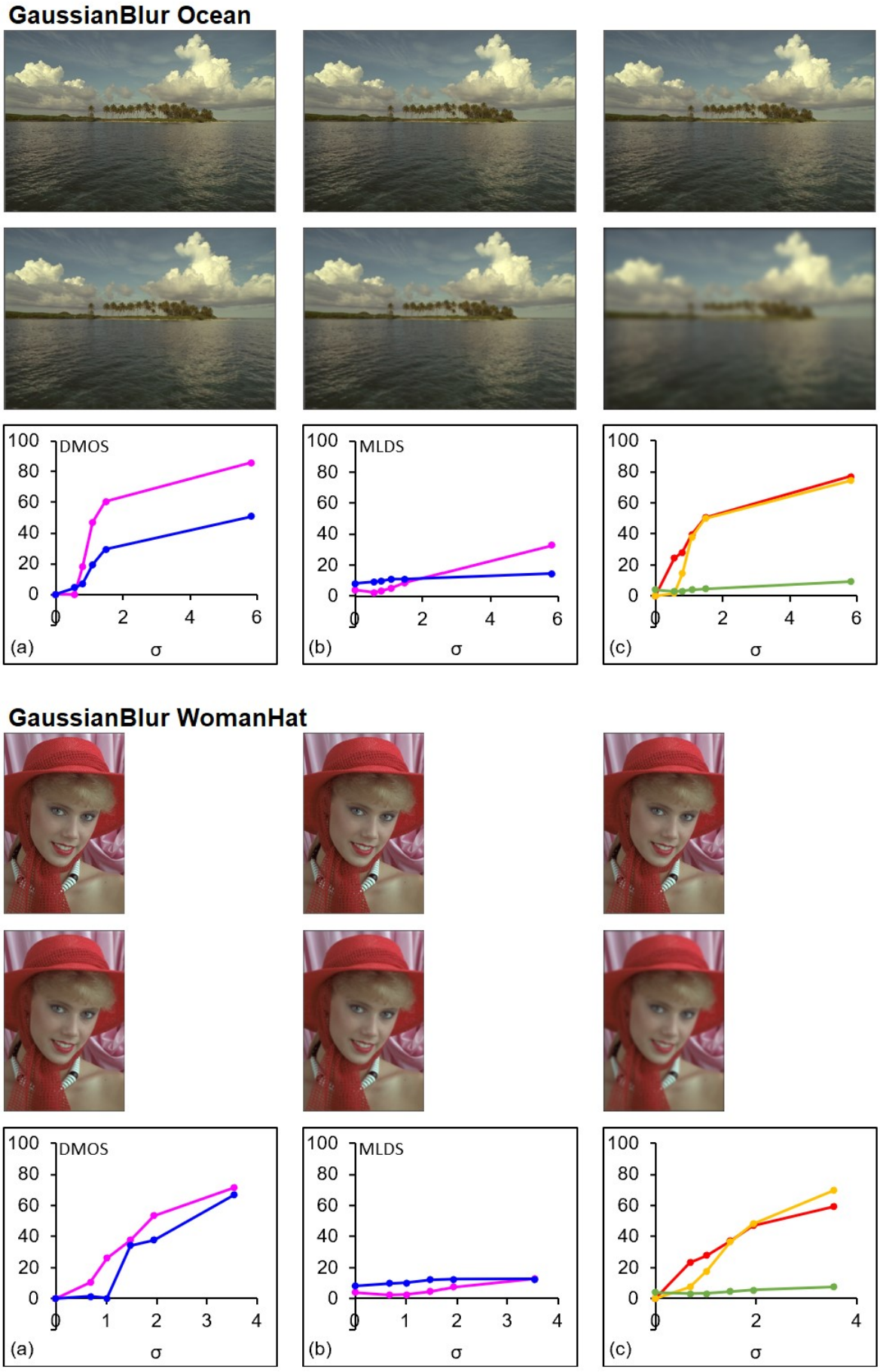


Abb. A10: GaussianBlur: *Ocean*, *WomanHat*

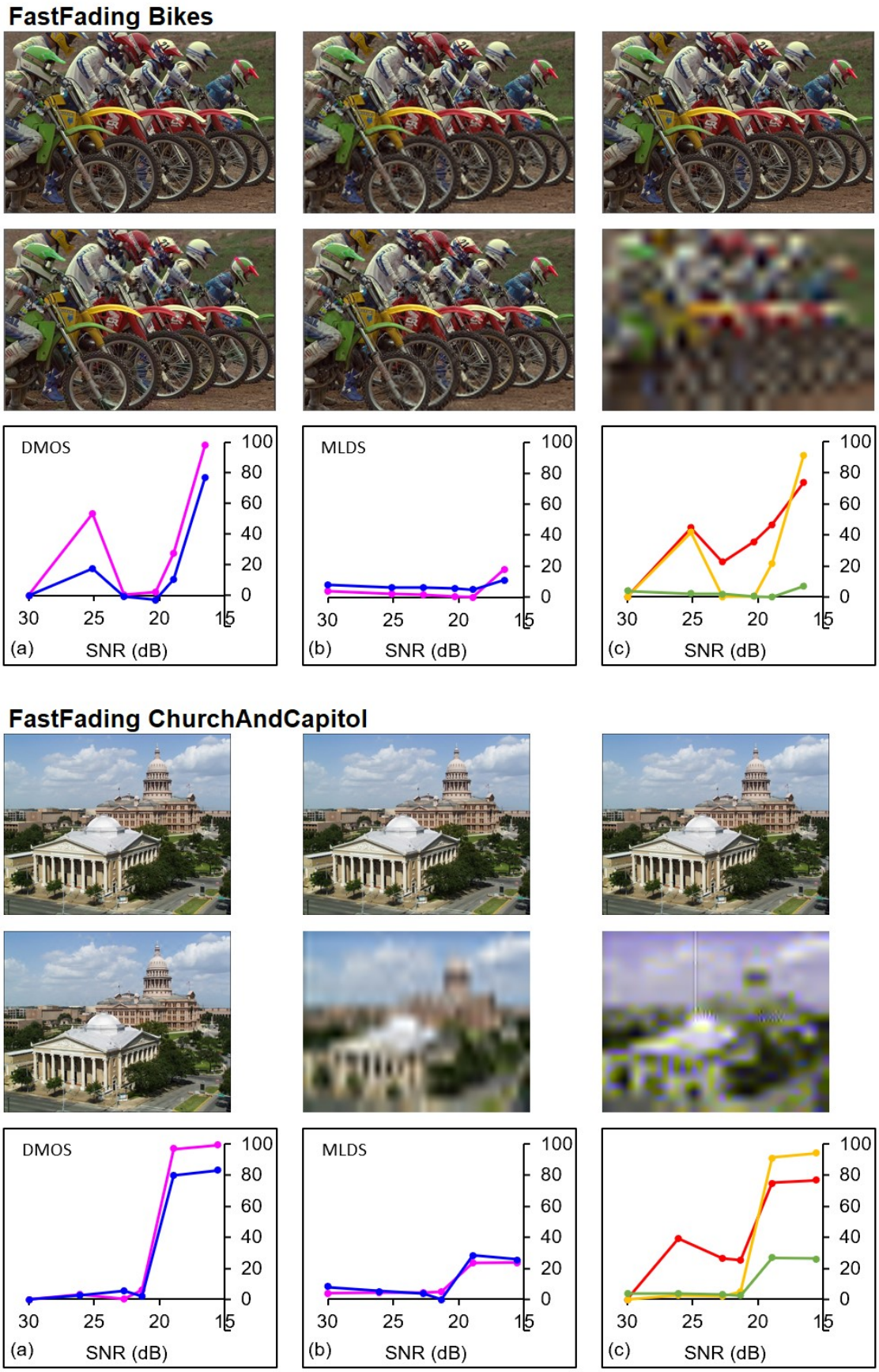


Abb. A11: FastFading: *Bikes*, *ChurchAndCapitol*

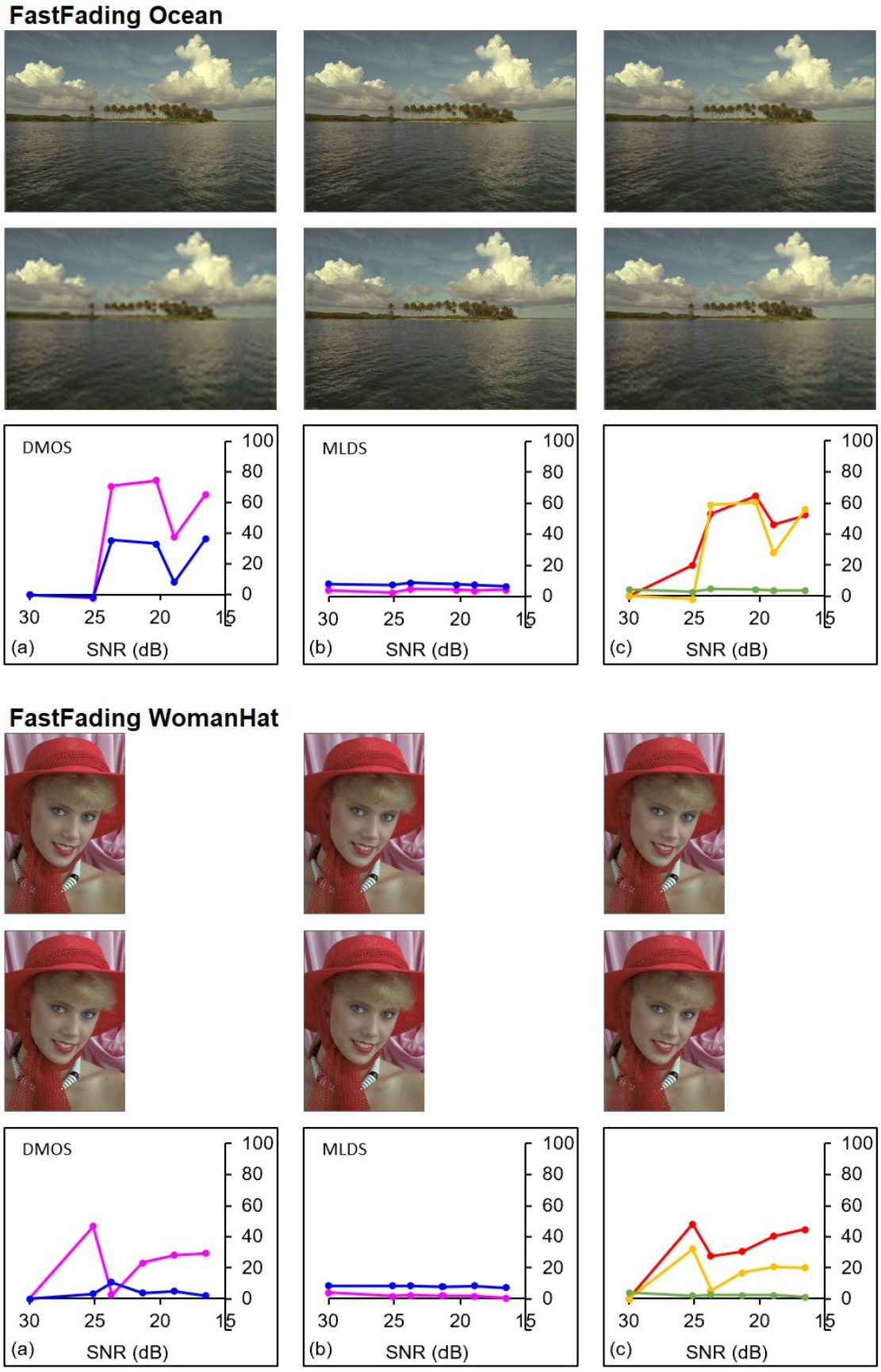


Abb. A12: FastFading: *Ocean*, *WomanHat*

5.3 Experimentelle Ergebnisse nach Verzerrungsrang

Gegenüberstellung der Bildserien und Ergebnisse der eigenen Wahrnehmungsexperimente gruppiert nach Verzerrungsart und Motiv. Die Abszissen der Grafiken geben den Verzerrungsrang der Bilder von eins bis sechs wieder.
In jeder der Darstellungen (Verzerrungsart, Motiv) befindet sich das nicht verzerrte Referenzbild in der linken oberen Kachel. Daneben folgen von links nach rechts die weiteren fünf Bilder der Serie mit aufsteigendem Verzerrungsrang, so dass in der rechten mittleren Kachel das Bild mit der stärksten Verzerrung wiedergegeben ist. Die unteren Kacheln zeigen links die Ergebnisse der beiden Versuchspersonen für das MOS-Experiment (a), in der Mitte die Ergebnisse beider Versuchspersonen für das MLDS-Experiment (b). Diese beiden Darstellungen ermöglichen eine Kontrolle der Messergebnisse zwischen den beiden Versuchspersonen. Die Kachel (c) unten rechts fasst alle Ergebnisse zusammen: DMOS der LIVE2 Datenbank (rot), DMOS beider Versuchspersonen (gelb), MLDS beider Versuchspersonen (grün).
Die Darstellung nach Verzerrungsrang erlaubt eine bessere Vergleichbarkeit zwischen den verschiedenen Verzerrungsarten, da der Wert des nicht verzerrten, besten Bildes mit Verzerrungsrang eins immer durch den ersten Datenpunkt ganz links auf der Abszisse wiedergegeben wird. Das am stärksten verzerrte Bild hat den Verzerrungsrang sechs unabhängig davon, ob der Verzerrungsgrad der jeweiligen Verzerrungsart numerisch auf- oder absteigend verläuft. Darüber hinaus ermöglicht die Darstellung nach Verzerrungsrang eine gute Unterscheidbarkeit der Datenpunkte auch bei Verzerrungsgraden, die sehr eng nebeneinander liegen (vgl. Kapitel 5.2).

