Developing a home monitoring system for patients with chronic liver disease using a smartphone

Miranda Helen Sarah Nixon-Hill

A dissertation submitted in partial fulfillment
of the requirements for the degree of
Doctor of Philosophy
of
University College London.

Department of Medical Physics and Biomedical Engineering
University College London

December 2022

Authorship Declaration

I, Miranda Nixon-Hill, confirm that the work presented in this thesis is my own. Where information has been derived from other sources, I confirm that this has been indicated in the work.

Abstract

Liver disease is a growing problem in the UK, and one of the major causes of working-age premature death. Patients with advanced liver disease are typically admitted to hospital on multiple occasions, where they are stabilised before discharge. At home, there is little or no monitoring of their condition available, making it difficult to time additional treatment. Here, a system for non-invasive assessment of serum bilirubin level is proposed, based on imaging the white of the eye (sclera) using a smartphone. Elevated bilirubin level manifests as jaundice, and is a key indicator of overall liver function. Smartphone imaging makes the system low cost, portable and non-contact.

An ambient subtraction technique based on subtracting data from flash/ no-flash image pairs is leveraged to account for variations in ambient light. The subtracted signal to noise ratio (SSNR) metric has been developed to ensure good image quality. Values falling below the experimentally-determined threshold of 3.4 trigger a warning to re-capture. To produce device-independent results, mapping approaches based on image metadata and colour chart images were compared. It was found that introducing a one-time calibration step of imaging a colour chart for each device leads to the best compatibility of results from different phones.

In a clinical study at the Royal Free Hospital, London, over 100 sets of patient scleral images were captured with two different smartphones and paired clinical information was recorded. A filtering algorithm was developed to tackle the high density of blood vessels and specular reflection observed in the images, yielding a 94% success rate. Strong cross-sectional and longitudinal correlations of scleral yellowness and serum bilirubin level were found of 0.89 and 0.72 respectively (both $p<0.001$). When the proposed processing was applied, results from the two phones were demonstrated to be compatible. These results demonstrate the strong potential for the system as a monitoring tool.

Impact Statement

The primary contribution of this thesis is the development of a non-invasive approach for assessing jaundice in adults with advanced liver disease. Patients with advanced liver disease typically end up in hospital needing urgent care on repeat occasions. They are treated and then discharged home, where they have very little monitoring available. It can be hard to know when to seek additional medical assistance, and thus they may end up back in hospital requiring in-patient treatment. This cycle of poorly timed repeat admissions is negative for their health and well-being as well as very expensive for the NHS.

This work focusses on assessing bilirubin level, which manifests as jaundice, in a cost-effective and non-invasive way. The bilirubin level is a key indicator of overall liver health and the ability to monitor it non-invasively rather than via a blood test would enable assessment outside of the hospital. The proposed approach to assess bilirubin level could be incorporated into an overall home monitoring system based around multiple indicators of patient health, such as heart rate variability and weight. This overall system could help patients receive additional care in a timely manner.

The proposed bilirubin assessment approach is based on smartphone imaging, and has a very simple image capture step. Images can be captured using the front or rear camera, making the method compatible with self-imaging or with use by a third party. The low cost of smartphones and ease of use may help the approach to be adopted.

Along with the future impact on patients with liver disease and their medical teams upon deployment of the system, this work makes several academic contributions. The use of an ambient subtraction technique to account for variation in ambient light was further tested and demonstrated. In particular, a simple metric to ensure that captured image data was appropriate for use with this technique was developed and tested. The ability to combine the subtraction technique with a one-time calibration to produce results which no longer depend on the capture conditions or device was demonstrated. This ability is key to avoid unnecessary recapture of data and enlargement of datasets through the introduction of new devices. The general two-step processing proposed may be of use to researchers aiming to quantify colour using smartphones in other applications within and beyond medicine.

Another contribution of this work is the clinical dataset produced. Around 100

image sets were captured using two different phones in parallel, including longitudinal captures, and detailed clinical information was also recorded. The use of several phones, longitudinal information and the clinical detail sets this dataset apart from previous work. In the future, it could be expanded further to improve the proposed system or re-analysed to yield alternative approaches for monitoring bilirubin level via smartphone images.

Publications resulting

Journal articles

- Konstantin Kazankov, Miranda Nixon-Hill (joint first author), Rahul Kumar, Ahmed Amin, Eman Alabsawy, Anmol Chikhlia, Terence S. Leung, and Rajeshwar P. Mookerjee. "A novel Smartphone scleral-image based tool for assessing jaundice in decompensated cirrhosis patients" *Journal of Gastroenterology and Hepatology* Under review (2022)
- Miranda Nixon-Hill, Rajeshwar P. Mookerjee and Terence S. Leung. "Assessment of bilirubin levels in patients with cirrhosis via forehead, sclera and lower eyelid smartphone images" *PLOS Digital Health* Under review (2022)
- Thomas Wemyss, Miranda Nixon-Hill, Felix Outlaw, Anita Karsa, Judith Meek, Christabel Enweronu-Laryea, and Terence S. Leung. "Feasibility of smartphone colorimetry of the face as an anaemia screening tool for infants and young children in Ghana" *PLOS ONE* Under review (2022)
- Christabel Enweronu-Laryea, Terence S. Leung, Felix Outlaw, Nana Okai Brako, Genevieve Insaidoo, Nana Ayegua Hagan-Seneadza, Mary Ani-Amponsah, Miranda Nixon-Hill, and Judith Meek. "Validating a Sclera-Based Smartphone Application for Screening Jaundiced Newborns in Ghana" *Paediatrics* 150(1):e2021053600 (2022)
- Miranda Nixon, Felix Outlaw, and Terence S. Leung. "Accurate device-independent colorimetric measurements using smartphones" *PLOS ONE* 15(3):e0230561 (2020)
- Felix Outlaw, Miranda Nixon, Oluwatobiloba Odeyemi, Lindsay W. MacDonald, Judith Meek, and Terence S. Leung. "Smartphone screening for neonatal jaundice via ambient-subtracted sclera chromaticity" *PLOS ONE* 15(3):e0216970 (2020)

Conference proceedings

- Miranda Nixon-Hill, Felix Outlaw, Lindsay W. MacDonald, Rajeshwar Mookerjee, and Terence S. Leung. "Minimising ambient illumination via ambient subtraction : smartphone assessment of jaundice in liver patients via sclera images" In *28th Color and Imaging Conference Final Program and Proceedings*, pages 307-312 (2020)

- Miranda Nixon, Felix Outlaw, Lindsay W. MacDonald, and Terence S. Leung."The importance of a device specific calibration for smartphone colorimetry" In *27th Color and Imaging Conference Final Program and Proceedings*, pages 49-54 (2019)
- Felix Outlaw, Miranda Nixon, Nana Okai Brako, Lindsay W. MacDonald, Judith Meek, Christabel Enweronu-Laryea, and Terence S. Leung."Smartphone colorimetry using ambient subtraction: application to neonatal jaundice screening in Ghana" In *Adjunct Proceedings of the 2019 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, pages 172–175 (2019)

Conferences attended

Presentations

- 28th Color and Imaging Conference, Online, November 2020

Posters

- London Imaging Meeting, Online, September 2020
- The International Liver Congress, Online, August 2020 [Abstract included in the "Best of ILC" which highlights the most noteworthy contributions to the scientific programme]
- 27th Color and Imaging Conference, Paris, October 2019 [Won the Cactus Award for the best interactive paper; awarded a Colour Group (GB) £500 DW Wright award for travel to a colour-related conference]
- BioMedEng Conference, London, September 2019

Academic meetings attended

- Colour Group (GB) Student Awards Presentations, November 2019
- UCL Institute for Healthcare Engineering Global Healthcare Engineering Symposium, London, July 2019
- Colour Group (GB) Colour in the Clinic, December 2018
- UCL Institute for Healthcare Engineering Autumn Symposium, London, October 2018 [Won a £200 prize for the best talk with a global healthcare engineering theme]

Contributions

Important contributions to the work presented in this thesis were made by Ahmed Amin (AA), Ana Barreira (AB), Alister Lam (AL), Alice Skelton (AS), Eman Alabsawy (EA), Felix Outlaw (FO), Fiona Young (FY), Judith Meek (JM), Josephine Windsor-Lewis (JWL), Konstantin Kazankov (KK), Lindsay MacDonald (LM), Miranda Nixon-Hill (MNH), Mohammed Sheikh (MS), Monica Mesquita (MM), Rahul Kumar (RK), Rajeshwar Mookerjee (RM), and Terence Leung (TL). A summary of areas they contributed to are given below.

Ambient subtraction technique
FO and TL came up with the concept with input from LM. MNH proposed the colour chart demonstration; MNH and FO collected the data; MNH performed the analysis. MNH, TL and LM planned the sclera demonstration; MNH collected data and performed analysis.

Subtracted signal to noise metric
MNH, FO and TL identified the need. MNH and FO designed the metric. MNH and FO collected data for the experimental testing; MNH wrote the software and performed the analysis to select a threshold.

One-time calibration step
MNH and FO proposed the concept with LM. MNH collected data, wrote the software and performed analysis to compare methods and select appropriate mappings.

Camera spectral sensitivity measurement
AS provided equipment and initial software, MNH collected data and performed all analysis.

Custom colour chart
LM, MNH and FO came up with the concept. MNH designed and produced the custom charts with input from LM. MNH collected data and performed analysis.

Sclera filtering algorithm
MNH, TL and LM identified the need. MNH developed and implemented the algorithm, designed the lab-based experiment, collected data, and performed analysis.

Royal Free Hospital clinical study

RM identified the need; RM and TL organised the study. TL and JM came up with the concept of sclera imaging for jaundice. FO, AL, JWL, and FY developed the smartphone image capture apps. MNH developed and produced the flash diffuser. MNH designed the image capture protocol; RM proposed repeat captures and suggested their spacing. RM and MNH oversaw data collection; KK, MM, AB, RK, AA, EA, and MS collected data. MNH managed the database. FO proposed the use of chromaticity in developing a metric. MNH and TL proposed capturing images to investigate haemoglobin. MNH wrote software and identified regions of interest in images; FO provided overarching analysis structure, MNH improved on; KK and RM directed clinical analysis; MNH performed all analysis.

Sclera spectral modelling

MNH and TL developed models, MNH implemented and tested models.

Sclera spectroradiometric measurement

MNH proposed the concept; MNH, LM and TL developed the methodology. MNH and FO collected data; LM provided initial software, MNH improved on software and performed all analysis.

Acknowledgements

There are many people without whom I would not have been able to finish this PhD. First and foremost I would like to thank my primary supervisor Terence Leung. His insight and guidance have been crucial to moving my research on and particularly during the pandemic when all of our plans got derailed his optimism and genuine enthusiasm for research gave me a much-needed boost. Thanks also to Raj Mookerjee, my clinical supervisor, who taught me about the clinical context and co-ordinated the clinical data collection alongside his already very busy main clinical role. And thank you to Lindsay MacDonald, my unofficial colour supervisor, for so many wonderful discussions about colour science and image processing, for the generous loan of his spectroradiometer, and for injecting a healthy dose of purism into this interdisciplinary research. Thank you to my PhD friends, particularly Felix Outlaw who guided me so well into the field and was a key collaborator; and Thomas Wemyss who I hope felt similarly welcomed - pass it on.

Thank you to the clinical fellows at the Royal Free Hospital who were involved with patient recruitment and data collection: Mohammed Sheikh, Eman Alabsawy, Ahmed Amin, Rahul Kumar, Monica Mesquita, Ana Barreira, and Konstantin Kazankov. And thanks to Alice Skelton at the University of Sussex for lending her equipment and time for spectral sensitivity measurements.

And thank you to my friends and family who never questioned that I would complete this PhD, even when I doubted it. Matilde - the best flatmate and friend anyone could ask for. Ros and Alice - always in my corner. Kate - ready to lighten the mood with silliness. Katharine - a sympathetic listener. Charlie and Anita, my uncle and aunt - for so much support and guidance. My mum Lucia and sister Elizabeth - for proofreading so many pieces of work over the years, providing their favourite silly technical words from my publications, and always believing in me.

Thanks to baby Samuel (who is really now a toddler but will always be my baby) for help keeping work in perspective, and providing extra motivation to get this PhD finished. A final special thanks to my husband Reuben for keeping me grounded, filling my life with joy and supporting me throughout this process despite undergoing treatment for brain cancer - this is for you.

Contents

List of Figures

List of Tables

1. Background

1.1. Motivation for monitoring

Liver disease is the third most common cause of working-age premature death in the UK [1], with a UK death rate of around 15 per 100,000 people per year as of 2000 [2]. Whilst mortality rates have greatly improved for many chronic diseases, the mortality rate for liver disease in the UK increased by 400% between 1970 and 2010 [1]. Even more concerningly, this value increases to nearly 500% for patients under 65. With an average patient age lying between 55 and 60 years old [3–5], patients are dying within their working life, leading to not only tragic losses to their families, but a large economic impact and high costs to the NHS.

Liver disease is a general term for damage caused to the liver through a variety of issues. Liver disease is most commonly caused by alcohol abuse, hepatitis C or both, with around 75% of cases due to alcohol abuse [5]. Early on, liver disease is asymptomatic [6] so the best option for detecting cases is via biopsy. This approach is clearly impractical for screening, and using non-invasive testing it is not possible to catch all cases [7, 8]. The result of this difficulty is most patients not presenting to the healthcare system until they have already developed advanced cirrhosis [1]. Cirrhosis is the term for the scarring of the liver caused by long-term issues. As the liver disease develops further, the deterioration in liver function can be described by a few key stages. An acute decompensation (AD) event is where a major complication presents alongside the general deterioration [4]. An example of a major complication is ascites, where fluid builds up on the abdomen, sometimes in the order of tens of litres. Acute-on-chronic liver failure (ACLF) is an extremely serious condition which affects around 30% of patients who have already had an AD [5]. ACLF presents with organ failure as well as an AD and has a very high mortality rate, with 30% of patients with ACLF dying within 30 days [5]. As well as high mortality rates, patient quality of life is reduced even further for those with ACLF compared to those with decompensated cirrhosis. Additionally, quality of life is also affected for close relatives proportionally to patient quality of life [9].

Liver disease can be unpredictable and since it is characterised by periods of relative stability followed by sudden deterioration, patients are repeatedly admitted to and discharged from hospital. After discharge, around 20% of patients are readmitted within 30 days [3]. This process is clearly undesirable for a number of reasons.

Whilst at home, patients' quality of life and independence, whereas in hospital there are risks of infection and costs for care are very high [10]. Upon readmission certain tests get repeated, often unnecessarily, which can affect patients' faith in and opinion of the treatment system [3, 10]. Since outside of hospital patients have little or no access to any monitoring of their condition, it is up to them to gauge when they need to seek treatment. Due to very reasonable fear, this can lead to costly unnecessary readmissions which waste patient time and expose them to risks described above [11].

The far larger issue with readmissions is the concerning statistic that the death rate for hospitalised patients is around 27% in 90 days, whereas for outpatients it is only 10% [3]. This suggests that readmissions are happening too late in the deterioration to be able to offer sufficient care, supported by the finding that the death rate is increased for unexpected readmissions [12]. The schema in Figure 1.1 demonstrates this concept. The predicted natural deterioration in liver function over time for patients with liver disease is shown. A sharp decline is observed at the occurrence of an AD event, however the majority of function can be restored when treatment for this event is received soon after its onset. Far less function can be restored when there is a delay in treatment, and any future AD events are likely to be fatal.

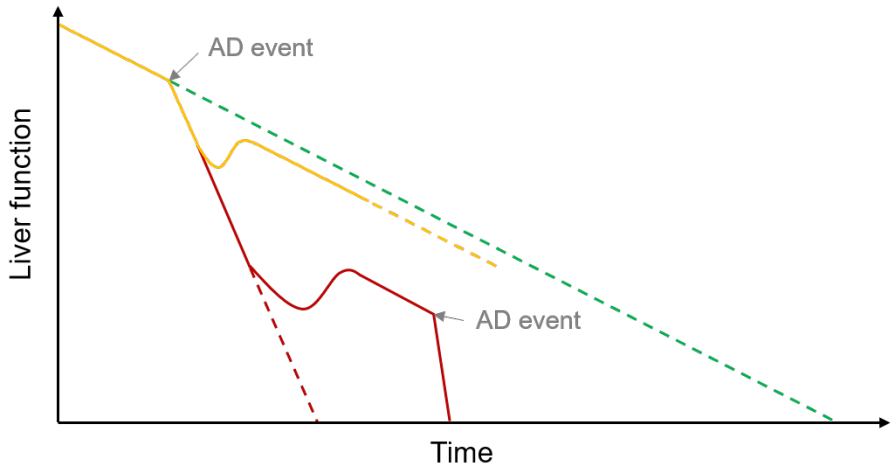


Figure 1.1: Typical decay in liver function over time is shown with a green dashed line. The rapid decline in function after an acute decompensating (AD) event is shown in orange, and the restoration of most function when treatment is rapidly received. The scenario when treatment is delayed is shown in red, where significant loss of function remains even after treatment and any further AD events are likely to be fatal.

It is therefore clear that early intervention is key to improving patients' quality of life and survival rate, with readmissions timed appropriately. There have been a variety of studies carried out aiming to reduce readmission. The kinds of interventions introduced can be split into specifically targeted or more broad, and simple to implement or more complex [10]. An example of a simple specific process is where paper and electronic checklists were introduced within the hospital setting targeting a specific subgroup of conditions [13]. The goal was to optimise inpatient care thereby reducing readmissions. A reduction in readmissions was found, however adherence to the process was low since it required a change to the standard workflow. Since our goal is to help time readmissions better, this approach is less relevant.

Broad, complex interventions involve reorganising care for all liver disease patients, for example by increasing inter-clinician communication and outpatient monitoring [10]. These kinds of interventions have been found to improve survival rates and simultaneously reduce readmission rates. A successful example of a broad complex intervention was implemented by Morando et al, and named the 'care management checkup' system [14]. The system involved an outpatient unit with an increase in sharing of information between the hospital and this unit, the use of a day hospital, and real time parameter evaluation. A reduction in readmissions was found. Whilst it is desirable to reduce readmission rates purely for patient care, the healthcare provider would still like to see a lowering in cost. A general study covering attempts to reduce readmission rates across a range of diseases and conditions found that the economic benefits of these interventions varied widely [15]. The interventions proposed by Morando et al found not only a reduction in readmissions and decrease in the 12 month mortality rates, but also a reduction in the overall costs involved because of the lower rate of costly hospital stays.

Success in these studies, combined with the desire to help time and target interventions, led to the goal of creating a low-cost monitoring system appropriate for use in the patients' homes. Additionally, the COVID-19 pandemic led to reallocation of resources in combination with a healthcare system already under immense strain, highlighting the need to move away from traditional treatment [16]. To be appropriate for use outside a hospital, the system must be simple to use and be based on one or more indicators of liver condition that can be measured non-invasively. The system would monitor these indicators, and through linking with a clinician who could critically review results, the system would be able to warn patients when it may be necessary to seek medical attention. This would hopefully reduce the time between, for example, AD events and treatment, and so both improve survival rates and decrease costs for the NHS.

1.1.1. Smartphones in medicine

Smartphones are becoming ever more ubiquitous worldwide with over 8 billion subscriptions by 2019 [17]. Their widespread availability and portability means that a large amount of research is being carried out into leveraging smartphone capabilities within different areas of medicine [18]. Some examples include GPS monitoring to support dementia patients with a tendency to wander [19], passive monitoring of factors such as sleep duration and kinesthetic activity to perform behavioural monitoring for mental health assessment [20], and performing a common test for lung function based on forcible exhalation, spirometry, using the microphone [21].

Use of smartphones within care for patients with liver disease has also been investigated [22, 23]. Examples include passive monitoring of sleep patterns and activity to try to predict urgent visits and hospitalisations [24], assessment of hepatic encephalopathy via an app version of the Stroop test, involving naming the colours of presented words not what the words say [25, 26], and monitoring of ascites via bluetooth connected weighing scales [27]. Importantly, a recent study found that the majority of cirrhosis patients both have a smartphone and are willing to use them as part of their disease management [28].

Kazankov et al recently reported the first work trialling home monitoring of cirrhosis patients using a smartphone app [29]. Existing monitoring devices were linked to the CirrhoCare® app to enable users to report heart rate, blood pressure, weight, % body-water, cognitive function (via the CL-Animal Recognition Test (CL-ART) App), self-reported well-being, and intake of food, fluid and alcohol on a daily basis. The group of 20 patients who had used the CirrhoCare® system were found to have fewer unplanned abdominal fluid drains as well as both reduced and shorter hospital admissions compared to an equivalent control group over 10 weeks of follow-up. This work perfectly demonstrates the potential improvement to patient care in the home when incorporating smartphone apps. Further improvements to patient outcomes may be possible when incorporating additional biomarkers, for example those used in existing clinical scores.

1.2. Clinical scores

In order to help clinicians gauge the severity of a cirrhosis patient's condition and predict mortality at set time points, a number of clinical scores based on different biomarkers have been developed. Of these, we will describe four commonly used scores in greater detail, highlighting the biomarkers they depend on in order to mo-

tivate a choice of biomarker for this system.

The first score is the Child-Pugh score. This was proposed in the 1970s and was originally developed to assess the risk of surgery in cirrhosis patients [30]. The score depends on five factors: bilirubin, albumin and INR levels, and hepatic encephalopathy and ascites grades. Bilirubin is a normal breakdown product from red blood cells, and is usually processed by the liver into a form which can be excreted. However, when the liver is unable to process the bilirubin, it builds up in the body, leading to elevated levels. Albumin is a protein generated by the liver, so decreased levels can indicate a struggling liver. INR stands for International Normalised Ratio for prothrombin time, which provides a measure of clotting. These first three factors are measured via a blood test. Hepatic encephalopathy (HE) is a decline in brain function caused by a drop in liver function. The severity of the effect varies, and is assessed for this score via the West Haven criteria which give a categorical level based on observed changes in factors such as behaviour and intellectual function [31]. The final factor is an assessment of the level of ascites, a build-up of fluid on the abdomen. These five factors were chosen empirically, based on clinical experience of managing patients with liver disease. Points for each factor are assigned based on thresholds and then summed to give a final score, with details shown in Table 1.1. The final score ranges from 5-15, where a higher score indicates a more severe illness. The Child-Pugh score is now commonly used in assessing prognosis, and is simple to calculate. Its downsides include that the factors were chosen empirically, the presence of a 'ceiling' effect where for example drastically increasing bilirubin still produces the same score, and the inclusion of the somewhat subjective factors of HE and ascites [32].

The second score is the MELD-Na score, which stands for Model for End-stage Liver Disease with sodium. The MELD score was originally developed for predicting survival after a shunt operation [33], with further development and focus on assigning

Factor	+1	+2	+3
Bilirubin (μmol/L)	<34	34 - 51	>51
Albumin (g/L)	>35	28 - 35	<28
INR	< 1.7	1.7 - 2.2	> 2.2
Encephalopathy (West Haven criteria)	None	1 - 2	3 - 4
Ascites	Absent	Slight	Moderate

Table 1.1: Definition of the Child-Pugh score. Contributions of the five component factors are assigned 1, 2, or 3 points depending on their level and these points are summed to give an overall score between 5 and 15 [30].

liver transplants [34]. Multivariate analysis was used to select three key parameters and a regression model with survival produced the formula for the MELD score as

$$\text{MELD} = 9.57ln(\text{Cr}) + 3.78ln(\text{Bi}) + 11.20ln(\text{INR}) + 6.43 \tag{1.1}$$

where Cr represents creatinine level (mg/dL), which is measured via a blood test and is linked to both kidney function and amount of muscle, and Bi is bilirubin (mg/dL) [34]. Later work by another group found sodium level was another important predictor of survival, and created the MELD-Na score based on the original score

$$\text{MELD-Na} = \text{MELD}(1 - 0.025(140 - \text{Na})) + (140 - \text{Na}) \tag{1.2}$$

where MELD is the original score given by Equation 1.1, and Na is the sodium level (in mmol/L, thresholded to the 125-140 range) measured via a blood test [35]. The MELD-Na score falls into the 6-40 range, with higher scores indicating more severe illness. Both the original MELD score and the MELD-Na score are now commonly used to assess prognosis, along with the Child-Pugh score.

The third score is the CLIF-C AD score, which stands for Chronic LIver Failure-Consortium Acute Decompensation score [4]. This score was developed specifically for patients with acute decompensation. As described in Section 1.1, the presence of acute decompensation is known to significantly affect mortality rates particularly within a short time frame, and so the scores described so far may not be as accurate for this patient population. Multivariate analysis was used to select predictors of mortality for this patient group, yielding a score definition of

$$\text{CLIF-C AD} = 0.3\text{Age} + 6.6ln(\text{Cr}) + 17.1ln(\text{INR}) + 8.8ln(\text{WCC}) - 0.5\text{Na} + 80 \tag{1.3}$$

where Age is the patient age (years), Cr is the creatinine level (mg/dL), WCC is the white cell count (10^9 cells/L), and Na is the sodium level (mmol/L). The CLIF-C AD score is thresholded to the 0-100 range, however in the original paper they found no values outside the 23-82 range [4]. As with other scores, higher values indicate more serious conditions - for this patient population, the CLIF-C AD score was found to have an improved predictive power compared to Child-Pugh or MELD-Na [4].

The final score is the ACLF grade [5]. As described in Section 1.1, ACLF is a very serious condition whereby organ failure occurs on top of acute decompensation. Patients with ACLF have even higher mortality rates than those with 'just' acute decompensation. The ACLF grade was developed to help assess patients specifically with ACLF. A grade of 0 indicates no ACLF, and values of 1-3 denote increasing

levels of severity. The grades are based around the number of organ failures, with criteria for organ failure for the relevant organs/systems summarised in Table 1.2. A patient is considered to have grade 1 ACLF if they fall into one of these three groups:

1. Single kidney failure
2. Single failure of the liver, coagulation, circulation or respiration with a creatinine level of 1.5-1.9 mg/dL and/or mild to moderate hepatic encephalopathy
3. Single cerebral failure with a creatinine level of 1.5-1.9 mg/dL

A patient is considered to have ACLF grade 2 or 3 if they have two or three organ failures respectively [5].

Organ or system	Criteria for failure
Liver	Bilirubin $\geq$12.0 mg/dL
Kidney	Creatinine $\geq$2.0 mg/dL or use of renal replacement therapy
Cerebral	Hepatic encephalopathy grade $\geq$3
Coagulation	INR $\geq$ 2.5 or platelet count $\leq 20\times10^9$/L
Circulation	Use of dopamine, dobutamine or terlipressin
Respiratory	$PaO_2/FiO_2 \leq$200 or $SpO_2/FiO_2 \leq$214

Table 1.2: Criteria for organ failure used in determining the ACLF grade [5]. Respiratory terms are as follows: PaO_2 is the partial pressure of arterial oxygen (mmHg), FiO_2 is the fraction of inspired oxygen, and SpO_2 is the pulse oximetric saturation (%).

1.3. Choice of biomarker

Each of the clinical scores described in the previous section is commonly used to help in clinical management, with some targeted to specific subsets of patients. A summary of the dependencies of the four scores is shown in Table 1.3. To help choose a useful biomarker to help assess a patient's condition, we consider those occurring repeatedly across common clinical scores. Table 1.3 shows that there are three parameters which appear in three or more of these scores. The first is INR, which appears in all four. However, there is no clear way to assess INR other than via a blood test. In addition, even if a non-invasive method were found, results have been shown to vary between labs, putting its reliability into question [36].

The second parameter is the creatinine level. This can be reliably measured via blood test, but in this patient group the results themselves may not be as meaningful. Creatinine depends on both kidney function and muscle mass, so a loss in

Factor	Child-Pugh	MELD-Na	CLIF-C AD	ACLF grade
INR	✓	✓	✓	✓
Bilirubin	✓	✓		✓
Creatinine		✓	✓	✓
HE grade	✓			✓
Sodium		✓	✓	
Albumin	✓			
Ascites level	✓			
White cell count			✓	
Age			✓	
Oxygen ratio				✓
Use of vasoconstrictor				✓

Table 1.3: Summary of the dependencies of the four clinical scores considered for this patient cohort. Each score is described in more detail in the main text.

muscle mass, which is common in very unwell patients, may result in healthier levels despite poor kidney function.

The final biomarker is the total serum bilirubin level (TSB). As described in Section 1.2, bilirubin only builds up in the body when the liver is unable to process it correctly, making it a clear indicator of a drop in liver function. As bilirubin builds up, it causes a visible yellowing of the skin and whites of the eyes known as jaundice. This visible change makes TSB a strong contender for non-invasive measurement. We have therefore selected the TSB level as the primary new biomarker for a home monitoring system, and have focussed around measuring it non-invasively.

1.4. Measuring bilirubin

When in hospital, the gold standard for measuring bilirubin is via a blood test. These are carried out daily as standard practice via a blood draw and provide an accurate reading. However use of blood tests outside a hospital is clearly less convenient owing to their reliance on a professional to perform a blood draw and subsequent lab analysis. A variety of approaches using a finger-prick test and advanced microfluidics exist, for example using a portable centrifugal analyser [37], phototreatment and image analysis [38], and three-dimensional tape-paper [39]. All the approaches have the drawback of being invasive, and typically require a highly trained operator due to their complex nature.

Measurement of bilirubin level via the skin discolouration, known as transcutaneous bilirubinometry, is a well-developed technique which has been widely adopted

in screening for neonatal jaundice [40]. There are different competing devices on the market, but in general they work in similar ways: the skin is exposed to light of multiple wavelengths and the ratio of incident to returning light is measured. The returning light will have been affected by various light-absorbing molecules, or chromophores, in the skin, such as bilirubin, melanin and haemoglobin. It is then possible to extract only the contribution of the bilirubin and so obtain an estimate for the TSB level. The use of transcutaneous bilirubinometer (TcB) devices is attractive since they are simple to operate, non-invasive and provide real-time results. However, they are prohibitively expensive for outpatient use, with each device costing in the order of £3,400 (NHS costing report, 2010 [41]). More concerningly, it has been found TcB readings are not accurate for adult populations [42], and it is not clear that TcBs provide consistent readings across different ethnic groups even for neonates [43].

1.4.1. Image-based measurements

Doctors regularly use a visual assessment as a pre-screening method for jaundice, and colour is used qualitatively in other areas of medicine such as using the pallor of the lower eyelid as a check for anaemia. However, this approach requires experience to notice subtle changes and becomes progressively harder for higher levels of jaundice [44, 45]. Ideally this qualitative measure would be upgraded to quantitative measurements, and one way to do this is via non-invasive images. There are a range of areas which seek to quantify colour via images, for example quantifying colorimetric urine tests for measurements of pH and glucose [46–49] or determining saliva alcohol concentration [50]. Applications continue beyond medicine, for example in testing water quality [51], improving the rigour of marine monitoring [52], and performing quality control of beer colour [53]. There are also applications within medicine which aim to quantify colorimetric biomarkers of the human body to detect conditions such as anaemia [54] and the eye condition anterior blepharitis [55].

For imaging, digital cameras provide very high image quality but they can be expensive and bulky to transport. Smartphones, on the other hand, are incredibly portable and are becoming even cheaper and more ubiquitous, with over 8 billion subscriptions by 2019 [17]. With a smartphone, unlike with a digital camera, it is possible to create an app which combines the image capture with any required additional processing to produce results in real time. All of these factors, combined with the continued increase in smartphone image quality, makes them the ideal candidate for use in quantifying colour via images.

The first major work on quantifying bilirubin via images came from de Greef et al, who imaged 100 neonates and found a 0.85 rank order correlation with the blood test value for TSB [56]. Images containing the forehead, sternum, and a basic custom colour chart to carry out colour correction were captured using an iPhone. The sternum was selected as the final region of interest, due to its lower likelihood of sun exposure which can slightly reduce the bilirubin level in the skin, and regions of interest were manually segmented. Results from a conversion to a series of colour spaces were fed into a machine learning regression algorithm to produce the TSB estimate. The same group did a larger follow-up study on 530 neonates from a range of ethnic backgrounds and obtained a 0.91 correlation with the blood test values [57]. Another group, in collaboration with the spin-out company Picterus, again used a custom colour chart placed on the baby's sternum to colour balance and then used a database of paired colour values and TSB values to produce a TSB estimate [58]. With a sample size of 220 neonates they achieved a 0.84 correlation.

Whilst the skin may seem like the ideal imaging target, owing to its large area and visible discolouration, it is highly affected by melanin as described when discussing TcB devices. An alternative imaging site is the white of the eye, or sclera. The sclera becomes visibly yellow for elevated bilirubin levels, but for humans does not contain melanin, meaning that the healthy baseline colour is the same across different ethnicities [59]. The first study using the sclera as the imaging site was performed by Laddi et al, using a digital camera, ring light and housing to block ambient light. They collected images of 330 patients with jaundice and 90 volunteers, but the results presented are minimal and do not describe the range of bilirubin levels considered [60]. A more detailed demonstration of sclera imaging for neonatal jaundice screening was done by Leung et al, who used a digital camera to capture images under roughly controlled ambient conditions [61]. They carried out white balancing using a chart included in each image of the sclera, and then used a multiple linear regression to obtain TSB estimates. With a sample size of 110 neonates this simple approach obtained a correlation of 0.75 with TSB values. The concept of sclera imaging was picked up by other groups, with a group in Saudi Arabia obtaining a 0.73 correlation for a sample size of 50 neonates [62].

The first major paper focussing on an adult population was published by Mariakakis et al [42], which aimed to screen for very early stage jaundice as a biomarker for pancreatic cancer using smartphone imaging. Two different approaches for dealing with variations introduced by ambient light were considered: a pair of card glasses incorporating a custom colour chart, and a box in the style of a VR headset which blocked all ambient illumination and allowed the phone to provide the lighting. An

automatic sclera segmentation algorithm was used followed by a machine learning approach based around random forest regression to obtain TSB estimates. The box approach was found to be more successful, and for the sample size of 70 patients a 0.89 correlation was obtained. Other groups have attempted to develop low cost methods for adult jaundice screening. Sammir et al used a similar style of goggles to Mariakakis with a built-in light source and a webcam to capture images [63]. They used a hue-based metric with 25 participants to try to assess bilirubin level. Miah et al tried to develop an ultra-low cost system for detecting liver disorder in rural cohorts, also using a webcam to capture images, with lighting provided by a flash-light [64]. Based on images of 25 participants, using a variety of machine learning techniques, the jaundice level was categorised as mild, moderate or severe. These studies demonstrate the interest in the engineering community in tackling the issue of adult jaundice, but the existing studies have small sample sizes, and often do not present the range of bilirubin levels considered or the underlying causes of elevated bilirubin levels.

1.5. Objectives

In this work a system focussed on tracking bilirubin is presented, in order to provide home monitoring for liver disease patients. The system is based on smartphone images, since smartphone imaging is cheap, portable, and non-invasive. It uses the sclera as its region of interest, since the sclera is free from the complicating factor of melanin found in the skin. The system has the following key aims:

- To provide a method of processing such that image capture is simple and does not require accessories in each image
- To develop the calibration and processing such that the readings the system produces are independent of the device and imaging location used for image capture
- To produce useful results over the large range of bilirubin levels found in liver disease patients.

The cheap and non-invasive nature of imaging, combined with a simple image capture protocol and the ability to use different phones for image capture, should make the system appropriate for use in a patient's home.

Chapter 2 discusses the challenges of quantifying colour, focussing particularly on accounting for changes in ambient light and obtaining device-independent measurements. General image theory and an overview of the different approaches used to

try to overcome these challenges are presented. Chapter 3 details our approach for obtaining accurate colour values, including experimental testing and validation. Chapter 4 then moves on to the challenge of linking colour to bilirubin, presenting results from a patient study of adult liver patients. The majority of patients with chronic liver disease have high bilirubin levels, so it is necessary to go beyond screening for abnormally high levels. The goal is to be able to use this approach not just to screen for jaundice, but to differentiate between different levels of bilirubin in order to target interventions more effectively and hence reduce mortality rates. Finally, Chapter 5 discusses work on obtaining spectral information for sclerae to further improve the system.

2. Why is it difficult to quantify colour?

One goal of this research is to be able to obtain accurate colour values across devices and lighting conditions so that the derived link between colour and bilirubin can be applied to data from different phones under different illuminations. In this chapter, the reasons that this is difficult will be explored. Figure 2.1 gives a basic schematic of the components of image formation, highlighting the challenges of separating information about the scene (here, the sclera). Since the scene, lighting and camera combine to give the final recorded image, changes to any component will lead to different recorded colour values. The theory behind acquiring an image will be discussed in detail in the next section to enable a deeper understanding of the challenges of measuring colour. Approaches to standardise for lighting and device variations will also be outlined. Our proposed approach and associated validation experiments will then be presented in Chapter 3. Parts of the work presented in this chapter are published in a PLOS ONE article [65] and in the Color and Imaging Conference 27 and 28 proceedings [66, 67]. These papers, licensed under the creative commons license CC BY 4.0, have been modified to form part of this thesis. Additionally, there is some overlap in the content of the theory presented here with the MRes thesis of the same name which formed the first year of this PhD program.

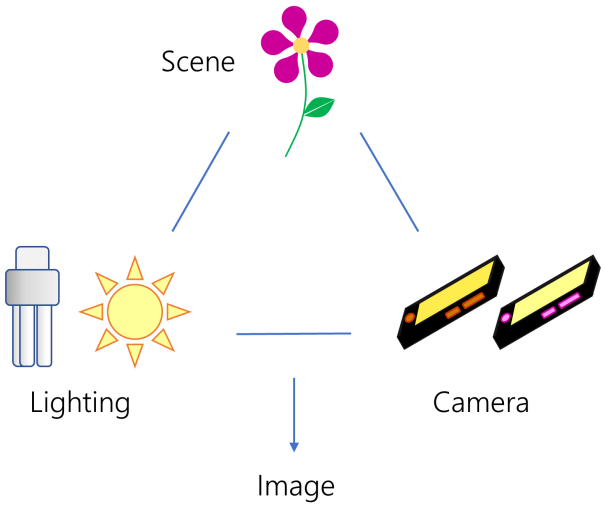


Figure 2.1: A schematic of image formation highlighting the challenges of quantifying colour. The final image is influenced by the scene, the lighting and the camera. So even if the scene remains constant, a change in lighting or use of a different camera leads to different values being recorded. In order to obtain stable colour values, these changes must be accounted for.