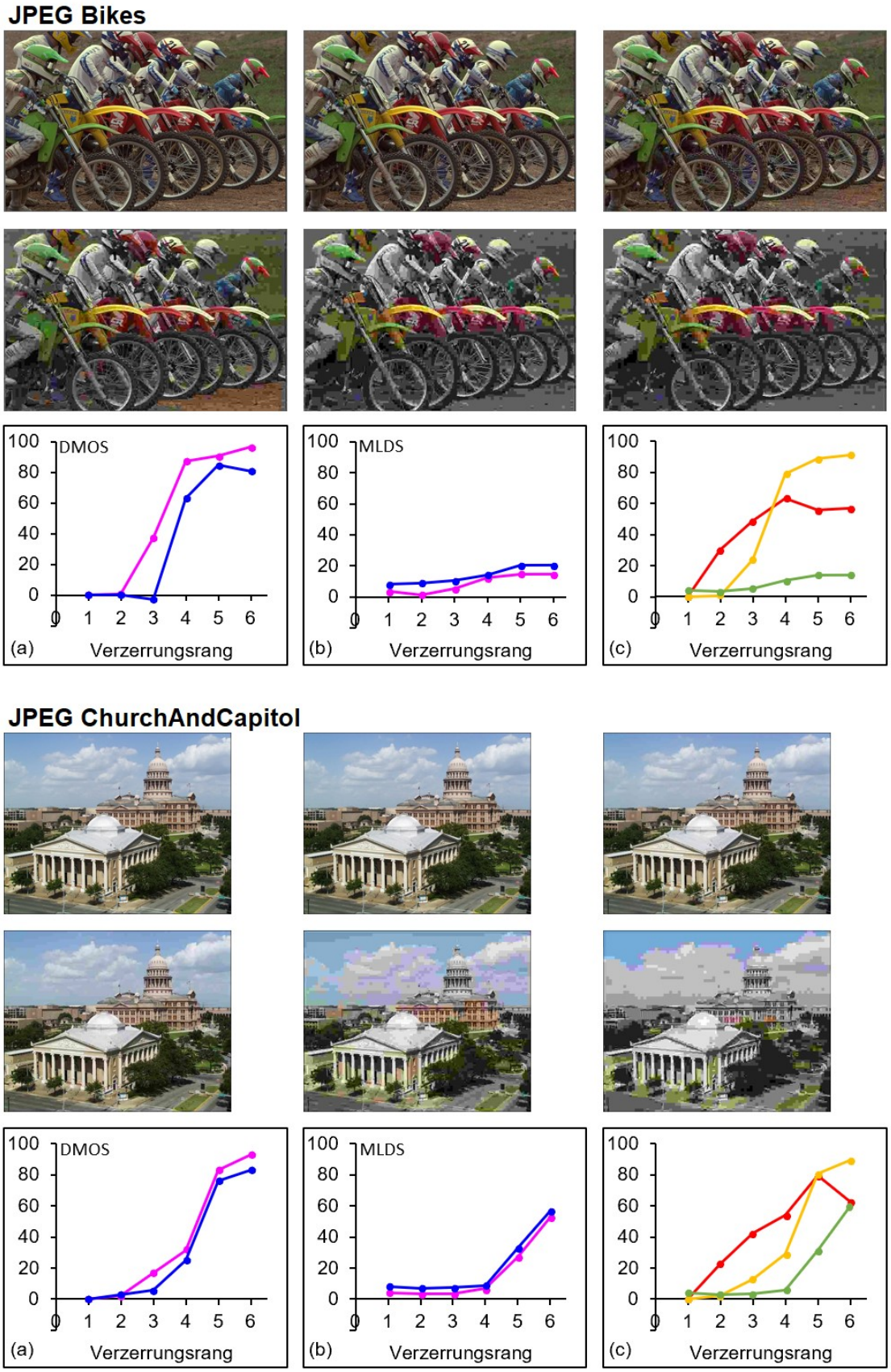


Abb. A13: JPEG: *Bikes*, *ChurchAndCapitol*