All analysis, unless otherwise stated, was carried out using MATLAB (MathWorks r2020b).

2.1. Image formation

As depicted in Figure 2.1, there are three main components of image formation. The first is the ambient lighting, which depends on wavelength and may also depend on space and time. The relative intensity of light at different wavelengths is what changes the colour of the light that we see, for example the blueness of fluorescent lighting compared to the redness of candlelight. If a room contains light sources with different spectral characteristics, for example fluorescent light and a candle, then the spectrum of the ambient lighting will depend on position within the room. Additionally, if one of the light sources is removed during measurement, for example the candle is blown out, then clearly time will also be a factor. For the purposes of this work, we will assume that ambient light depends only on wavelength given the time-frames and scales of capturing images of the human face.

The second component of image formation is the scene - what you are trying to image. The optical properties of the objects in the scene determine what proportion of incident light at each wavelength is reflected back. As well as wavelength, properties of the scene may depend on space, for example imaging several different surfaces, or time when imaging something moving. As for the lighting, here we will assume that the scene does not depend on space or time. As well as wavelength and space dependence, the relative angle between incoming light and the particular surface will affect the returning light so must be considered. The energy reflected from a surface can be described in equation form as

$$E(\lambda, \boldsymbol{x}) = m(\boldsymbol{x})s(\lambda, \boldsymbol{x})e(\lambda) \tag{2.1}$$

where $E(\lambda, \boldsymbol{x})$ is the energy as a function of wavelength, λ, and space, $\boldsymbol{x}$, $e(\lambda)$ is the lighting, $s(\lambda, \boldsymbol{x})$ is the surface reflectance, and $m(\boldsymbol{x})$ represents the geometric dependence of the reflectance [68].

The final key component of image formation is the camera itself. Rather than recording the reflected light at every wavelength, cameras instead use just three channels to represent the information. By using channels in the red, green, and blue parts of the visible spectrum, referred to as RGB from now on, the majority of colours that humans are able to perceive can be reproduced. If the observed intensity is assumed to be independent of the viewing angle, known as the Lambertian

model, then the values recorded by each colour channel $c \in \{R, G, B\}$ are given by

$$f^c(\boldsymbol{x}) = \int_\omega E(\lambda, \boldsymbol{x})\rho^c(\lambda)d\lambda = m(\boldsymbol{x}) \int_\omega s(\lambda, \boldsymbol{x})e(\lambda)\rho^c(\lambda)d\lambda \tag{2.2}$$

where ω represents the visible spectrum, ρ^c the spectral sensitivity of each colour channel, and all other terms are as previously described [68]. This common model only accounts for the effects of diffuse reflection, but specular reflection may also occur. Diffuse reflection refers to when light interacts with the surface, whereas specular reflection occurs when light is reflected off the surface. In the case of specular reflection, its spectral properties are determined only by the lighting. In an analogous way to Equation 2.1, the specular energy can be described as

$$E_s(\lambda, \boldsymbol{x}) = m^s(\boldsymbol{x})e(\lambda) \tag{2.3}$$

where the subscript s refers to the influence of specular reflection, and all other terms are as previously described [68]. The effect of specular reflection can be taken into account using a dichromatic model, where recorded colour values are given by

$$\begin{aligned} f^c(\boldsymbol{x}) &= \int_\omega (E(\lambda, \boldsymbol{x}) + E_s(\lambda, \boldsymbol{x}))\rho^c(\lambda)d\lambda \\ &= m(\boldsymbol{x}) \int_\omega s(\lambda, \boldsymbol{x})e(\lambda)\rho^c(\lambda)d\lambda + m^s(\boldsymbol{x}) \int_\omega e(\lambda)\rho^c(\lambda)d\lambda \end{aligned} \tag{2.4}$$

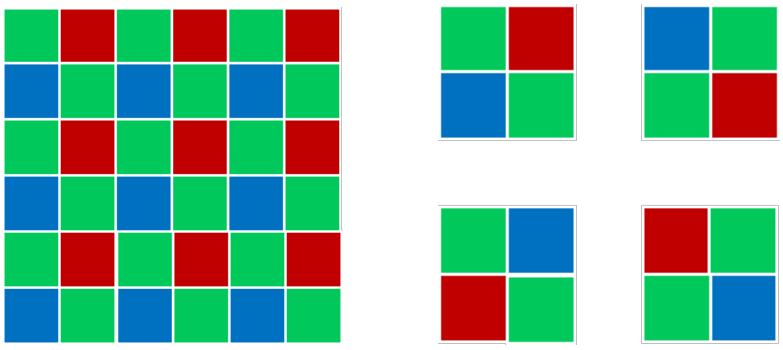
where all terms are as previously described, and it can be seen that specular reflection simply produces an extra term compared to Equation 2.2 [69]. When capturing images of the eye, specular reflection is very common due to the layer of tear protecting the surface of the eye. It is therefore very important to consider its impact on recorded images. With appropriate exposure times and ISO values for the rest of the scene, the specular reflection often saturates the camera's sensors. In this case, it is straightforward to remove its effect by simply removing saturated pixels from analysis. However, in some cases specular reflection does not saturate the sensors and so the reflection modifies the resulting colour of the surface and is harder to exclude.

In order to identify and remove these pixels, a filtering algorithm was developed which is described in more detail in Chapter 4.2. Regions affected by specular reflection are therefore excluded from analysis, and so the simpler Lambertian model for image formation in Equation 2.2 is more appropriate to describe results. An inspection of this equation confirms that the factors of image formation are not directly separable, and so an alternative approach is necessary to separate out the contribution from the scene alone.

2.2. Recording images

The previous section described the process of image formation. Here, the process of recording these values will be discussed. Each pixel in an image is described using a combination of three RGB values. However, it would be too complex and expensive to record an R, G, and B value for each sensor site so in practice only one is recorded at each location. This is achieved by using a combination of sensors and overlaid filters, which allow light through in just the red, green or blue wavelength bands. Later, interpolation is used to recover full RGB values for each pixel. The filters for recording each of the channels are arranged using what is known as a Bayer pattern [70], designed to enable a high quality reconstruction through even distribution of R, G and B sensors across the sensor area. An example Bayer layout is shown in Figure 2.2 (a). Note that there are twice as many green sensors compared to red and blue since the human visual system is more sensitive to green. This means that the final image is perceived to be higher quality when green is recorded most accurately. Figure 2.2 (b) shows all four permutations of Bayer layouts. Since different manufacturers use different layouts, it is important to be aware of which layout the camera of interest uses.

So far, the RGB values have been described as though they were measured using sensor-filter combinations covering fixed non-overlapping wavelength regions. In reality, this is not the case. Whilst each filter sensitivity peaks in either the red, green or blue region of the visible spectrum, there is some overlap between filters and the exact shape varies between different cameras. The combined wavelength-dependent sensitivities of the sensors and filters are called the camera spectral sensitivity (CSS). An example normalised smartphone CSS is shown in Figure 2.3. The relative heights



(a) (b)

Figure 2.2: (a) Example Bayer filter layout of R,G,B sensor sites (b) The four kernel permutations for Bayer arrangements, shown as different layouts are used by different manufacturers.

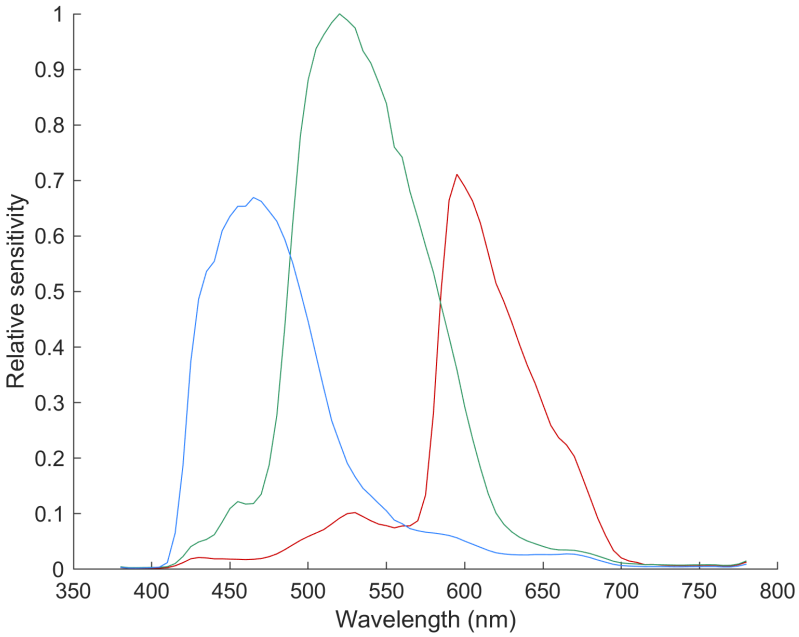


Figure 2.3: Normalised camera spectral sensitivity for an example LG Nexus 5X smartphone camera. Sensitivities for each colour channel from long to short wavelength are shown in red, green and blue respectively. The data presented here was measured using the monochromator method, with more details presented in Appendix A.

of the curves represent the likelihood of a photon of any given wavelength to be detected by a red, green or blue sensor site. Note that towards the edges of the sensors variations in CSS can occur, along with vignetting effects, so it is advisable to use the central region of the sensor only [71].

From the recorded Bayer pattern image, various stages of processing are then automatically carried out by the camera to produce the images we are used to seeing. RGB values for every pixel are produced through a form of interpolation known as demosaicing, and various other scaling and compression steps are carried out. As the popularity of smartphone imaging in everyday life has increased, so too has the complexity of post-processing. A huge amount of work is done behind the scenes to yield visually pleasing images for users. Unfortunately for our purposes, an aesthetically pleasing photo is not an accurate one, and every manufacturer has their own typically secret processing pipeline to obtain these images. In this work we have therefore chosen to analyse the raw Bayer pattern images directly, accessed via dng images which many smartphones have the ability to save, allowing us complete control over any processing applied.

2.3. Colour constancy

The process of accounting for ambient lighting so that our perception of colours remains stable is an automatic part of the human visual system. Whether we read a book outside or inside under fluorescent lighting, we are able to see the pages as white. This correction is not an automatic process for a camera where very different pixel values for the pages of the book would be recorded under these different illuminations, as seen in Equation 2.2 and depicted in Figure 2.1. Correcting images taken in different ambient conditions is crucial here and in other areas which aim to quantify colour in and between images - we must be confident that a registered change is due to a change in patient condition and not just in the background lighting.

There are a huge variety of approaches which aim to remove the effects of ambient light. It is possible to avoid the issue entirely by blocking ambient light and providing a fixed illumination [42, 50, 51]. An alternative approach is to include a standardised white card in every image, using white balance to account for lighting changes [56]. This second approach and the majority of approaches which follow are vulnerable to metameric illuminations: where the white point is recorded as the same, but for different spectra. Both of these approaches require additional pieces of equipment for the capture process to work which significantly complicates image capture.

One category of approaches is those aiming to estimate the ambient light in order to correct for it, appropriate here since these approaches maintain RGB images rather than reducing to colour constant quantities [72]. The most simple approach is MaxRGB (also known as Scale by Max or WhitePatch). In this approach the ambient illumination is estimated as the maximum value in each channel, calculated independently across the channels [72]. MaxRGB performs well provided there are surfaces which are maximally reflective for each channel, and provided there is no clipping due to channel saturation. A similar reliable approach is Gray World, where the average for each channel across all pixels provides the ambient estimation under the assumption that the average of an image is approximately neutral grey [73].

An alternative way to estimate the illumination for images containing the human face is to use the pixel values of the sclera as the ambient estimate [74]. This approach has yielded good results, however this is clearly not appropriate for us as the variation in sclera colour is the exact quantity we are interested in. Specular reflection in the image could be exploited as a way to determine the illumination,

since by definition it does not depend on the surface. However specular values are often clipped due to the limited dynamic range of the camera, which would skew the results [75].

Approaches described so far fall into the static category - they are based on analysis on a per-image basis. An alternative category of approaches uses training data to obtain illumination estimates for new images. Within trained approaches, probabilistic methods calculate the probability of observing the RGB responses for a library of possible illuminants and then select the most likely illumination from the library [75]. Gamut mapping is an alternative strategy, where a set of plausible illuminants is determined by checking which produce values within the expected gamut for all pixels in the image [72]. A final choice for the scene illumination can then be selected in a variety of ways. The final and very common subset of trained approaches is machine learning, in which there is a large amount of work being done [76]. Machine learning has the ability to yield good results but requires very large training sets which continue to increase in size the more different lighting conditions are needed. Despite all the different options, it is challenging to find an approach to deal with ambient light which is both general and simple enough for our application.

2.3.1. Ambient subtraction

All of the approaches for colour constancy described so far rely on a single image of the scene. This is desirable since it maintains a simple capture process, however it is not much more time consuming to capture two images of the scene in quick succession. There is a growing interest in leveraging flash/ no-flash image pairs for illumination correction [77–80]. Once image pairs are registered and any difference in exposure accounted for [78, 79], the no-flash image can be subtracted from the flash image. The pixel values of the flash image result from a sum of the two effective light sources present - the flash and the rest of the ambient light, whereas the no-flash values are only influenced by the ambient light. When the images are subtracted, the result is an image as though captured under a pure flash illumination. For a given device, results are therefore standardised over different ambient conditions and so can be directly compared.

More mathematically, the flash image values can be represented as

$$f^{flash} = f^F + f^A \tag{2.5}$$

where the values of the flash image f^{flash} are influenced by both the flash F and ambient A illuminations. The no-flash image is only influenced by the ambient lighting, $f^{noflash} = f^A$. The subtracted image is therefore given by

$$f^{flash} - f^{noflash} = (f^F + f^A) - f^A = f^F \tag{2.6}$$

and is the equivalent of capturing an image with no ambient lighting and flash illumination only. When using smartphones, images can be captured using the rear-facing camera with the camera flash, or using the front-facing camera with the screen as the 'flash'. In both cases, it is important that the ambient lighting remains constant over the short time needed to capture the two images, else the contribution of f^A to the flash image would change. Additionally, it is important that the response of the sensors across the three channels is linear with increasing intensity - doubling the intensity should double the pixel values. The linearity of the sensors of the phones used in this study were checked. The sensors were found to be linear, and the results are presented in Appendix C.

One issue that remains is motion between the capture of flash and no-flash images. For real-world image capture, a tripod is not appropriate so hand-held images are captured. There will always be some movement between the two images due to hand motion, and this misalignment becomes even more inevitable when imaging a human subject who may also move slightly. To consider the impact on obtaining subtracted data, we consider an example physical point within the scene, p. For the flash image, p appears at position $\boldsymbol{x}$. With no motion, p would also appear at $\boldsymbol{x}$ in the no-flash image. However, in reality p appears at a slightly different location, $\boldsymbol{y}$. If these points are known, then the subtracted image value for p is given by

$$f_p^F = f^{F+A}(\boldsymbol{x}) - f^A(\boldsymbol{y}) \tag{2.7}$$

where f denotes image values. To obtain an average value for the sclera, a region of interest would be defined. Again, if the corresponding regions are known in both images then the average value can be written as

$$f_R^F = average(f^{F+A}(\boldsymbol{X}) - f^A(\boldsymbol{Y})) \tag{2.8}$$

where R denotes the region of interest, and $\boldsymbol{X}$ and $\boldsymbol{Y}$ denote the image co-ordinates for the flash and no-flash images respectively. In order to obtain these corresponding regions of interest, an extremely good image registration would have to be carried out. Any mismatch between the images could otherwise cause systematic offsets for this pixel-wise subtraction approach. An alternative much simpler approach has

been proposed and demonstrated by our group in the context of neonatal jaundice [81]. The goal is to obtain a single value for the colour of the sclera which is independent of ambient lighting, so we do not need information on a per-pixel basis. Instead, the average across the region of interest is carried out for the flash and no-flash images before subtraction such that the final value is given by

$$f_R^F = average(f^{F+A}(\boldsymbol{X})) - average(f^A(\boldsymbol{X})) \tag{2.9}$$

where all terms are as in the previous two equations. Note that the same region of interest is used in both images, meaning that slightly different physical points are considered in the two images. Carrying out the averages before subtraction means that any small shifts between the images are mitigated, provided the shift is small compared to the size of the region of interest. It is also possible to use a different region of interest for the ambient image to help when there is more movement, however unlike in Equation 2.8, when using the form given in Equation 2.9 there does not need to be a pixel-wise correspondence between the regions of interest.

The ambient subtraction approach is generalisable across different lighting conditions, does not require additional equipment at the time of image capture, and is comparatively simple to implement, so it has been selected for use here. Experimental testing and validation is presented in Chapter 3.1.

2.4. Chromaticity

When capturing an image, in order to make the best use of the dynamic range of the camera, the exposure time and ISO should be varied to ensure that the image is neither saturated nor overly dark. Variations in the intensity of the ambient illumination mean that different settings are optimal for different conditions. Whilst varying these settings can improve the signal to noise ratio, the recorded colour values will clearly be affected. Additionally, the region of interest is likely to be affected by geometric shading, observed as pixels further from the camera or on an angle having lower pixel values. What these two issues have in common is that the RGB channels will be scaled equally. The issues of shading and illumination intensity are therefore each summarised by a single unknown scaling factor.

In order to remove the impact of the scaling, rather than using the raw RGB values, we can instead use chromaticity values. Chromaticity values are a way of describing the relative amount of a colour, and are defined as a channel value divided by the

sum of all three channels,

$$r = \frac{R}{R+G+B}, \quad g = \frac{G}{R+G+B}, \quad b = \frac{B}{R+G+B} \tag{2.10}$$

where lower-case letters are used to denote a chromaticity value, and any scaling factor across the channels will cancel out [68]. If the exposure time is changed by a factor of α and we consider red chromaticity as an example we see that the final chromaticity value is independent of the scaling factor

$$r = \frac{\alpha R}{\alpha R + \alpha G + \alpha B} = \frac{R}{R+G+B} \tag{2.11}$$

Using chromaticity therefore allows us to account for these shading and illumination intensity changes. Chromaticity values always sum to 1, meaning that our data has been reduced from three dimensions to two. Whilst using chromaticity means a loss of some information, the standardisation across these different scenarios is key.

2.5. Device independence

As discussed in Section 2.2, the colour values returned by two different phones for the same object, even under identical illuminations and using raw images, will be different. This difference is caused by variations in the spectral sensitivity of the cameras and in the spectral power distribution of the flashes. In order to understand how to obtain measurements that are device-independent, some additional theory on colour spaces will now be outlined.