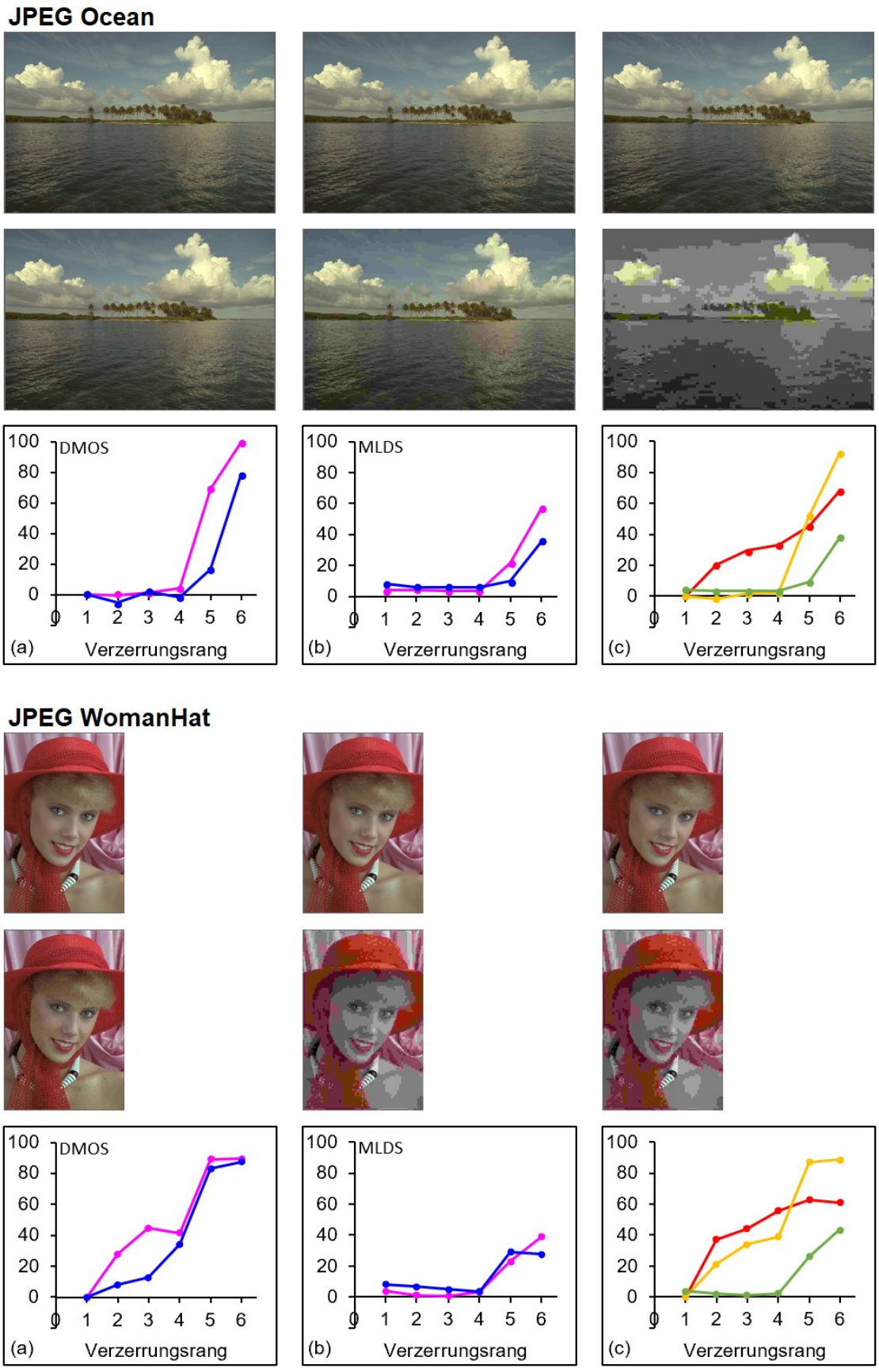


Abb. A14: JPEG: *Ocean*, *WomanHat*

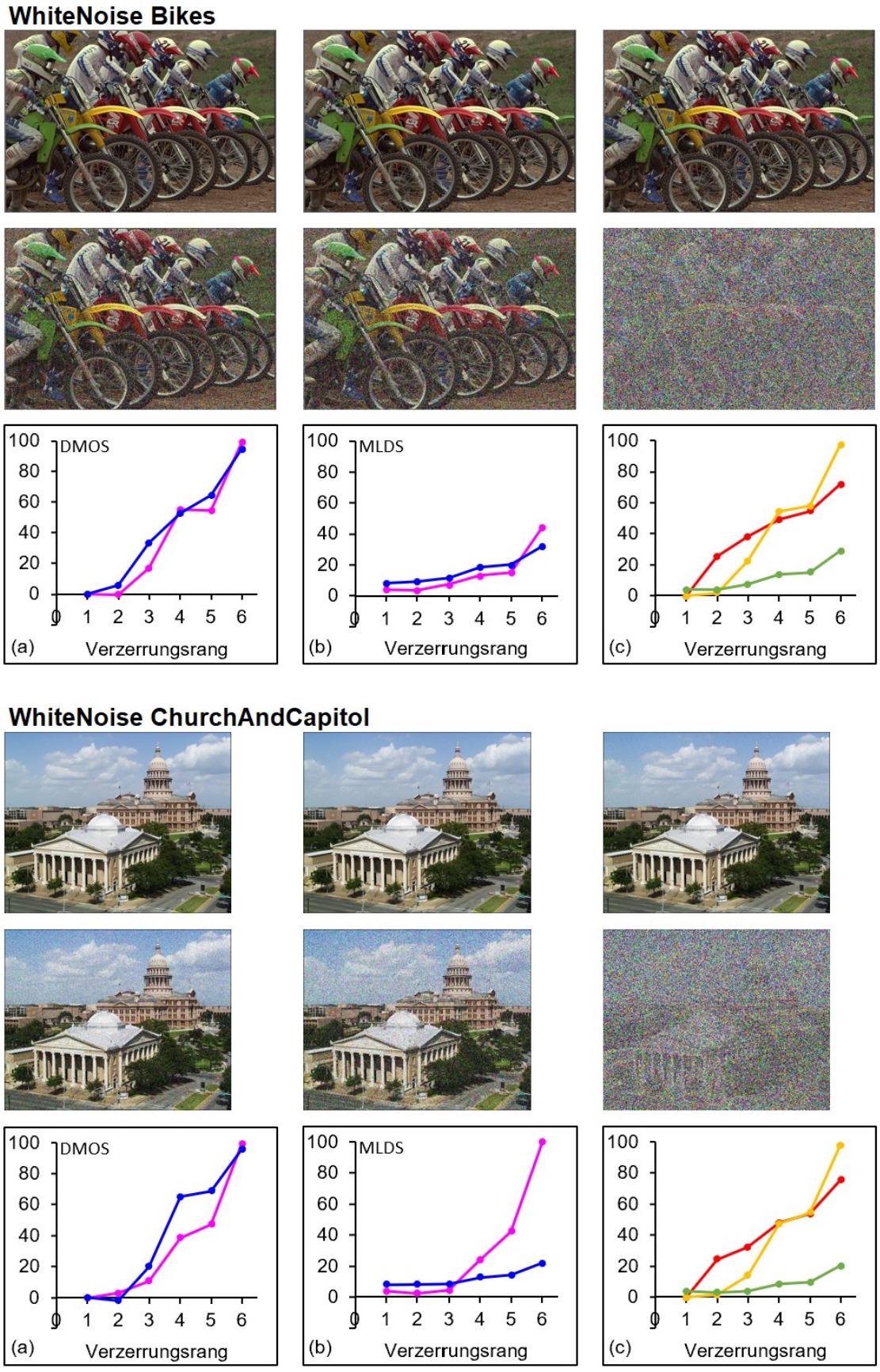


Abb. A15: WhiteNoise: *Bikes*, *ChurchAndCapitol*

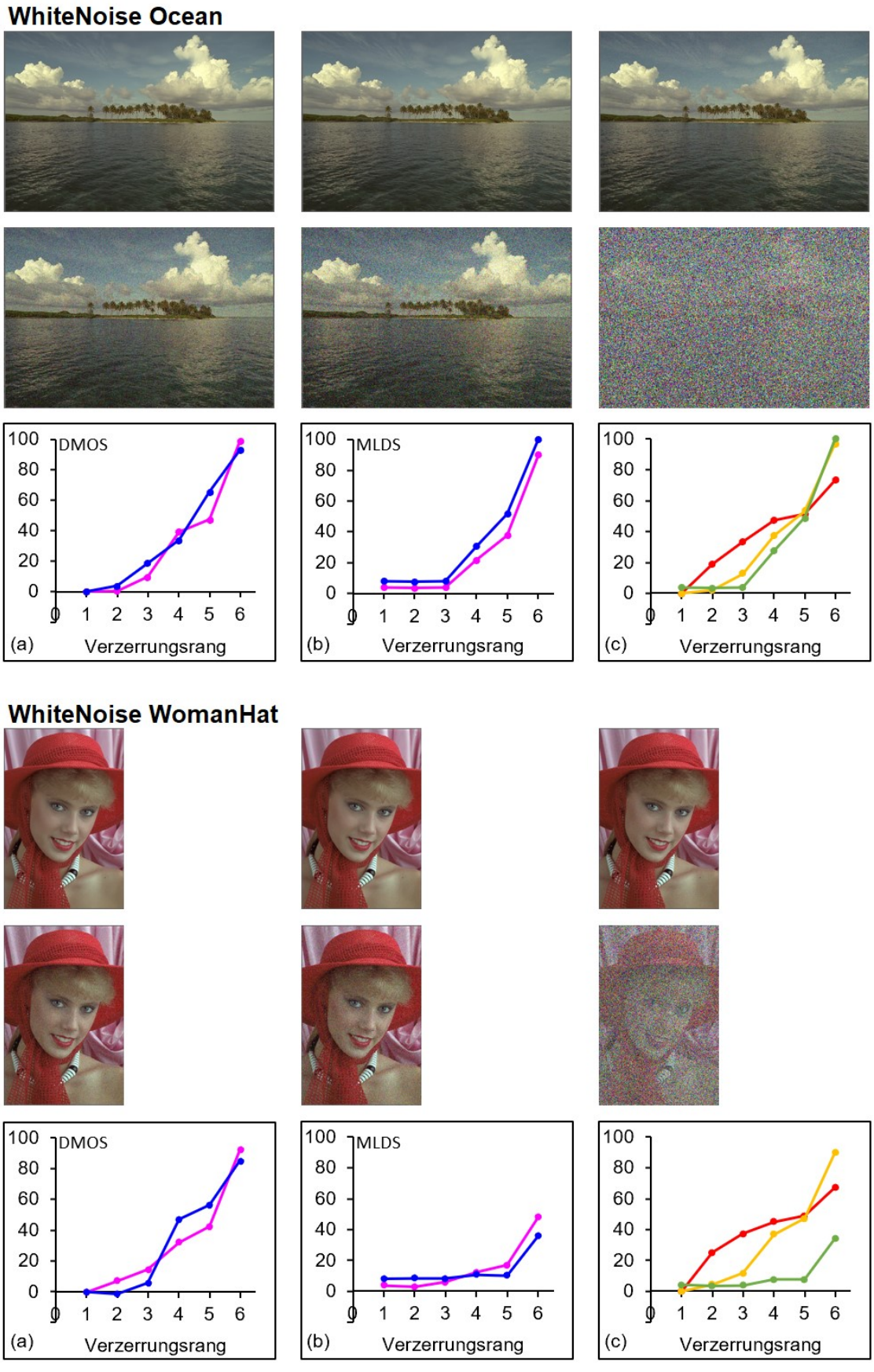