2.5.1. Colour spaces

It is possible to gain understanding of colour spaces by considering traditional mathematical vector spaces. The traditional unit vectors are here given by three linearly independent light sources known as primaries. Rather than a vector space, the span of the colours described by these primaries is known as the colour space, where these colours are described by numbers known as tristimulus values [82]. The range of colours described by a particular space with respect to another space is known as its gamut. A two dimensional diagram or plot is a useful way to visualise three dimensional colour spaces, so a chromaticity diagram is commonly used. xy chromaticity values are plotted on the horizontal and vertical axes respectively, where chromaticity has been defined in the previous section. An example chromaticity diagram is shown in Figure 2.4, where a triangle connecting the three primaries of an example space denotes the gamut of the space. Colours made from a single wave-

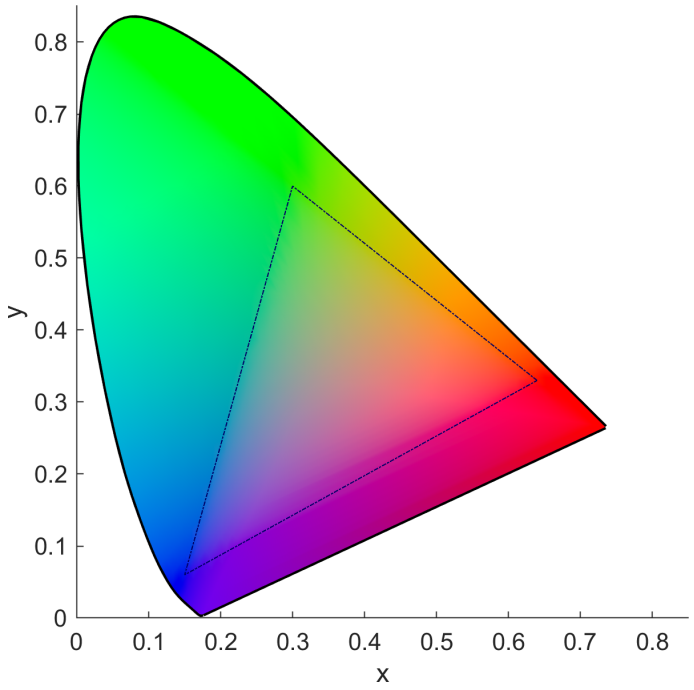


Figure 2.4: The xy chromaticity diagram is shown, along with the gamut of the sRGB space marked using a dot-dash line. xy horseshoe generated using code from [83].

length of light are found around the edges of the horseshoe. More specific detail on the diagram and the spaces shown will now be discussed.

A variety of colour spaces exist, based around different primaries. The space most commonly used by displays is known as sRGB space, whose triangular gamut is marked on Figure 2.4. The gamut of this space is relatively small, but is nevertheless commonly used and is device-independent. Some device-independent spaces contain all possible colours, a property which gives them the additional title of reference colour spaces. In order to compare results obtained using different phones, we would like to present the data using a device-independent reference space. From inspecting Figure 2.4, it is clear that to encompass the whole horseshoe non-visible, or imaginary, primaries must be chosen.

The most common reference space is CIE XYZ space, whose chromaticity diagram is shown in Figure 2.4. This space was derived from the human visual system, with the aim of producing XYZ tristimulus values describing which combinations of light appear the same for a standard observer. Since its introduction in 1931, other spaces such as CIE L*a*b* space have been designed for increased perceptual uniformity [68]. This means that differences in the L*a*b* space have a more direct relationship to what the human eye would be able to discern - a higher difference

in value corresponds to a larger difference to the eye. The aim of the research presented here is not to mechanize human colour judgements but rather to obtain repeatable digital colour descriptors that can then be linked to the application specific scale, total serum bilirubin. Therefore, XYZ space has been chosen for use as a standard device-independent colour space. Additionally, colour space conversions amplify noise as well as introducing colorimetric error [84], so where possible it is good to minimise the number of conversions. The conversion to XYZ space accounts for variations in the phone spectral sensitivities, but it is also necessary to account for the different flash illuminations. To do this a set standard illuminant, here CIE D50, is chosen for the XYZ values. This means that XYZ values resulting from a conversion from two different phones should match.

When a raw image is captured using a camera or smartphone, it is in a space we have not discussed yet. The image is in a space specific to the device, known as the native camera space. A linear transformation to XYZ space should be possible since the camera native space theoretically includes all colours [82]. Therefore the conversion from native RGB values to device-independent XYZ values should be straightforward. However, metameric error is present due to filter design and image noise. This error means that some pairs of surfaces which should have identical tristimulus values are recorded with different values. Therefore, the linear transformation can only ever be approximate [82, 85, 86]. Applying the transformation from native RGB to a reference space will introduce error, and could reduce the range of measurable colours. However the benefit outweighs the cost, since after applying the mapping, the device-independent space allows direct comparison of results from different phones.

2.6. Mapping to XYZ

There are a number of different ways to convert from camera native space to XYZ space which vary in accuracy and simplicity to determine. The theory behind the three most common approaches is outlined here, with a comparison of the accuracies for the two overall viable approaches given in Chapter 3.2.3. All of the approaches involve using a series of corresponding pairs of native RGB and XYZ values to develop a link between them.

2.6.1. Spectral method

The most theoretical approach, commonly known as the spectral method, involves measuring camera and illumination properties and then modelling both RGB and

XYZ values. The camera spectral sensitivity (CSS), how sensitive the camera is to light of different wavelengths for the red, green and blue channels, must be measured. Additionally, the spectral power distribution of the illumination, here provided by the phone, must also be determined. Finally, in order to generate RGB and XYZ values, a series of target colours must be chosen and their reflectances measured. As discussed in Section 2.1, the product of the reflectance, spectral power distriubtion (SPD), and CSS produces RGB values. XYZ values can be produced in a similar way, using a slightly modified form of Equation 2.2

$$X, Y, Z = \int_{\omega} s(\lambda) e_{D50}(\lambda) \, CMF^{X,Y,Z}(\lambda) d\lambda \tag{2.12}$$

where $s(\lambda)$ still represents the reflectance of a given target colour, $e_{D50}(\lambda)$ represents the standard CIE D50 illuminant (another illuminant could be used, for example CIE D65), and $CMF^{X,Y,Z}(\lambda)$ is the XYZ colour matching functions. This formulation of XYZ values is presented using an integral, however in reality reflectance of physical surfaces are measured at discrete intervals. Equation 2.12 is therefore rewritten as a sum

$$X, Y, Z = \sum_{i} s_i(\boldsymbol{x}) e_{D50,i} \, CMF_i^{X,Y,Z} \Delta\lambda \tag{2.13}$$

where i represents measurements at each wavelength interval, and other terms are as described above. When using the spectral method, the surfaces considered and the target illumination can be changed very simply. A drawback of this method is that each device of interest must be fully profiled. The SPD of the phone's illumination can be measured quite straightforwardly using a spectrometer. However, the CSS is more complex to measure. Typically CSSs are measured by exposing the camera to the output of a monochromator and recording its response as the full wavelength range is scanned (see Appendix A for details), a process which is both time consuming and requires extremely expensive equipment. Equipment based on a single capture of illuminated interference filters to form a 'test chart' has been developed which simplifies this process, however the device is still expensive [87]. Cheaper alternatives have been proposed [88, 89], but these come with a significant loss of accuracy in the measurement. Use of a smartphone add-on [71] or estimation via a ColorChecker image [90] have also been proposed but are error-prone. The processes of measuring the SPD and CSS of a phone accurately require multiple pieces of expensive equipment, and would significantly limit the simplicity of introducing a new phone for image capture. The spectral method was therefore not considered in further detail for this application.

2.6.2. Image metadata methods

At the other extreme of the complexity scale from the spectral method are the metadata methods. One way to achieve the conversion from phone native space to XYZ space is to use pre-determined mapping information stored in the metadata to do the conversion - this information is optimised only for images under a particular illumination and for a generic phone of that model, but may still provide adequate results. Using metadata methods would make the calibration process extremely simple so they have been considered here.

dcraw approach

In practice, the simplest way to utilize metadata information is to use the widely used open-source software dcraw [91]. The method implemented by dcraw uses one of two calibration matrices stored in the raw metadata – these matrices map from XYZ space to phone native space, based typically around illuminant D65 and one other standard illuminant. The dcraw implementation uses the D65 calibration matrix, and so the following description refers to D65 for clarity, however the same approach could be used for the other matrix. The overall mapping is achieved by applying the following matrices to the recorded RGB values [82]

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{D65} = s\underline{C}^{-1}\underline{D}_{D65}^{-1}\underline{D} \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{scene} \tag{2.14}$$

where C is the stored colourmatrix moving from XYZ to native space; s is a scaling factor determined such that the whitepoint of D65 in XYZ [0.9504, 1, 1.0888] is mapped to a native space green channel value of 1 upon application of $\underline{C}$; and $\underline{D}$ is a 3x3 diagonal matrix of white balance multipliers, designed to move the recorded values for a white object to [1,1,1]. These multipliers can be obtained through knowledge of the RGB values for white under the scene illumination - a test image of a neutral patch could be captured with little ambient light with the phone illumination. $\underline{D}_{D65}$ is another 3x3 diagonal matrix, this time inverted to move [1,1,1] to the white point of D65 in native space. In other words, the diagonal entries of $\underline{D}_{D65}^{-1}$ are simply the white point of D65 in native space. These values can be obtained as follows

$$\begin{bmatrix} R(\text{WP}) \\ G(\text{WP}) \\ B(\text{WP}) \end{bmatrix} = \frac{1}{s}\underline{C} \begin{bmatrix} X(\text{WP}) = 0.9504 \\ Y(\text{WP}) = 1 \\ Z(\text{WP}) = 1.0888 \end{bmatrix}_{D65} \tag{2.15}$$

where WP represents the white point, and s and $\underline{C}$ are applied to the XYZ value of the D65 whitepoint to move to native space. Throughout this research we use a D50 whitepoint, so the final step is to shift the whitepoint to D50. For this a chromatic adaptation transform (CAT) is used. There are a wide range of options, here the commonly used Bradford transform was applied [82].

Upon implementing the dcraw method in MATLAB it was discovered that in line with a mention in the literature [82], dcraw makes the assumption that the colourmatrix for D65 illumination is always stored as the second colourmatrix in the metadata. Whilst this may well be the case for digital cameras, by considering the CalibrationIlluminant1 and 2 metadata tags, it was found that for these smartphone models the D65 calibration matrix is in fact the other matrix. The dcraw software for conversion to XYZ space with smartphones should therefore be used with care. Experimental results presented in Chapter 3.2.3 demonstrate the impact of using the wrong colourmatrix.

ForwardMatrix approach

Also stored in the image metadata for dng files are what are called 'Forward matrices'. According to the Adobe dng specification the ForwardMatrix approach behaves better for more extreme values, so has been considered here as an alternative to the dcraw approach [92]. These forward matrices convert directly from phone native space to XYZ D50

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix}_{D50} = \underline{M}_{FM}\underline{D} \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{scene} \tag{2.16}$$

where $\underline{M}_{FM}$ is the forward matrix and $\underline{D}$ is the white balance matrix as before [92]. The metadata includes two forward matrices, taken at high and low colour temperatures. If the colour temperature of the illumination is known, then an interpolation may be carried out to determine the optimal combination of the two forward matrices. Here, where the details of the illumination are unknown, a simple mean of the two matrices was used.

2.6.3. Standard colour chart method

An alternative approach is to use a standard colour chart such as the Macbeth ColorChecker Classic, which has 24 patches covering a wide range of colours and neutral shades. The chart was developed primarily for photographers, and has inherent advantages here owing to its carefully produced colours and portability. The XYZ values for each patch are provided by the manufacturer, or can be measured

with a spectrophotometer, and by capturing an image of the colour chart using each phone under the flash illumination a corresponding set of RGB values is produced. It is then possible to obtain a device specific mapping, M, from native RGB to XYZ optimised for the smartphone illumination. Note that the smartphone flash is used since this is the resulting illumination after carrying out ambient subtraction, the proposed method to account for variations in ambient light discussed in Section 2.3.1. An overview of the whole process required to obtain a mapping from RGB to XYZ using the standard colour chart method is shown in Figure 2.5.

The images of the colour chart should be captured with no ambient light, and with a 45° angle between the phone and chart to minimise reflection. To develop the mapping, it is crucial to know the relative values of the patches. However, in most practical settings, and especially when using the phone's flash as the illumination, there is a high level of intensity non-uniformity, or shading, across the resulting image. An example figure demonstrating this effect is shown in Figure 2.6. The ideal imaging setup would result in a uniform illumination field (a), and yield the correct relative colour chart patch values (b). A more realistic imaging setup could result in a shading field as depicted in (c), with a ratio of the brightest to darkest point of 2.8. This results in colour chart patch values in (d) which have visibly incorrect values compared to the uniform setup.

The use of chromaticity, as discussed in Section 2.4, would remove the effect of shading. However in order to develop the mapping, we need to know the full RGB values and their relative values so this is not a viable option here. It is therefore necessary to correct for the spatial variation in the intensity of the illumination of the chart. To carry out this intensity non-uniformity correction (INUC), when using the standard colour chart method images of a neutral grey chart are captured along with the colour chart images. The neutral chart image allows the recovery of the shading field, as depicted in Figure 2.6 (c). The green channel values of these grey chart images are used to capture the intensity variation across the region of interest since the green channel has the largest signal values and hence the largest signal to noise ratio. Initially a simple white sheet of paper was used instead of a neutral grey chart, however it was found that the paper was not uniform enough and hence a dedicated neutral chart has been used. The following correction was applied

$$c_{corr}(\boldsymbol{x}) = \frac{c_i(\boldsymbol{x}) - c_d(\boldsymbol{x})}{n_{G,i}(\boldsymbol{x}) - n_{G,d}(\boldsymbol{x})} \tag{2.17}$$

where c and w represent the colour chart and grey chart images respectively, the superscript denotes the colour channels where $c \in \{R, G, B\}$, and the subscripts i,

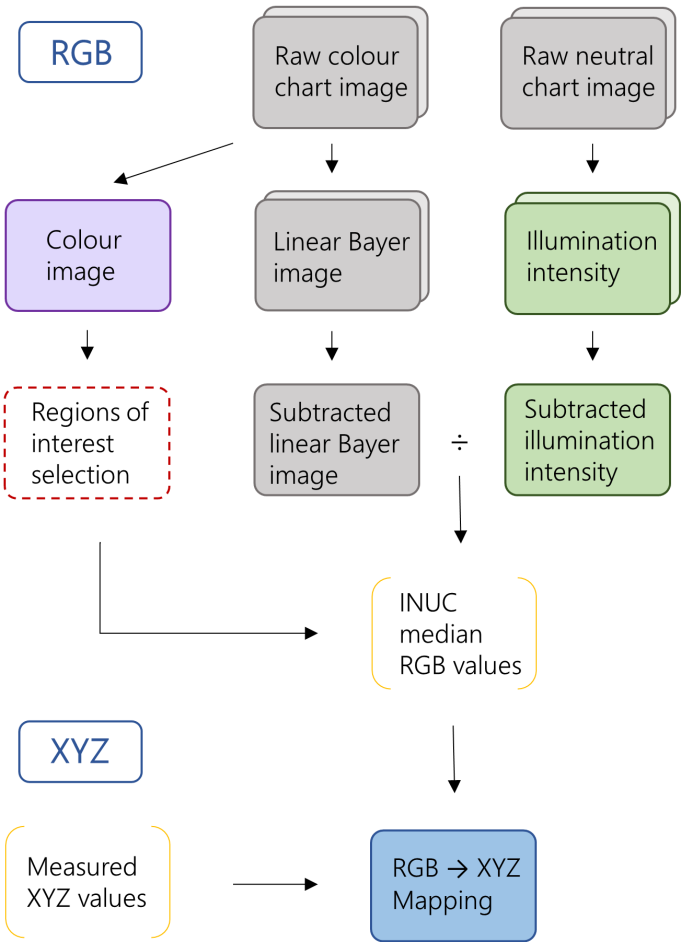


Figure 2.5: The overall pipeline to generate a device-specific mapping from RGB to XYZ using images of a colour chart and grey chart is laid out. RGB values: Two pairs of raw images are captured with the phone in the same position - one of the colour chart and one of a neutral chart. The raw images are converted to linear Bayer images and the illumination intensity is obtained by selecting the green channel. The images are then downsampled by a factor of four and a pixel-wise subtraction carried out. Finally the subtracted colour chart image is divided by the illumination intensity to carry out the INUC (intensity non-uniformity correction) and manually selected regions of interest from the 'sRGB' version of the colour chart flash image enable median RGB values for each patch to be found. XYZ values: the XYZ value for each patch is obtained through direct measurement using a spectrophotometer. The RGB and XYZ values are then used to determine the mapping.

d and *corr* refer to the illuminated, dark and corrected images respectively. A dark image correction is incorporated as well as the main intensity correction, using this simplified version of the form from [93]. Applying the INUC corrects for the shading field present, and produces patch values of the form shown in Figure 2.6 (b), rather than the example shaded values shown in (d). To reduce the impact of noise, both the colour chart and grey chart images are downsampled by a factor of four before

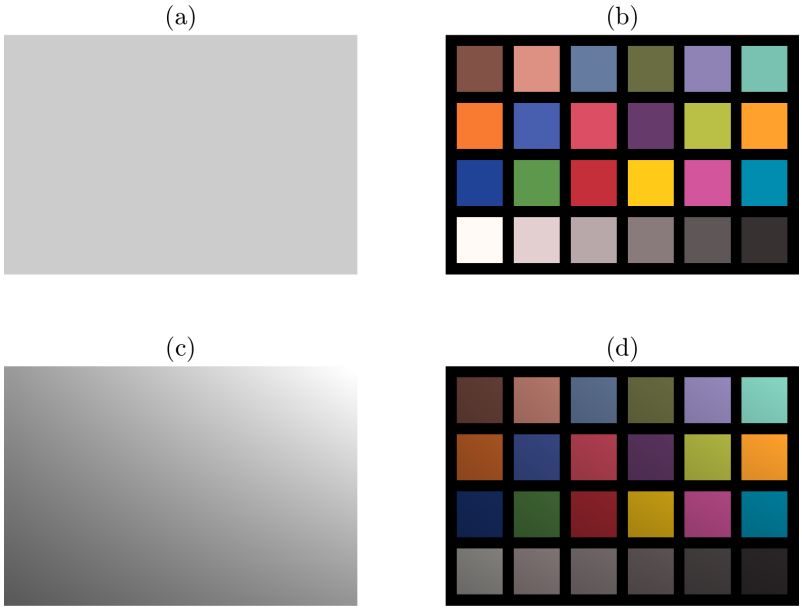


Figure 2.6: The impact of shading on the relative value of colour chart patches is demonstrated. A uniform and non-uniform field are contrasted in (a) and (c) respectively. The Classic colour chart is rendered under the uniform and non-uniform field in (b) and (d). The relative patch values have clearly been affected in (d), highlighting the need for a correction for variations in illumination intensity.

then applying the pixel by pixel correction. See Appendix B for a demonstration of the impact of applying the INUC on real data.