Abb. A16: WhiteNoise: *Ocean*, *WomanHat*

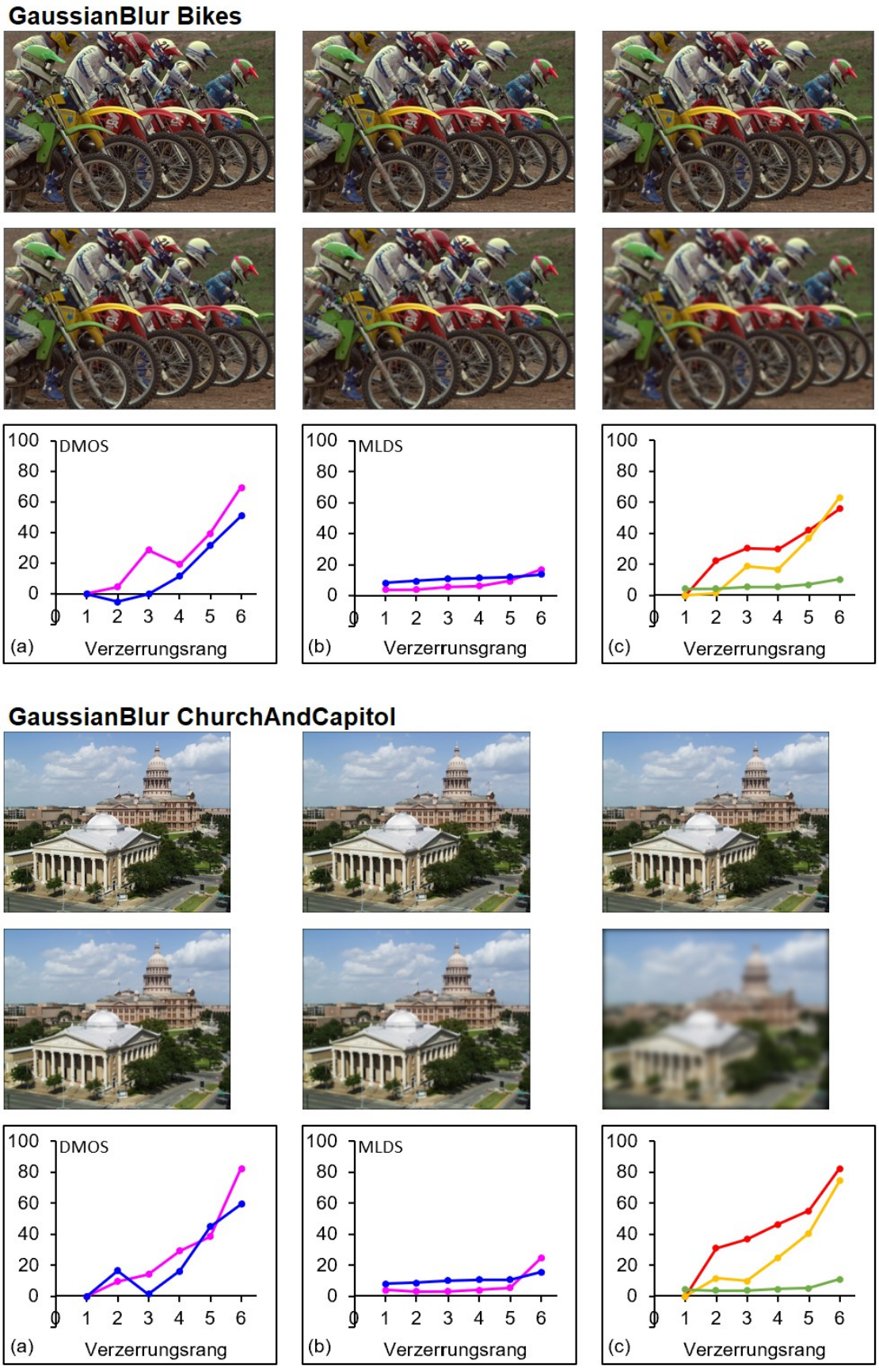


Abb. A17: GaussianBlur: *Bikes*, *ChurchAndCapitol*

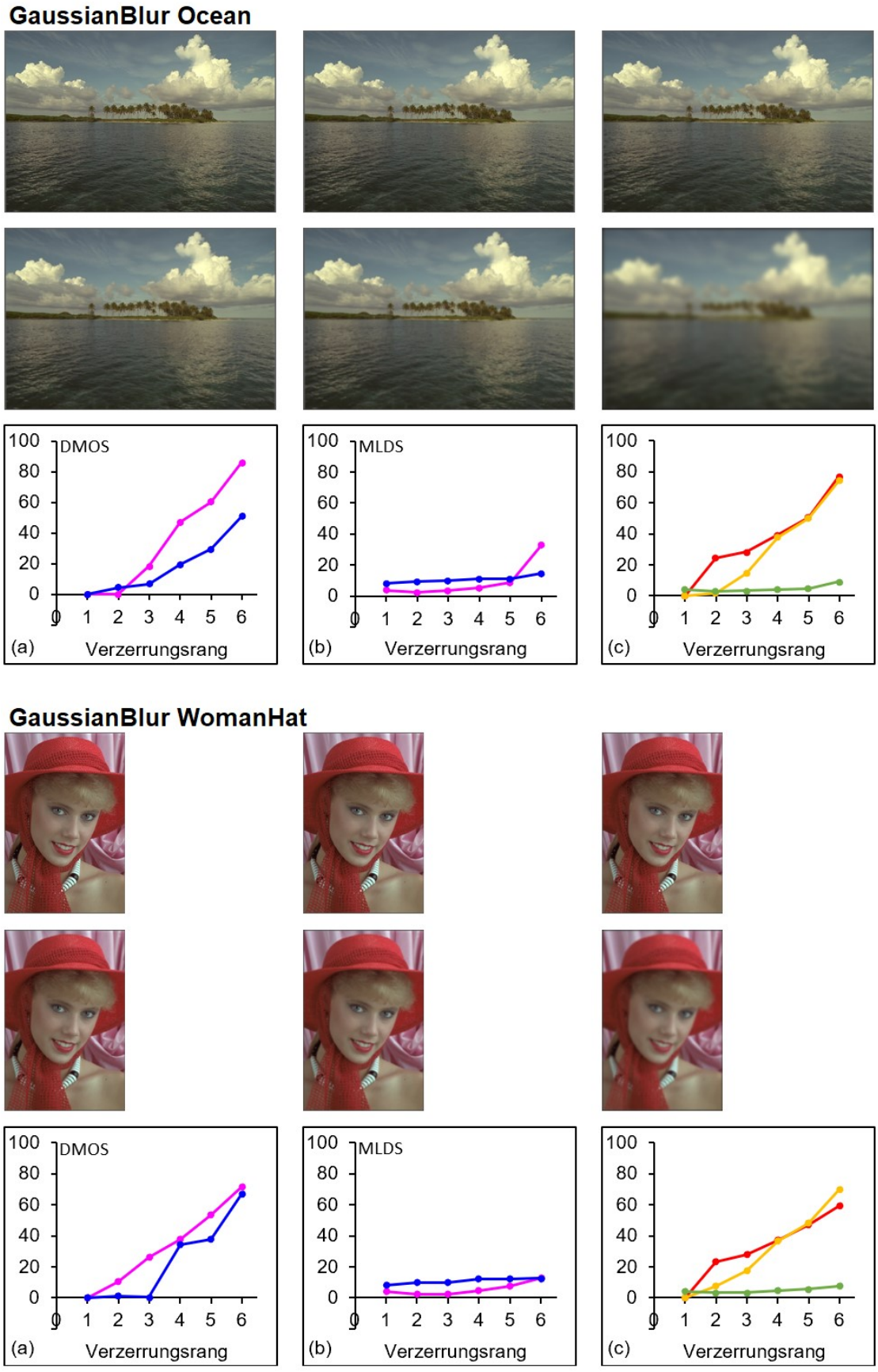
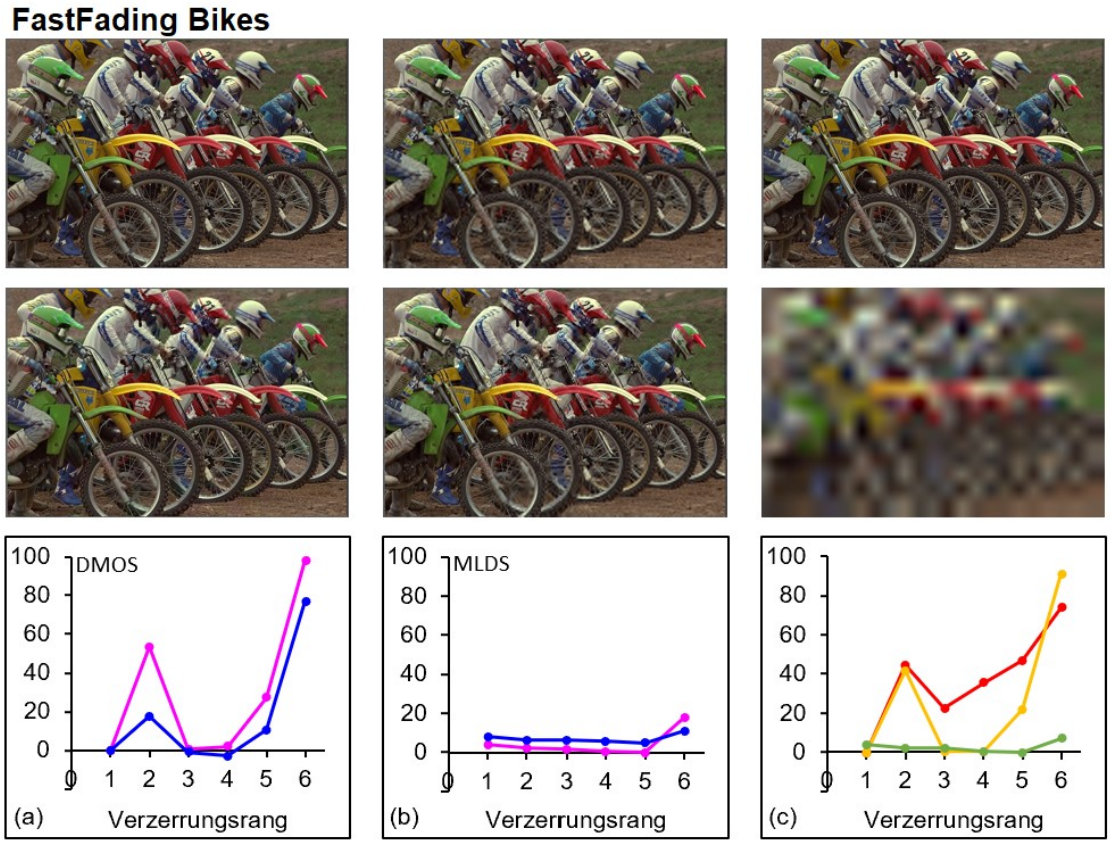


Abb. A18: GaussianBlur: *Ocean*, *WomanHat*



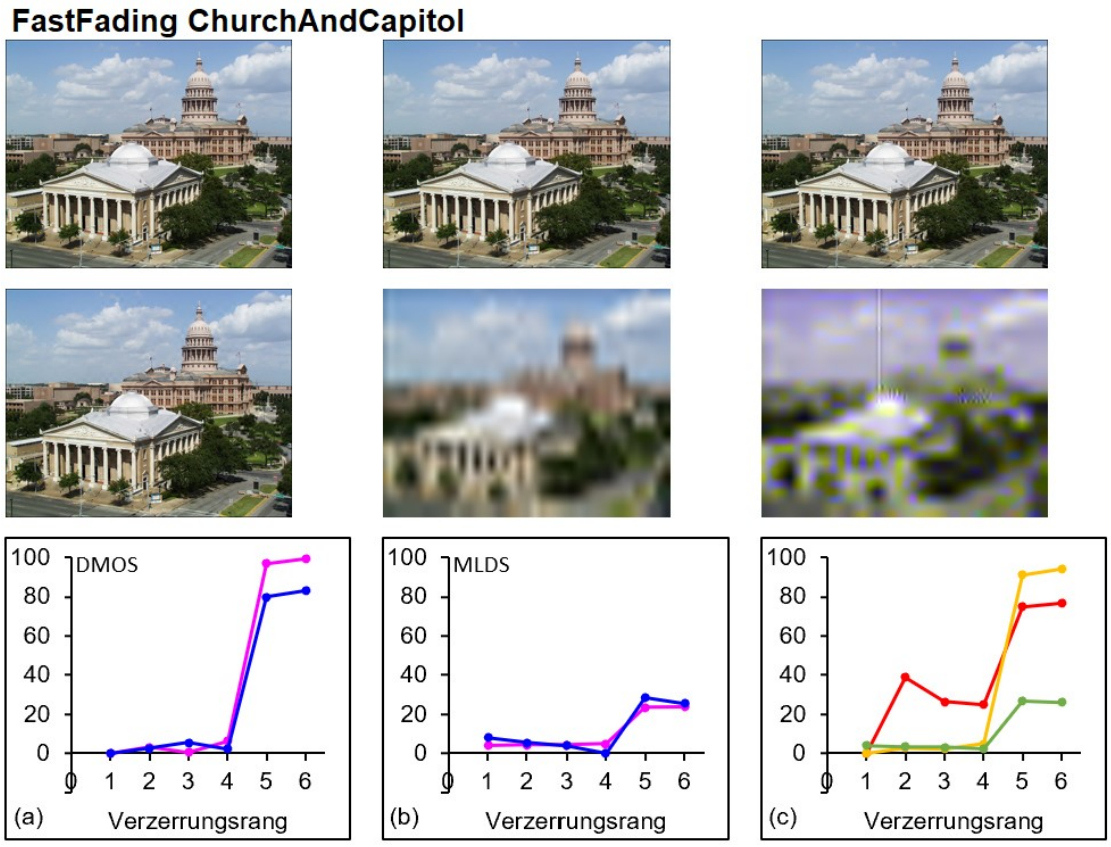


Abb. A19: FastFading: *Bikes*, *ChurchAndCapitol*

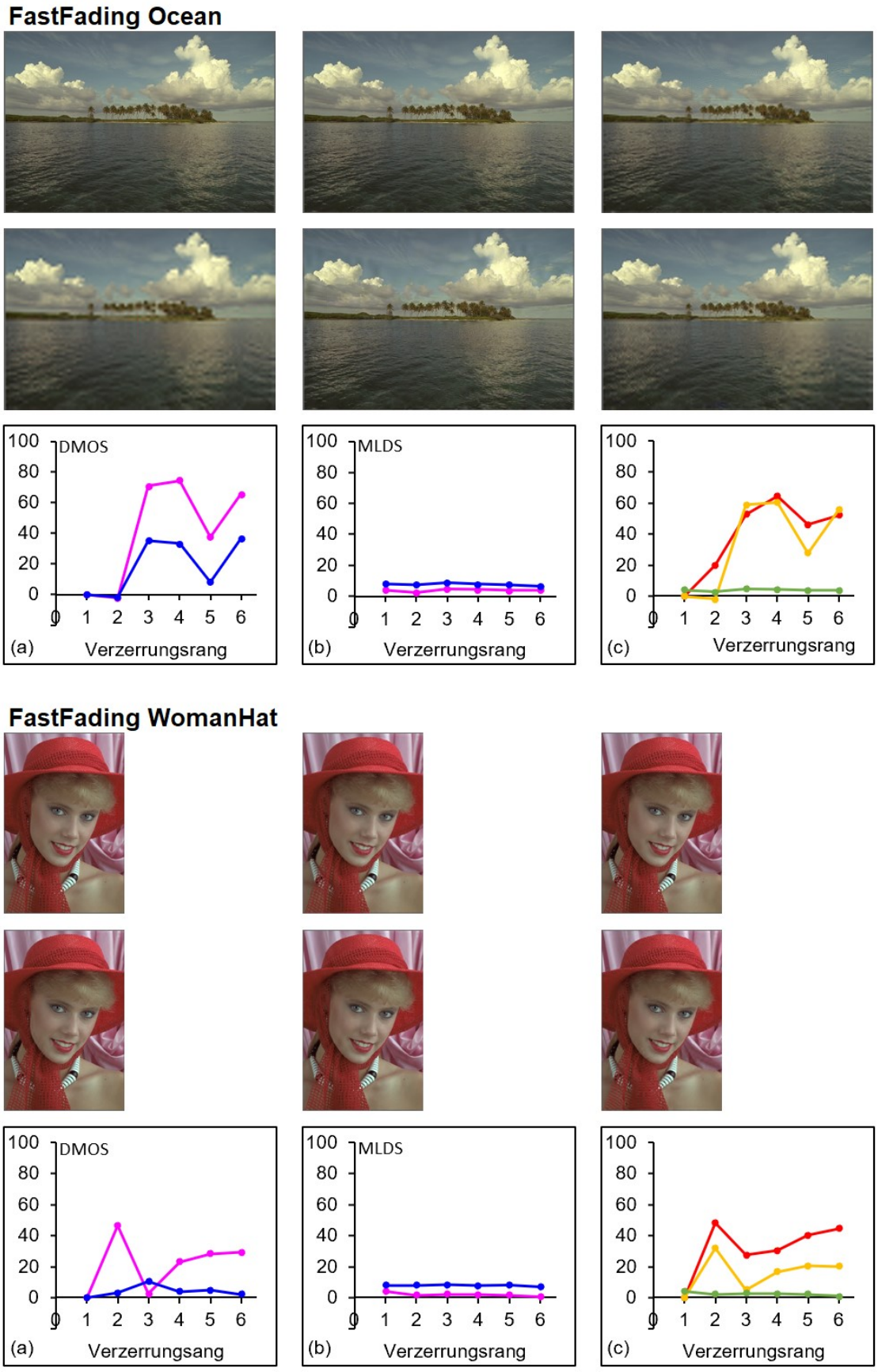


Abb. A20: FastFading: *Ocean*, *WomanHat*