Finally, the corresponding XYZ values were measured directly from the colour chart using the X-Rite ColorMunki spectrophotometer to ensure that values used were precise. Four repeats of each colour patch were taken and the results averaged to give a final XYZ value for each patch.

Having obtained XYZ and corrected RGB values for each patch, the final step is to determine the conversion from RGB to XYZ. The different methods of conversion fall into roughly two groups - linear and polynomial regressions, and multidimensional look-up-tables (MLUTs) [94]. MLUTs rely on having a very large set of known pairs of values, and constructing mappings based on local neighbourhoods. They are only appropriate when the spectral approach is used, so have not been considered further here.

In the case of regressions, mapping matrices are determined such that XYZ values can be obtained from RGB values through a simple multiplication. The simplest

case is a linear regression with

$$[X\ Y\ Z] = [R\ G\ B]\, M_{lin} \tag{2.18}$$

where M_{lin} is a 3×3 matrix. This simple approach has the advantage of being exposure time invariant [95]. Exposure invariance means that if we have an RGB value $\boldsymbol{r}$ which has a corresponding XYZ value $\boldsymbol{h}$, and then we change the exposure time by a factor of k we would obtain an adjusted RGB value of $k\boldsymbol{r}$ which is in turn mapped to an XYZ value of $k\boldsymbol{h}$. Crucially, there is no colour shift for RGB values taken using a different exposure time. This is a particularly important quality for the selected mapping method since the same calibration mapping is applied to results from many different images, taken using different exposure settings.

The downside of a linear regression is that they can yield quite high errors overall [95]. Polynomial regressions reduce errors by expanding including higher order polynomial terms in RGB. For example, for second order

$$[X\ Y\ Z] = \left[R\ G\ B\ R^2\ G^2\ B^2\ RG\ GB\ RB\right] M_{poly} \tag{2.19}$$

where M_{poly} is expanded to a 9×3 matrix. Unfortunately, inspection of Equation 2.19 reveals that polynomial regressions are not exposure invariant due to the higher power and cross terms. They are therefore not appropriate for use here. An alternative approach which enables the use of higher order terms whilst maintaining exposure invariance is a root-polynomial regression [95]. Again, for second order

$$[X\ Y\ Z] = \left[R\ G\ B\ \sqrt{RG}\ \sqrt{GB}\ \sqrt{RB}\right] M_{root} \tag{2.20}$$

where M_{root} becomes a 6×3 matrix. Of the options considered, both linear and root-polynomial regressions are appropriate to our application.

The implementation to determine M is effectively the same in both cases so will be presented for both at once, using labelling conventions from [94]. First, matrices of the matching sets of RGB and XYZ values are constructed. For both linear and root-polynomial regressions, the XYZ matrix H has shape N×3, where N is the number of pairs of values. For linear regressions, the RGB matrix R has shape N×3 and has the larger shape N×6 for the root-polynomial approach. In each case, the aim is to find the optimal mapping M such that $H \approx RM$. This is achieved by minimising

$$\mathfrak{F} = \|H - RM\|^2 \tag{2.21}$$

The Moore-Penrose pseudo-inverse expression for M is obtained by expanding this expression and taking partial derivatives, yielding

$$M = \left(R^T R\right)^{-1} R^T H \tag{2.22}$$

This closed-form solution can be very easily implemented. Practical testing of the standard colour chart method is presented in Chapter 3.2.

2.6.4. Shading independent colour chart method

The standard colour chart approach has the obvious advantage of capturing the intensity variation across the colour chart as close to directly as possible. However, the approach also has several downsides. The first is the need for more equipment - rather than just needing the colour chart, a grey chart and phone tripod are also required. When trying to develop a cheap, portable method this is an important consideration. The second is the requirement of obtaining images of the colour chart and grey chart with as little movement as possible, which significantly complicates the calibration image capture process. The process is quite straightforward with some experience, but is very prone to user error for example by knocking the tripod between captures. When carried out by less experienced users the resulting data is often of poor quality.

Several papers have been published proposing methods to obtain a colour chart mapping avoiding the need for grey chart image capture, by indirectly accounting for the fact that there may be intensity variation across the chart. The first papers published proposed developing a mapping by minimising a quantity which is independent of the relative magnitudes of the RGB values [96, 97]. In this way, the issue of the intensity variation can be side-stepped. However, initial testing of the approaches yielded very poor results likely due to the reliance on a minimisation search which is highly dependent on the initialisation. Later work by Finlayson et al kept the goal of avoiding a grey chart capture but tackled the intensity variation more directly [98]. They proposed an elegant solution of estimating both the shading across the colour chart and the mapping using an alternating least squares approach.

Two algorithms are presented in the paper, which will both be discussed mathematically and a visual description presented afterwards to aid understanding. When considering the mapping from RGB to XYZ, with intensity corrected data the goal was to find M such that $H \approx RM$. Here, the intensity variation can be brought directly into the colour correction equation as $H \approx DRM$, where D is an N×N diagonal matrix containing shading correction terms for each patch of the colour

chart. Unlike the previous form, there is no closed-form solution hence the use of an alternating least squares (ALS) approach. With ALS, the estimate for M is updated keeping D fixed, then D is updated keeping M fixed. This process is iterated, with parameters improving until convergence at a solution. This first algorithm can be described mathematically in just a few steps, as laid out in [98], using notation consistent with this work. Here, R^0 is the initial matrix R, and R^k is R after iteration k.

1. Update D: The diagonal elements of D are given by

$$d_{jj} = \frac{r_j \cdot h_j^T}{\|r_j\|^2} \tag{2.23}$$

   with j ranging from 1 to N, and r_j and h_j representing the jth row of R^{k-1} and H. The authors note that in a least-squares sense D is therefore the optimal diagonal transform between R^{k-1} and H.

2. Update M:

$$M = (DR^0)^+ H \tag{2.24}$$

   where the + denotes the Moore Penrose pseudo-inverse, i.e. $A^+ = (A^T A)^{-1} A^T$

3. Update R:

$$R^k = R^0 M \tag{2.25}$$

4. Repeat steps 1-3 until convergence.

This first algorithm is very simple and lightweight to implement, and converges within a few tens of iterations. However, the algorithm estimates shading on a per-patch basis. In reality, the shading across the colour chart will be smooth, with shading factors for adjacent patches being similar. Finlayson et al therefore propose a second ALS algorithm which models the shading across the whole image using a weighted sum of shading basis functions, thus ensuring smoothness. The authors use Discrete Cosine Transform (DCT) functions as the bases [99]. These increase in complexity as the order increases, but as an example the first three bases are constant, increasing left-right and increasing top-bottom. In the following mathematical description, P represents the colour chart image, J is the per-pixel shading correction, and as above H and R represent the XYZ and average RGB values for each patch. Superscript indexes denote the iteration number, so for example P^0 is the initial colour chart image P. The shading field and mapping M are determined as follows

1. Update J:

$$J = \sum_{g=1}^{M} w_g G_g \tag{2.26}$$

where M is the number of DCT basis functions used, G_g represents the gth basis function, and the ws are the scalar weights. The weights are updated using a least squares regression $w = E^{+}u$, where as for the previous algorithm the + signifies a Moore-Penrose pseudo-inverse. Here, u is a $3N \times 1$ vector of the XYZ values stacked on top of each other - $[X_1 \dots X_N, Y_1 \dots Y_N, Z_1 \dots Z_N]^T$, and E is a $3N \times M$ matrix based on RGB values and DCT basis functions. This matrix is formed in several steps. Firstly, for each colour channel i, M images are formed by pixel-wise multiplication of the colour chart image slice with the basis functions, i.e. $P_i * G_1, \dots, P_i * G_M$, where $*$ represents pixel-wise multiplication. Second, the average pixel values of these images for the N patches are determined and placed into three $N \times M$ matrices E_i. Finally, the $3N \times M$ matrix E is formed by stacking the matrices for each colour channel - $[E_{\text{red}}; E_{\text{green}}; E_{\text{blue}}]$.

2. Update M:

   (a) $P^k = P^0 * J$, where as above $*$ represents a pixel-wise multiplication, i.e. apply the shading correction to the image.

   (b) Determine R^k - average for each colour patch in P^k

   (c) $M = (R^k)^{+} H$

3. Update P:

$$P_j^k = P_j^0 M \tag{2.27}$$

where j represents each pixel, i.e. apply the mapping M to every pixel in the original image.

Note that this description of the second ALS algorithm is from [98], but corrects several crucial typos found in the paper and uses notation to match this work. The second algorithm also converges in tens of iterations for up to 21 DCT bases.

A more visual understanding of the differences between the two algorithms is shown in Figure 2.7. The top row shows the input non-uniform field and resulting non-uniform image (as in Figure 2.6). The second two rows show the shading corrections and corrected images obtained using the first and second approaches respectively. The first approach yields independent, per-patch shading correction factors whereas the second yields a smooth, per-pixel shading correction field. In this example case both approaches yield good correction estimates, producing shading

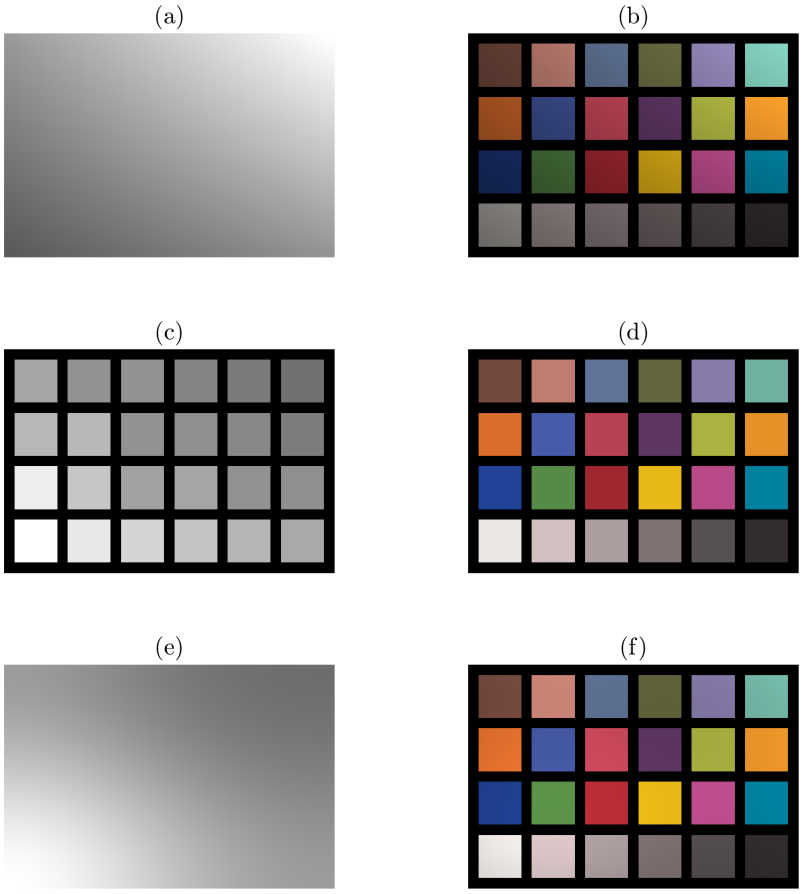


Figure 2.7: The same non-uniform field as in Figure 2.6 is shown in (a) along with the Classic colour chart rendered under this field in (b). The shading correction factors produced by the first ALS approach are shown in (c), where there is a single correction per patch which is independent of all other patches. The shading correction field produced by the second ALS approach is shown in (e). Rather than discrete, per-patch correction factors the second method produces a smooth shading correction field across the whole image. The shading corrected images for approach one and two are shown in (d) and (f) respectively. Note the uniformity of the resulting images.

corrected images which are visually very similar and which both virtually eliminate the intensity non-uniformity. Note that both approaches are compatible with linear, root-polynomial, or polynomial mappings. However, since we are using a colour chart with a low number of patches, linear mappings are most appropriate with ALS approaches here.

An overview of the calibration process when using either of the two ALS methods is shown in Figure 2.8. The crucial difference compared to the standard colour chart calibration process depicted in Figure 2.5 is that the aligned images of a grey chart are no longer required. This means that the phone can be handheld when collecting images of the colour chart, making the image capture process simpler and

less prone to user error. The task of determining the shading across the colour chart and thereby the true mapping has instead been moved to the processing stage. A comparison of the two ALS approaches and the standard colour chart method is presented in Chapter 3.2.2.

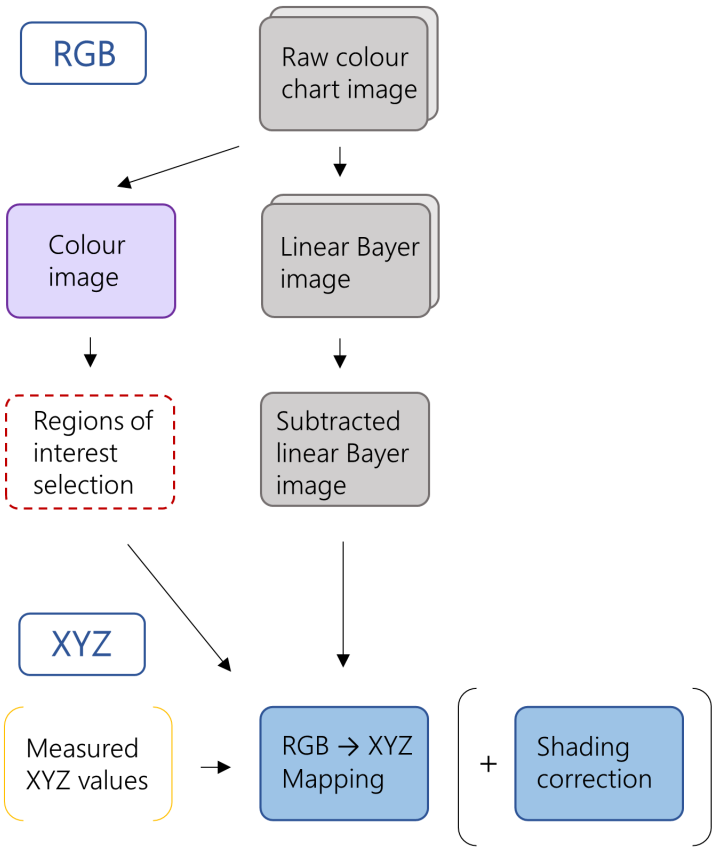


Figure 2.8: The pipeline to generate a shading-independent device-specific mapping from RGB to XYZ using images of a colour chart is laid out. A pair of flash/no-flash raw images are captured of the colour chart in a dark environment, and regions of interest are selected from the 'sRGB' flash image. Measured XYZ values for the colour chart patches are used as inputs to either ALS algorithm along with the colour chart image and regions of interest. The algorithm outputs the mapping from RGB to XYZ, using the shading correction which it also determines.

3. Obtaining quantified colour values

The components of our approach to overcome the challenges of quantifying colour are tested experimentally, before an overview of the calibration and data collection procedures, and a summary of the overall approach. Parts of the work presented in this chapter are published in a PLOS ONE article [65] and in the Color and Imaging Conference 27 and 28 proceedings [66,67]. These papers, licensed under the creative commons license CC BY 4.0, have been modified to form part of this thesis.

3.1. Ambient subtraction

As discussed in Chapter 2.3.1, the subtraction of data from flash/ no-flash image pairs enables the impact of ambient light to be minimised. This enables data from a given phone captured in different lighting conditions to be compared. Here, experimental demonstration data for the technique and a metric for determining the suitability of image pairs for use with the technique are presented.

3.1.1. Colour chart ambient subtraction demonstration

The ambient subtraction method was tested for a wide range of colours by imaging 172 patches of the Macbeth ColorChecker DC chart (excluding the repeating neutrals from the boundary of the chart and the reflective patches). An image of the DC chart is included in Figure 3.2 (a). The DC chart was imaged under daylight and fluorescent lighting, as well as with no ambient lighting to provide a ground truth. As discussed in Chapter 2.4, to discount the effect of shading and varying exposure time chromaticity values were used. The rg chromaticity values for each patch were calculated before and after subtraction, and the distance to the corresponding ground truth rg values (GT) was calculated. The distance was defined as the Euclidean distance

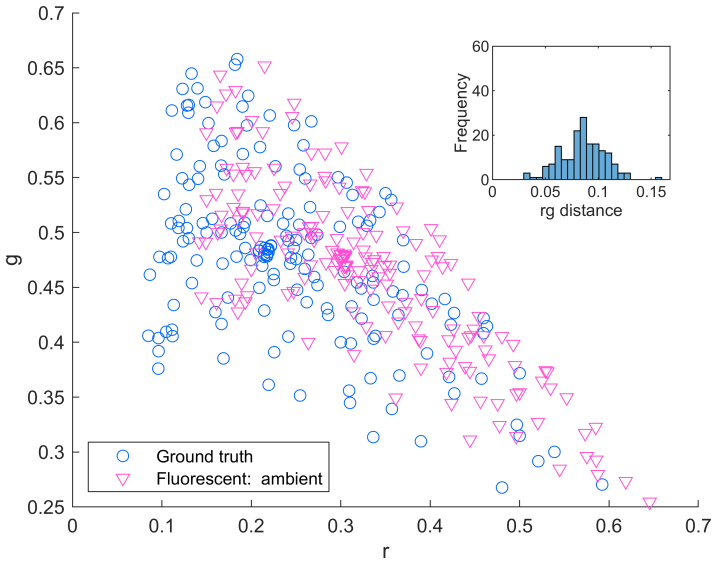
$$\text{rg distance} = \sqrt{(r_{test} - r_{GT})^2 + (g_{test} - g_{GT})^2} \tag{3.1}$$

where r and g are red and green chromaticities for the patches, and the $test$ and GT subscripts refer to data under ambient light (before or after subtraction) and no ambient light respectively.

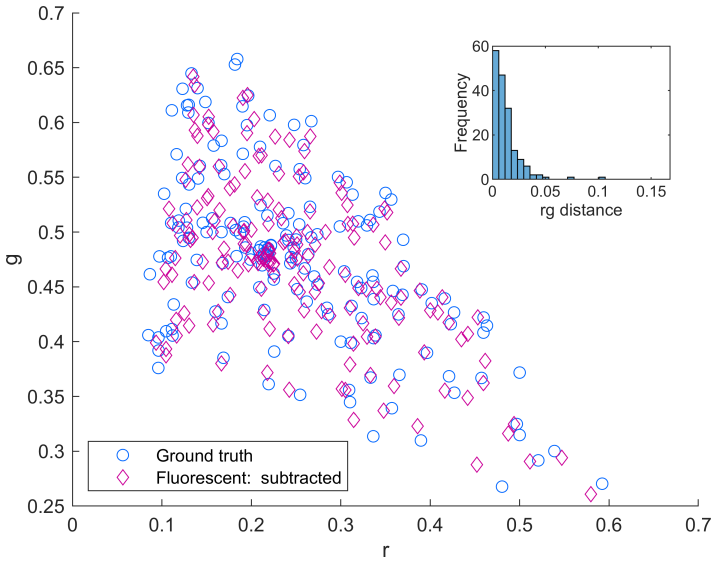
Data was collected using the two models of phones used in the study. These are the Samsung S8, referred to simply as the S8, and the LG Nexus 5X, referred to as the

Nexus. The S8 uses the rear-facing camera in combination with the flash, and the Nexus uses the forward-facing camera with the screen as the flash. Further details on their use in clinical data collection are given in Chapter 4.1. Figure 3.1 shows the DC chart results before and after subtraction for a Nexus phone under fluorescent lighting - results for all phones were similar in form, and fluorescent lighting results have been presented here as an example. From Figure 3.1 it is clear to see that the patch values after subtraction move towards the ground truth values, as the subtraction removes the impact of the ambient light on the pixel values. The average rg distance between corresponding pairs of ground truth and ambient light influenced values decreases significantly after subtraction, as demonstrated by the histograms inset in Figure 3.1.

To aid a clearer visualisation of the impact of ambient subtraction, a subset of 10 points from the DC chart is shown in Figure 3.2. The outer bounds of the total group of ground truth patch values from Figure 3.1 is shown with a blue dashed line for reference. The subset of patches was chosen to be representative of the whole spread of points, with each ground truth value shown as a filled coloured circle corresponding to the actual colour of the patch. The values before and after subtraction are shown connected by an arrow. With the smaller set of points it is now even easier to see how much the ambient subtraction helps to standardise the data. The data presented here is again for fluorescent lighting, but similar results were found for daylight.

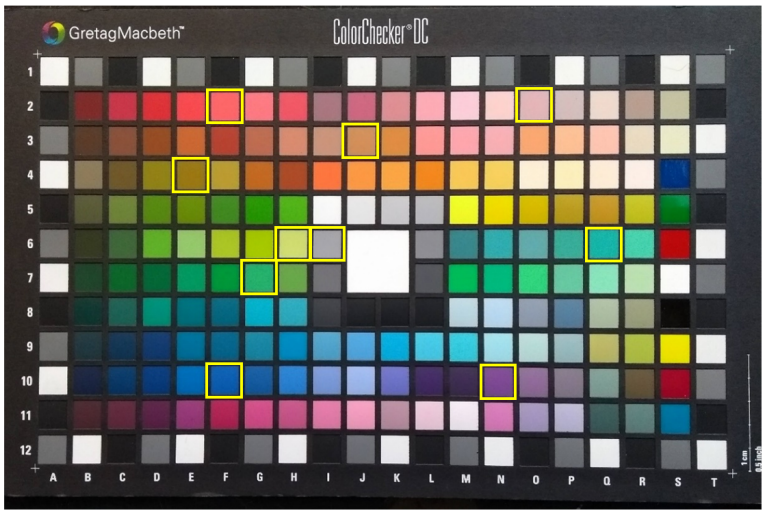


(a)

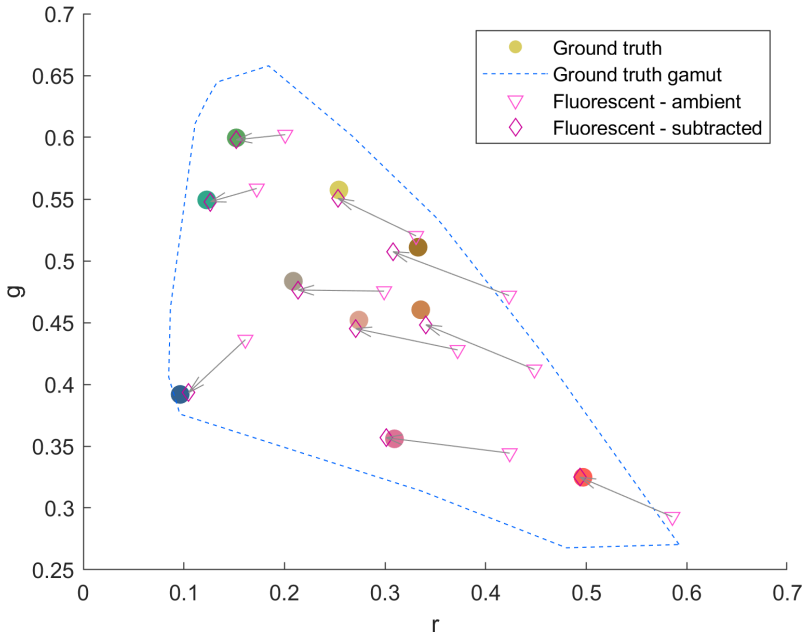


(b)

Figure 3.1: Macbeth ColorChecker DC chart rg chromaticity patch values for ambient fluorescent lighting are shown for an example phone before (a) and after (b) subtraction, as pale pink triangles and dark pink diamonds respectively. The ground truth rg values are denoted by blue circles, and the histogram of rg distances from the ground truth is shown as an inset in each subfigure. Note the sevenfold decrease in the average rg distance to the ground truth as the ambient subtraction minimises the effect of ambient light.



(a)



(b)

Figure 3.2: A subset of the example phone DC chart patches shown in Figure 3.1 are shown to enable a clearer visualisation of the impact of ambient subtraction. An image of the DC chart is shown in (a) with the selected patches outlined in yellow. The impact of ambient subtraction on these patches is shown in (b). The outer limit of the ground truth rg values is shown with a blue dashed line and the subset of points have been selected to cover the gamut. The ground truth values are shown using large filled coloured circles, where the circle colour is given by the ground truth colour of the patch. The values before and after subtraction are denoted by pale pink triangles and dark pink diamonds respectively, as before, and corresponding points are joined by an arrow. Note how in all cases the subtracted points move into close proximity of the ground truth value.

3.1.2. Subtracted Signal to Noise Ratio (SSNR)

In order to avoid time-consuming recapturing of data, or loss of data, a metric to give an indication of whether the images captured are suitable is required at the time of capture. For the ambient subtraction method to yield good results, the flash must dominate over the ambient light. A simple intensity ratio of the flash to no-flash image appears to be a good option, however this does not take into account additional noise introduced if the overall signals are small. For a pixel value in the midrange of the sensor, as is typical when auto-exposure is used, shot noise dominates which can be described by a Poisson distribution [100]. In this case, the noise is simply given by the square root of the signal.

We propose the Subtracted Signal to Noise Ratio (SSNR) as a suitable metric, given by the signal to noise ratio of the flash only pixel values obtained after subtraction

$$\text{SSNR} = \frac{I^F}{noise(I^F)} = \frac{I^{F+A} - I^A}{\sqrt{I^{F+A} + I^A}} \tag{3.2}$$

where I^{F+A} and I^A are the intensity values recorded in the flash and no-flash images respectively, and the positive sign in the denominator is due to the summation of errors in quadrature [101]. To avoid introducing error from motion between images, this calculation is not performed pixel-wise but instead the average signal for the region of interest is calculated for flash and no-flash images, and a global SSNR calculated. Note that demosaiced images should be used to avoid biasing the results towards the green channel. We use the most simple demosaicing option, where the overall image size is halved and each kernel becomes one pixel. The definition of the SSNR makes simplifying assumptions about sources of noise so that the calculation can be based simply on the pixel values and no other information is needed.

3.1.3. SSNR threshold

Since there is not an intuitive value for the SSNR above which the images will be reliably useful, the following experiment was carried out to get a gauge of what the lower limit SSNR cutoff for ambient light-independent colorimetric measurements should be. The phone was held static at a 45° angle to a Macbeth ColorChecker Classic chart, which has 24 patches, and flash/ no-flash image pairs were captured. A TaoTronics TT-DL09 LED desk lamp was used to provide a controlled ambient light with a correlated colour temperature of 3850K, and image pairs were captured as the intensity was gradually increased. For each image pair the rg chromaticity values for each patch after subtraction were calculated and the distance to the corresponding ground truth rg values, from a no ambient light image set, was cal-

culated according to Equation 3.1. The smallest rg distance achievable in practice was estimated by taking a series of no ambient light images of the Classic chart over different capture sessions, and the average rg distance between the same patches imaged multiple times was calculated. This average rg distance was then used as a baseline for each phone to determine the suggested SSNR cutoff for practical use.

Figure 3.3 shows the rg distance for each subtracted patch value from its corresponding ground truth value as a function of the SSNR for that patch, for an S8 phone. The experimentally determined baseline rg distance is shown in Figure 3.3 along with the value plus one standard deviation. For a practical limit, we deem that once all points are within this higher threshold they will not be limited by the SSNR. The inset of Figure 3.3 shows the ground truth rg value for an example patch along with some subtracted results with varying SSNR. The practical rg distance threshold is also marked in the inset, and it can be seen that once a certain SSNR is reached the results remain within the threshold. For each phone used in our research, the threshold SSNR was calculated using the baselines specific to that phone. The SSNR values yielded were similar for each phone and an average over the four phones results in a threshold SSNR of 3.4, marked in Figure 3.3. When capturing data using ambient subtraction, images should be retaken until the region of interest has an SSNR above this threshold.

3.1.4. Sclera ambient subtraction demonstration

So far, ambient subtraction has been demonstrated for a wide variety of colours using the DC colour chart, however the overall aim is for subtraction to be incorporated into the system for jaundice monitoring via human images. At this stage, therefore, the subtraction technique was tested for the target scenario of sclera imaging. Test data was collected for five healthy volunteers under five ambient lighting conditions, using a single Samsung Galaxy S8 phone. For each lighting condition, three repeat flash/ no-flash image pairs were collected of the same portion of sclera. For each image pair, the SSNR was checked for the sclera region and if the value fell below the threshold defined in Section 3.1.3 then the images were re-captured. Along with these images, the spectrum of the ambient light was also measured to compare the different lighting conditions. A white tile was held near to the sclera, and a SpectraScan PR-650 spectoradiometer was used to measure the spectrum of light at the tile. Finally, images of the sclera were captured under no ambient light (flash only) to provide a ground truth sclera colour value for post-subtraction comparison. The image capture process under ambient LED lighting is shown in Figure 3.4. Permission from the subjects were obtained to use these images.

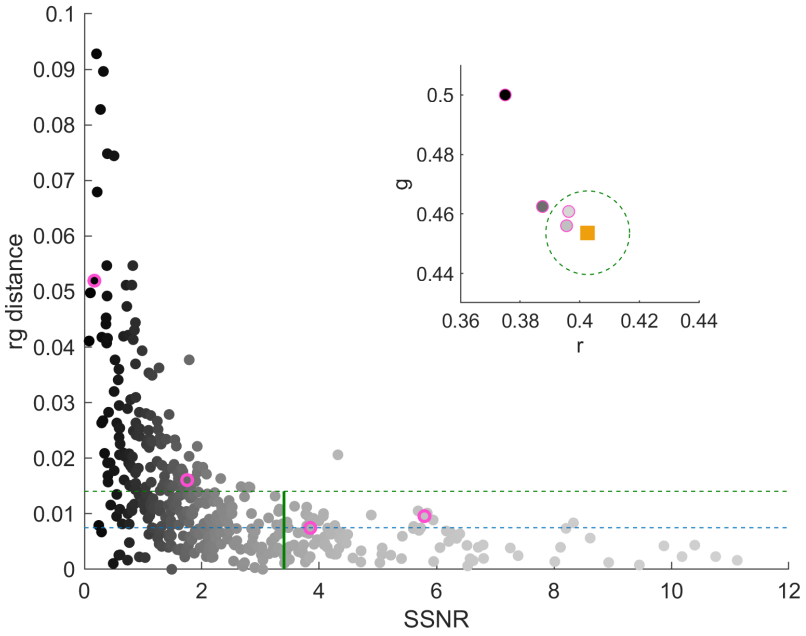


Figure 3.3: The rg distance of each Classic patch from the ground truth, found using images under no ambient light, is plotted as a function of the SSNR value for that patch as the ambient light level was varied. The colour of the points varies from black to pale grey according to increasing SSNR. The baseline rg distance and the baseline plus one standard deviation are shown with dashed lines in blue and green respectively. The required SSNR for useable data is defined as the point at which all points are below the upper line, and therefore not limited in accuracy by SSNR. The overall experimentally determined SSNR threshold across phones is 3.4, shown as a solid vertical green line. Additionally, an inset shows a subset of data points for an example patch. The square pale orange point represents the ground truth rg value for the selected patch, where the colour of the square is given by the ground truth colour of the patch. The green dashed line shows the baseline rg distance determined for adequate SSNR. Finally, the greyscale circles outlined in pink (as also outlined in pink in the main figure) show the initial large impact of increasing SSNR on the rg distance until the baseline where the results are comparable even for increased SSNR.

After capture, the scleral region of interest was manually segmented in each image and the median RGB values before and after subtraction were extracted. As with the colour chart demonstration, rg values were used to compare the values before and after subtraction. The results are shown in Figure 3.5 (a), and the associated ambient spectra are in (b). It is clear from Figure 3.5 (a) that there is a sizeable variation in recorded sclera values under the different ambient illuminations. The ground truth value for the sclera is marked with a black square, and was determined by capturing images purely under the phone's illumination in a dark room. The

(a) (b)

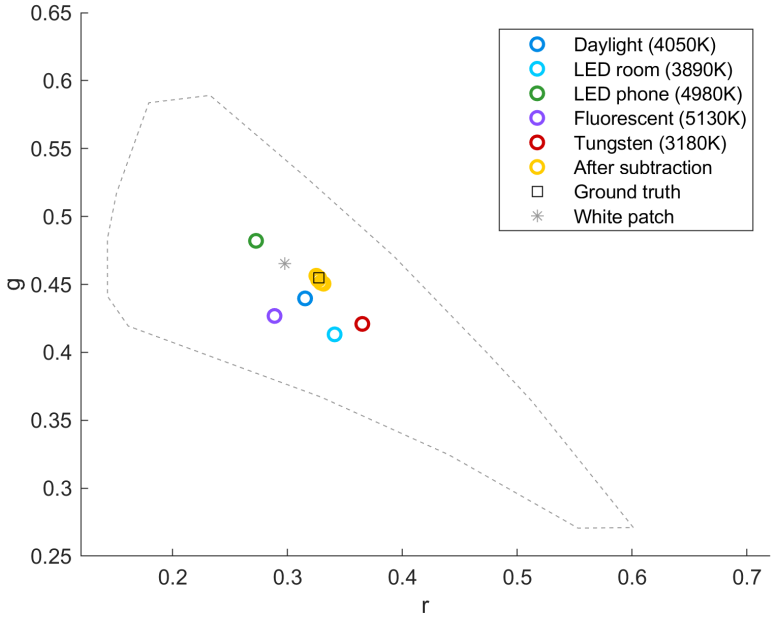
Figure 3.4: (a) Image capture for the subtraction demonstration is shown, with permission from the subject, under ambient LED lighting. The subject's eye is positioned in the centre of the field of view, at a distance of a few inches. Specular reflection coming from the phone flash is carefully positioned over the iris. An example pair of flash/ no-flash images are shown in (b), cropped to the eye region.

data for all ambient lighting conditions after the ambient subtraction has been performed is shown in yellow. The variability has been hugely reduced, and the values are closely clustered around the no ambient light ground truth value.

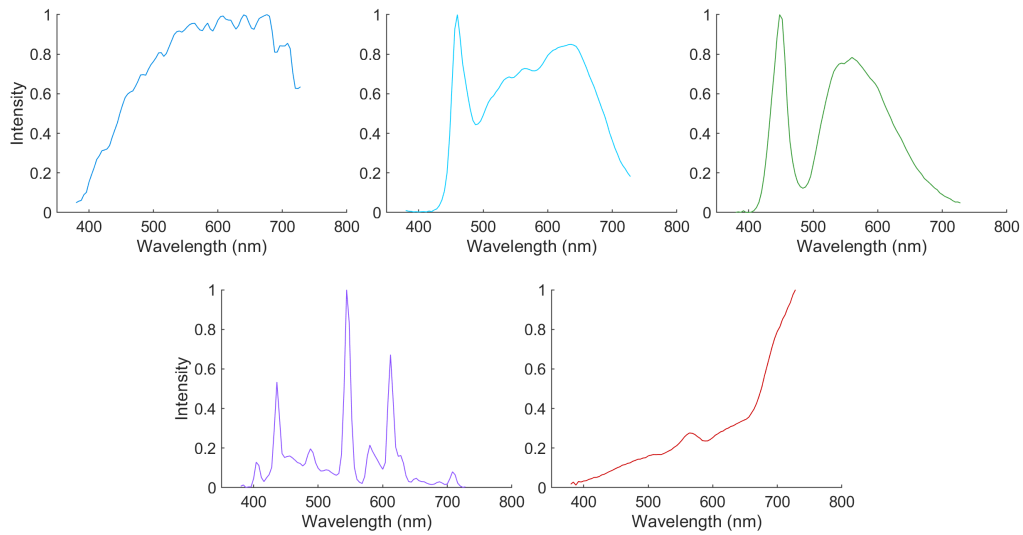
Similar trends were observed for the five participants. The average rg distance data across all participants for the five different lighting conditions is shown in Table 3.1 before and after subtraction. The data in Table 3.1 confirms what is visually shown in Figure 3.5: that carrying out ambient subtraction greatly reduces the rg distance from the ground truth, to less than a tenth of the uncorrected distance for all lighting conditions. This confirms the ability of the subtraction technique to provide stable results under varying room lighting conditions, even for handheld capture of human subjects.

Lighting	rg distance pre subtraction		rg distance post subtraction	
	Mean	SD	Mean	SD
Daylight	0.055	0.023	**0.003**	**0.003**
LED room	0.037	0.008	**0.004**	**0.002**
LED phone	0.057	0.008	**0.003**	**0.002**
Fluorescent	0.051	0.006	**0.002**	**0.002**
Tungsten	0.058	0.012	**0.002**	**0.002**

Table 3.1: Average rg distances for sclera image values under different ambient lighting conditions to the no ambient light ground truth value before and after ambient subtraction. The mean and standard deviation (SD) for the data is presented across the five healthy volunteer subjects.



(a)



(b)

Figure 3.5: (a) The recorded colour values for a healthy adult sclera are shown using an rg chromaticity diagram before and after ambient subtraction. The ambient recorded data is shown for daylight (blue), LED room lighting (turquoise), LED phone lighting (green), fluorescent (purple) and filtered tungsten (red). The correlated colour temperature of each ambient lighting condition is provided in the legend, obtained from the spectral power distribution of the ambient light. After subtraction, the data is shown in yellow, along with the ground truth sclera colour obtained from images captured with no ambient light (black square). The outer limits of the Macbeth DC chart patch values as imaged by this phone are shown for context (grey dashed line), along with the DC chart white patch value (grey star). (b) The spectrum of each ambient lighting condition is shown with colours as in (a).

3.2. Mapping testing

In order to compare data from different phones, a mapping from phone native space to a device-independent space must be carried out, as discussed in Chapter 2.5.1. Here, different methods for developing a mapping from pairs of known RGB and XYZ values are compared, before a larger scale comparison of mapping approaches based around metadata information and a colour chart. Finally, the potential for a custom colour chart is explored.

3.2.1. Standard linear vs root polynomial mapping

As discussed in Chapter 2.6.3, two viable options for developing the mapping from native RGB to XYZ space for the standard colour chart method are linear and root polynomial mappings. These two approaches are viable since they are independent of the exposure time, with root polynomial including some cross terms. A leave one out method was used to compare the quality of the mapping resulting from each approach. With this method, the mapping is developed excluding one patch of the colour chart at a time, and then testing on this final patch by applying the mapping and comparing the result to the known expected value. This process is repeated for each patch, and the resulting errors averaged. Matching with convention, the errors are presented in L*a*b* space using the ΔE metric

$$\Delta E^*_{ab} = \sqrt{(L^*_2 - L^*_1)^2 + (a^*_2 - a^*_1)^2 + (b^*_2 - b^*_1)^2} \tag{3.3}$$

Using this formulation, an error above around 2 means that a human is able to visually detect the difference between those two colours [95].

Table 3.2 presents the mean, median and 95th percentile ΔE errors for linear and root polynomial mappings for a representative Nexus phone based on images of a ColorChecker Classic chart. The errors for the linear mapping are slightly higher than seen in the literature, although in the same region [95]. However, the root polynomial mapping does not provide the expected decrease in ΔE errors across the three metrics, with a more pronounced effect for the third order version. This is likely due to the comparatively low number of patches on the colour chart, which may be resulting in overfitting. Using a chart with more patches is not feasible in this context - reducing the patch size would maintain the chart physical dimensions and so maintain portability, but it would increase the noise. To maintain generalisability, the simpler linear mapping was therefore deemed more reliable.

Nexus	Mean	Median	95th percentile
Linear	2.8	2.5	5.4
Root polynomial order 2	2.6	2.2	7.6
Root polynomial order 3	3.7	2.4	11.8

Table 3.2: ΔE errors in L*a*b* space between expected XYZ values and those produced by linear and root polynomial mappings from an example Nexus phone native space. Mean, median and 95th percentile averages are presented from a leave one out analysis.

3.2.2. Standard vs ALS colour chart mapping

An alternative to using the standard colour chart approach, which involves a paired grey chart image, is to use a shading independent method. Here we compare two alternating least squares (ALS) methods against the standard method. The mathematical details were presented in Chapter 2.6.4, but the relevant details of both methods are summarised here. The ALS methods work by alternating between determining the mapping and a shading correction until convergence. The first method is based on average patch values and so produces shading estimates on a per-patch basis which are then fed into the mapping calculation. The second method is based on using smoothly varying basis functions to model the shading and so produces a per-pixel shading correction which is applied before calculating averages and calculating the mapping. As in the original paper, we use discrete cosine transform (DCT) functions as bases [98].

As with the comparison of linear and root polynomial mappings in the previous section, images of a ColorChecker Classic chart captured under the smartphone illumination were used. The mappings were generated using the original (non-uniform intensity illuminated) images, and the ΔE errors calculated based on the shading corrected image data. Table 3.3 presents results for the standard linear method and the two ALS methods. As expected, when the images are corrected using the paired grey card images before developing the mapping, and then tested on the corrected data, the errors are lowest. When no attempt is made to correct the shading across the colour chart, presented in Table 3.3 as Linear, the errors when the mapping is tested on intensity-corrected data are extremely high. This is expected since the level of shading across the colour chart images is significant, with maximum to minimum intensity ratios of 2.3 - 3.1 for the phones tested.

The errors for all the ALS data presented are lower than when shading is ignored, providing reassurance that the methods were implemented correctly. For the DCT

Nexus	Mean	Median	95th percentile
Linear - intensity corrected	2.4	2.3	4.6
Linear	10.2	10.0	16.9
ALS - per patch	**3.3**	**3.2**	**5.9**
ALS - DCT basis			
3 basis	6.5	5.9	12.0
6 basis	4.1	4.0	7.4
10 basis	4.3	4.2	7.5

Table 3.3: ΔE errors in L*a*b* space between expected XYZ values and those produced by linear and alternating least squares mappings from an example Nexus phone native space. Mean, median and 95th percentile averages are presented based on testing mappings on intensity corrected images. The first row contains results based on generating the mapping using intensity corrected data, all other mappings were generated using uncorrected image data.

basis method, use of 6 basis functions provided the lowest errors. Increasing the number of basis functions to 10 and beyond increased the errors again due to over-fitting. The DCT basis method is more computationally complex since calculations are based on a whole region of pixels per patch rather than a single average value, hence it is slower to run than the per-patch method. Constraining the shading estimate to be smoothly varying, since adjacent patches should have related intensities, makes logical sense, but in the real world there can be shading fields which are not well modelled by this approach. For example, the edge of a phone can cause a sharp edge in shading which leads to adjacent patches having distinctly different intensities.

Here, amongst the ALS results, the mathematically and computationally simpler per-patch ALS method provided the lowest errors. These are still higher than when using grey chart images to correct for shading, however the image capture process is significantly simplified. The standard method requires good alignment between colour and grey chart images, therefore meaning that the phone must be held in a tripod for image capture. Misalignment of the paired images will affect the quality of the mapping produced by the standard method, since the shading will not be fully corrected. The ALS approaches, on the other hand, do not require paired images and so can be captured using a hand-held device. The simplification of image capture means that a lower skill level is required to obtain good quality calibration images. Therefore, the per-patch ALS method is included for further testing in the following section to assess whether the higher ΔE errors presented in Table 3.3 have a practical impact.

3.2.3. Comparison of mapping methods

To compare how well the different colour chart and metadata mapping methods perform, a controlled experiment was carried out. To investigate the variability of the phones used in our research within a specific model, two devices of each model were used. The Macbeth ColorChecker DC chart was imaged under no ambient light, as for the ambient subtraction demonstration in Section 3.1.1, to isolate the mapping stage. The repeating neutrals, reflective patches and those outside the gamut of the Classic chart in xy space were excluded from analysis. The resulting ground truth values for the Classic chart and DC chart are shown in Figure 3.6, measured using the X-Rite ColorMunki spectrophotometer. The different mapping approaches were applied to the same set of DC chart images, and the results compared to the known ground truth values.

As for the ambient subtraction demonstration, the relative brightness information is not retained. This means that effectively the full XYZ values are not yielded, but instead just the chromaticity information. A conversion to L*a*b* space, as would be standard, is therefore not possible. The xy chromaticities could be easily converted to u′v′ values, from the CIE 1976 uniform chromaticity scale diagram which is designed to be more perceptually uniform [102]. However, this was deemed unnecessary since perceptual uniformity is not a priority. Instead, xy chromaticity space

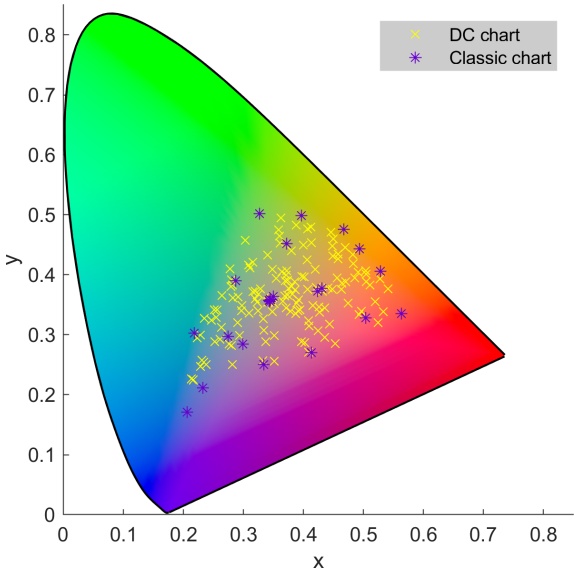


Figure 3.6: The ground truth xy values for the Macbeth DC chart used for testing (yellow crosses) are shown along with the ground truth xy values for the Macbeth Colorchecker Classic chart used for the colour chart approach (purple stars).

was used, in alignment with previous literature in this area [46,47], and similarly to Equation 3.1 the xy distance to the ground truth used as a performance measure.

It is important to be able to tell how well the different mapping approaches perform across all the phones considered here. A procedure to calculate the Classification Accuracy for Multiple Phones (CAMP_n, where n is the number of phones considered) was therefore defined in order to compare the three metadata methods and two colour chart methods across different phones. All mapping approaches were applied to the data and different subsets of the DC chart patches were selected with varying allowed minimum separations in xy chromaticity space. The mapped values for each phone and patch were classified to the ground truth values using a nearest neighbour classification - the classification was deemed successful only if all four phones correctly classified a patch. In other words, to calculate the CAMP_4 data presented next, these steps were followed:

- For a given minimum allowed xy separation
  - Select a subset of DC chart points which are all at least this far apart
  - For each phone (1 to n)
    - Do a nearest neighbour classification of the mapped data to ground truth data using the selected subset of points
  - CAMP_n = percentage of points for which all phones gave the correct classification
  - Repeat for a minimum of 1000 unique point set permutations and find the average CAMP_n

The CAMP_4 accuracies for the five methods are shown in Figure 3.7 for a range of average minimum xy separations. From an inspection of Figure 3.7, it is clear that for very high xy distance thresholds, or in other words for discriminating between very different colours, it does not matter which mapping is chosen as all give good results. The dcraw mapping gives the worst results but as discussed in Chapter 2.6.2 this is due to a bug in the current version of dcraw. When the corrected version is used, the results improve. The corrected ColorMatrix and ForwardMatrix approaches provide similar results to each other, with a 90% accuracy achieved for xy separations around 0.07 - 0.08. However, the two colour chart approaches reach a 90% accuracy for xy separations around 0.04 - 0.05. This means that xy values which are around 70% as far apart can be correctly classified using the colour chart methods, demonstrated by the example point separations for each approach below the main plot in Figure 3.7. This analysis demonstrates that on a practical level the

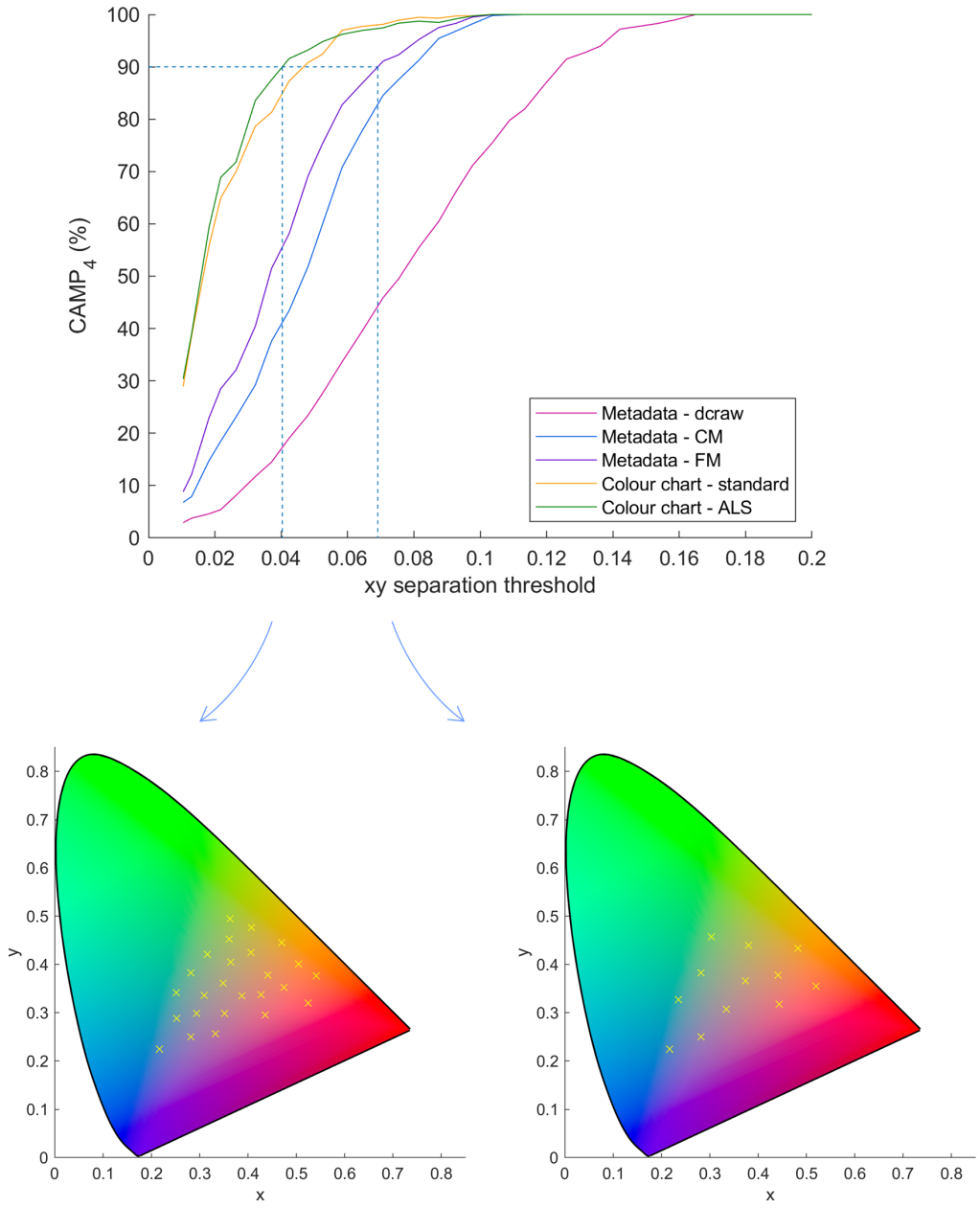


Figure 3.7: $CAMP_4$ (Classification Accuracy for Multiple Phones, 4 phones) for different subsets of patches from images of the DC chart under no ambient illumination. Results are shown for three metadata methods: dcraw (pink), dcraw-style using the correct ColorMatrix (blue), ForwardMatrix (purple). Results are also shown for the standard (orange) and ALS (green) colour chart methods. The blue dashed lines indicate the xy distance at which a 90% $CAMP_4$ is achieved for the ALS colour chart method and the best metadata method, with corresponding example DC patch subsets shown below to enable a visual understanding of different colours that can be discriminated using the two approaches.

mapping developed using the ALS method performs similarly to the standard mapping. The per-patch ALS mapping approach was therefore adopted for all further analysis, owing to its simpler calibration imaging process.

At this stage, it would be tempting to assume that the colour chart mapping developed for a phone of a given model would be applicable to another phone of the same make and model. It is known that variations in digital cameras prohibit this [103, 104], however results for smartphones have not been presented. Unfortunately, the combination of subtle variations in the CSS and variations in the SPDs of the LEDs or screens of typical smartphones is too great to allow a model-level calibration. Figure 3.8 demonstrates the impact of using a model-level mapping rather than a device-level. To produce the model-level results, the device-level ALS mapping for one phone of each model was selected as the model-level mapping and applied to both phones of that model. The $CAMP_4$ values were then calculated as before, and are presented here alongside the previous colour chart data and best metadata results for comparison. A significant drop in accuracy is observed, showing that a device-level calibration is necessary to maintain accuracy. However, since this device-specific calibration need only be carried out once per device it is not too onerous.

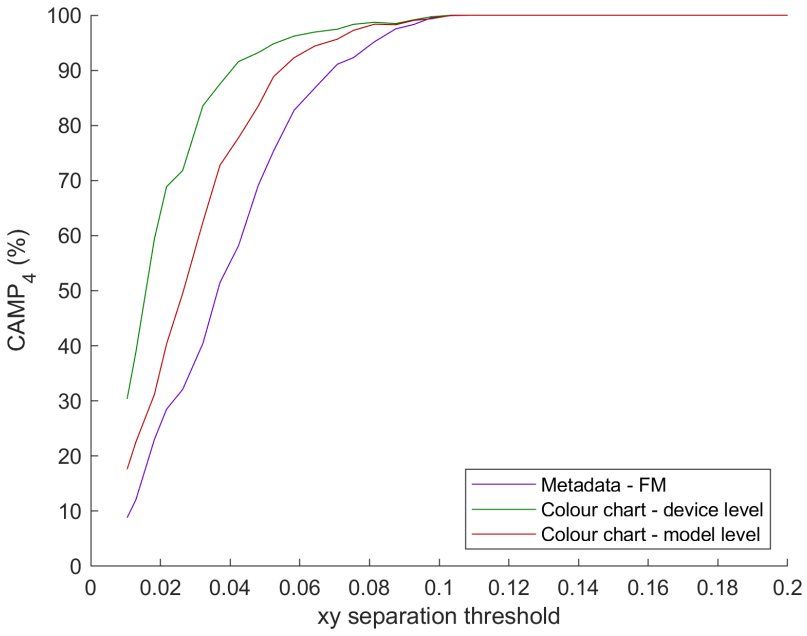


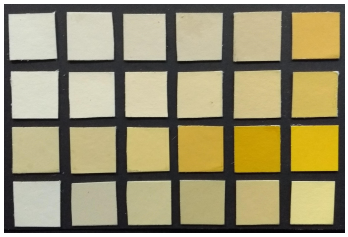
Figure 3.8: The $CAMP_4$ for different subsets of patches from images of the DC chart under no ambient illumination is shown again for the ForwardMatrix metadata method (purple) and the original ALS device-level colour chart method (green), along with results from a model-level calibration (orange). The loss in accuracy for small xy distances highlights the need for a device specific calibration.

3.2.4. Custom colour chart

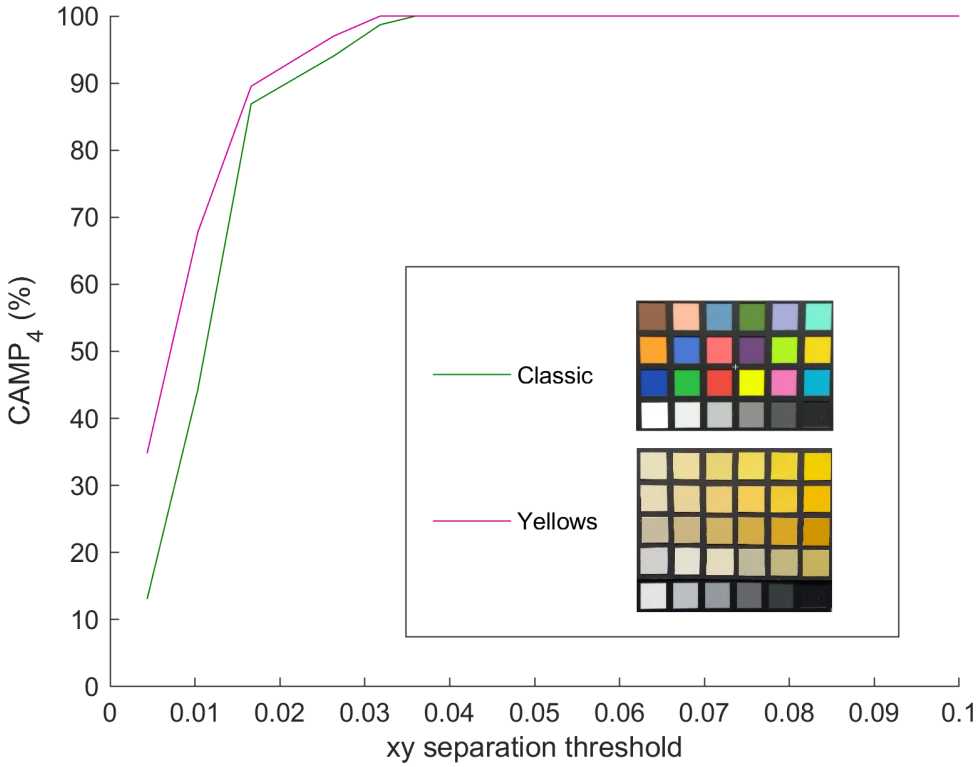
Figure 3.7 demonstrates that the classification accuracy across multiple phones is still low when the xy separation between points is small. This is not ideal here, since the task is to discriminate between different levels of yellow which will naturally have small xy separations. The use of a standard colour chart means that a variety of colours can be mapped with reasonable precision and accuracy, however as previously discussed there will be a reduction in accuracy for each patch to maintain a reasonable overall accuracy. If the colour in question is out of the gamut of the colour chart, the mapped value is likely to be highly inaccurate. Akkaynak et al suggested an approach for scene-specific colour calibration which involves measuring the radiance of different parts of the scene and then calculating the corresponding XYZ values, as well as simulating the RGB values through additional calibration of the camera or obtaining them through images [103]. This approach is ideal for complex environments, such as underwater marine monitoring, however for more typical environments the additional complexity and specialised equipment required outweighs the benefit. A simpler approach would be to develop a physical custom colour chart with patch colours targeted according to the application. Such charts have been developed for a variety of different applications, such as monitoring agricultural growth [105–107] or wound healing [108], and digitizing artwork [109–111].

Here, we are only interested in mapping white - yellow accurately, and do not mind if other colours are mapped poorly. It is therefore viable to create a physical custom colour chart with patches spanning the required colours, maintaining the simplicity of equipment required to carry out a calibration. Since we do not have ground truth reflectance spectra or even ground truth RGB/XYZ values for the colour of real patient eyes for different bilirubin levels, it is hard to choose the colours for the chart. Chapter 5 presents preliminary work in modelling and direct measurement of sclera reflectance spectra which could inform this work. As a starting point, a custom chart was created covering a range of shades of yellows and neutrals. A per-patch ALS mapping was developed using this chart, using the method described in Chapter 2.6.4. The aim was to test whether the use of a more targeted chart does indeed improve the accuracy of mapped results. To this end, a second yellows chart was constructed covering a smaller range in xy space as a testing chart, with patches from a different manufacturer and with different reflectance profiles. Images of this testing chart were captured, the Classic and yellows mappings were applied and the results compared using the $CAMP_4$ metric as before.

Figure 3.9 shows images of the charts used and the $CAMP_4$ as a function of the mean minimum xy separation for the subsets with no ambient light. The yellows chart approach provides a slightly higher classification accuracy than the Classic chart for all xy separations, with larger differences for small xy separations. This experiment demonstrates that a more targeted mapping can improve the resulting accuracy across phones. It does not necessarily mean that this particular yellows



(a)



(b)

Figure 3.9: An image of the yellows chart used for testing is shown in (a) with the $CAMP_4$ shown in (b). Results are presented for per-patch ALS mappings developed using the Classic chart and training yellows chart (with images of the charts in the legend) as a function of the average minimum xy separation for different point subsets of the test yellows chart with no ambient light. Note the slight increase in classification accuracy for small xy separations when the yellows chart approach was used, demonstrating the potential of using a colour chart with a smaller range of colours.

chart will improve accuracy for real patient data, since it was not designed based on a ground truth. Since the training chart covers quite a narrow range of yellows, it was decided that the Classic mapping should be used with patient data, to avoid poor mapping of out of gamut colours.

3.2.5. Individual phone accuracy

To compare the impact of different mapping approaches, the results for all four phones included in this research were combined together into an overall classification accuracy. When assessing the suitability of a new phone for use in a particular application, the individual colorimetric accuracy can be useful. Of course there are other important factors such as access to raw images and sensor linearity (see Appendix C), but we note that the individual colorimetric accuracy can provide useful insight.

It can be useful to look at performance both over a wide variety of colours and for a region of colour space specific to the application. Here, we use the DC colour chart patches shown in Figure 3.6 as a broad view. We also focus in on the yellow region since accuracy of measuring yellow is of particular interest. Selected patches are shown in Figure 3.10, and are a subset of DC chart patches supplemented with those from the two yellow charts discussed in Section 3.2.4.

As in previous sections, we use xy distance as a metric rather than the more conven-

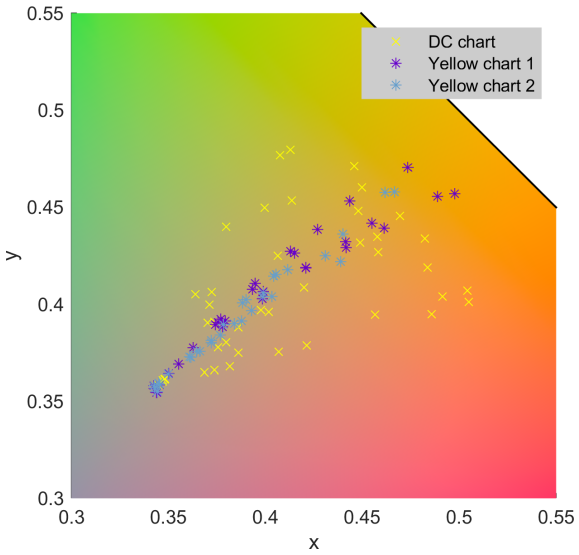


Figure 3.10: The ground truth xy values are shown for the Macbeth DC chart in the yellow region of the xy horseshoe (yellow crosses) and the two custom developed yellows colour charts (blue and purple stars).

All colours

	xy distance to ground truth	
Phone	Mean	Standard deviation
Nexus 1	0.01	**0.009**
Nexus 2	**0.008**	0.01
S8 1	0.01	0.01
S8 2	0.01	**0.009**

Table 3.4: Mean and standard deviation xy distances of mapped phone data to the ground truth DC colour chart patches shown in Figure 3.6 are presented for the four phones used in this research. The per-patch ALS mapping approach with the ColorChecker Classic was used in each case.

Yellow region

	xy distance to ground truth	
Phone	Mean	Standard deviation
Nexus 1	**0.006**	**0.007**
Nexus 2	**0.006**	**0.007**
S8 1	0.008	0.008
S8 2	0.008	0.009

Table 3.5: Mean and standard deviation xy distances of mapped phone data to the ground truth colour chart patches shown in Figure 3.10 are presented for the four phones used in this research. The per-patch ALS mapping approach with the ColorChecker Classic was used in each case.

tional metrics such as Delta E error since we only have chromaticity information. Mean and standard deviation xy distances between mapped and ground truth patch values are presented for the four phones used in this research. Results for the full range of colours and focussed yellow region are presented in Tables 3.4 and 3.5 respectively.

For the full set of colours the performance across phones is relatively consistent. For all phones, there was a slight improvement in performance for the focussed yellow region. It is reassuring that there is not a significant deterioration in accuracy in the area of interest, and that the results from all phones are similar - this implies that most phones may be viable for use. As new phones are considered for clinical data collection, it would be useful to benchmark them against these results since the corresponding clinical performance is known for the existing phones (presented in Chapter 4). If a significant deterioration is observed compared to the results in Tables 3.4 and 3.5, it may be possible to improve results using a custom chart (as discussed in Section 3.2.4) or it may be necessary to avoid use of this phone.

3.3. Proposed approach

Having presented the theory and relevant validation experiments, the proposed approach for obtaining accurate colour values from images over different environments and phones is summarised here.

3.3.1. One-time calibration

The first step when introducing a new phone for data collection is to carry out the one-time calibration. The approach detailed in Chapter 2.6.4 should be followed, as depicted in Figure 2.8. To summarise: a flash/ no-flash pair of images of the colour chart should be captured. The phone should be positioned at approximately 45° to the charts to minimise any specular reflection, and ambient light should be minimised. The per-patch alternating least squares algorithm, accounting for the intensity non-uniformity of the smartphone flash across the chart, should be used to develop the linear mapping M from phone native space RGB values to device-independent XYZ values. The use of this approach rather than the standard approach applying a grey chart to correct for shading means that the smartphone used to capture images of the colour chart can be handheld.

3.3.2. Data collection

Once the one-time calibration has been carried out, the data collection and following analysis process is extremely simple. Figure 3.11 shows the steps involved. First, an image pair containing the region of interest is captured, for example the patient's eye, ensuring that an SSNR of at least 3.4 is obtained. The particular scleral region of interest is selected from the flash and no-flash colour images, and a median RGB value calculated for each from the raw data. Then the no-flash median RGB value is subtracted from the flash median RGB value. This process accounts for any motion between images as well as the ambient light. The previously determined device-specific mapping (M) is applied to convert the subtracted RGB values into XYZ values. Finally the device and ambient light-independent xy chromaticity values are calculated. Datasets containing results from more than one phone are then compatible. It is possible at this point to begin developing the link between obtained colour values of the sclera and the corresponding blood test measurements of total serum bilirubin.

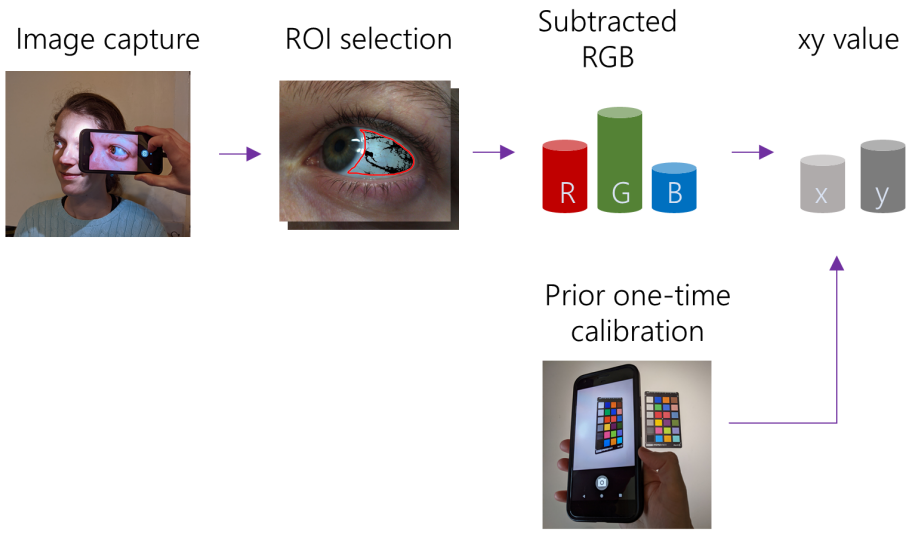


Figure 3.11: The simple process for obtaining colour values from a subject is depicted. First, a flash/ no-flash image pair of the overall region of interest is captured, here the patient's eye. The particular region of interest (ROI) is selected for each image using colour versions of the raw images. Median RGB values are calculated from the raw images for the flash and no-flash image ROIs, and the results subtracted. The previously calculated RGB to XYZ mapping (M) is then applied and xy chromaticity values are calculated, yielding device and ambient light independent colour values.

3.4. In-image colour chart alternative approach

For comparison, an alternative to the proposed approach was also used and is described here. Figure 3.12 depicts the method. For this method, images are captured under ambient light and a colour chart is included along with the patient's eye in every image. The colour chart is used to develop a mapping from RGB to XYZ and applied only to data from that image. A semi-automatic method can be used to extract patch data from images [112]. Even when under relatively spatially homogeneous ambient lighting there is likely to be a variation in the illumination intensity across the chart. For this reason, as for the proposed one-time calibration, the ALS method of determining the mapping is used. The median value for the sclera region of interest is then determined, and the per-image mapping applied to move to XYZ space. xy chromaticity values can then be calculated as for the proposed approach.

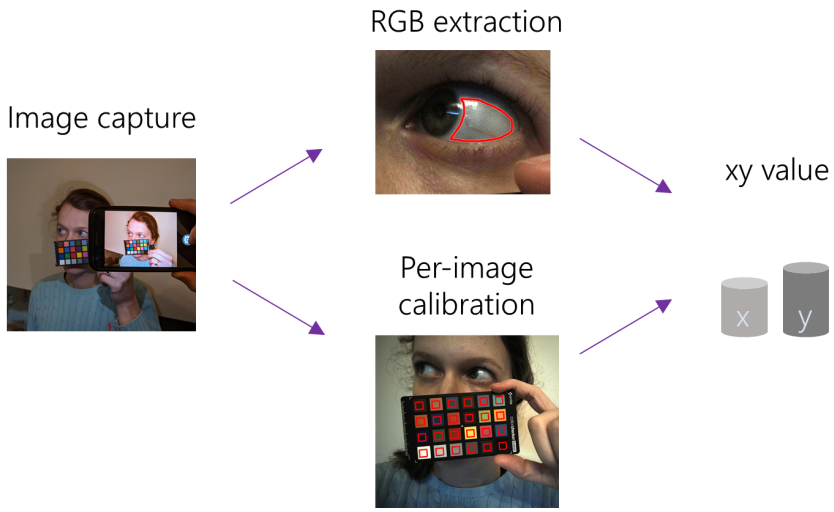


Figure 3.12: The in-image colour chart method is depicted. A raw image of the region of interest (here the patient's eye) is captured which also includes a colour chart. The colour chart is used to generate an image-specific mapping from RGB to XYZ. The median RGB value for the ROI is extracted, and the image-specific mapping is applied. Finally, xy chromaticity values are calculated.

4. Clinical study

Parts of the work presented in this chapter are to be published in a PLOS Digital Health article [113] and in the Journal of Gastroenterology and Hepatology [114]. These papers, licensed under the creative commons license CC BY 4.0, have been modified to form part of this thesis.

4.1. Data collection

A study was carried out on patients with advanced liver disease through the Institute of Hepatology at the Royal Free Hospital. Ethics for this study was granted as part of the DASIMAR study (amendment 6). The DASIMAR study was approved by the local governing Research Ethics Committee (London – Harrow; REC Ref: 08/H0714/8) and all patients provided informed consent. The data for this study was collected during the period of May 2017 to March 2021 by clinical fellows on the hepatology ward. Criteria for inclusion in the study were that the patient be over 18 years of age, have diagnosed cirrhosis either histologically or based on clinical, radiological and biochemical criteria, and have acute decompensation with no improvement of clinical state within the first 72 hours as assessed by the treating clinician.

Results from the daily blood tests carried out as part of in-patient care, including the TSB and haemoglobin levels as well as other parameters required to calculate clinical scores, were recorded along with capturing three types of images. A visualisation of the differerent aspects of the dataset is shown in Figure 4.1, with more detail provided in the following text. The primary type of images captured was images focussed on the sclera region alone, to test the proposed method for determining bilirubin. The second was images of the sclera region also including a ColorChecker Classic card, to allow a comparison against a standard alternative method. And the third type was images of the lower eyelid, to enable preliminary work into determining haemoglobin level.

Over the course of the collection period, a total of 57 patients were recruited. The first 11 patients were recruited during a pilot period and detailed information, other than their TSB level, was not recorded. Characteristics of the 46 patients from the main study are detailed in Table 4.1. To investigate the ability to track individual patients over time, repeat sets of data were collected, spaced by at least 3 days to

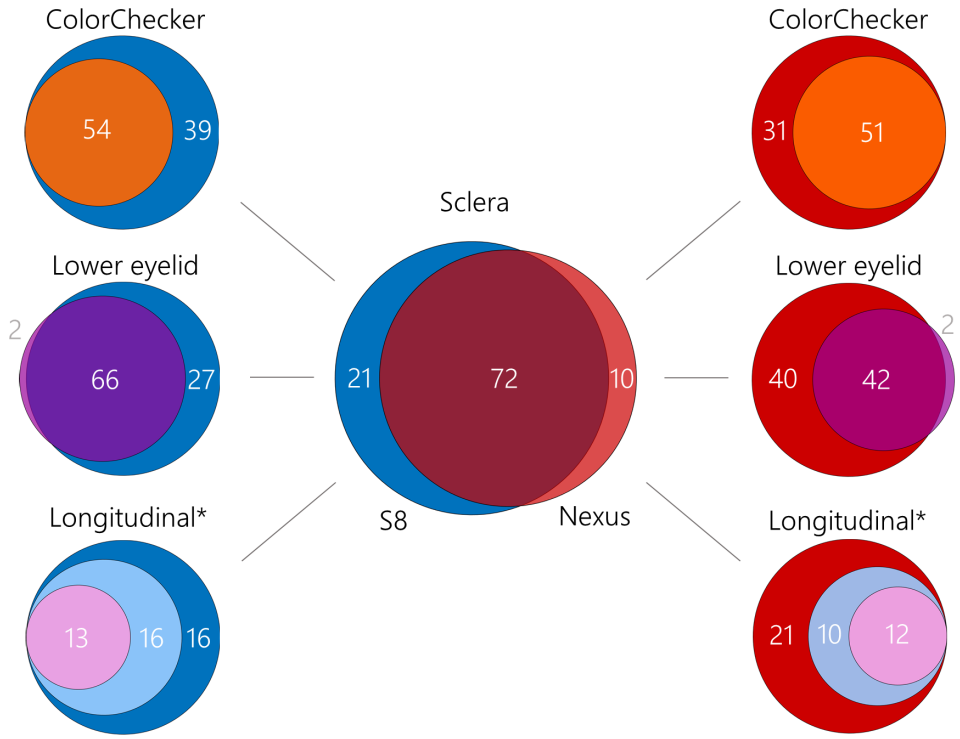


Figure 4.1: Visualisation of the patient image sets from the clinical study available after thresholding for the SSNR and applying filtering (see later section for in-depth discussion). The central Venn diagram shows image sets collected of the sclera alone using the two study phones S8 (blue) and Nexus (red) and their overlap. The left and right columns show the overlap of the sclera images with sclera ColorChecker images (orange) and lower eyelid images (purple) for the S8 and Nexus respectively. Note that in an ideal case, all image sets would be available for both phones and for all image capture types but for practical reasons this was not always possible. The longitudinal diagrams represent not image sets but the number of patients with repeat image sets available, with the outer circle representing those with a single set of images, followed by two (pale blue), and three or more (pink).

allow the bilirubin level to change. In total sets of patient images were collected at 112 timepoints, with 32 patients with two or more repeats and 16 patients with three or more repeats.

In order to test the ability of the system to obtain device-independent values and to demonstrate the two modes of subtraction, images were obtained in parallel using two phones. Figure 4.2 shows the process of sclera image capture for the two phones, where patient consent was acquired to use these images. Ideally all timepoints would have all three types of images captured using both phones, however this was not always possible. The breakdown of image availability and overlap of sclera images is detailed in Figure 4.1. The phones were briefly introduced in Chapter 3.1.1 but will be described in more detail now. The first phone used was the LG Nexus 5X,

		Number of patients, n = 46
Female : Male		14 : 32
Age (years)		49 (37 - 61)
Child-Pugh score		10.5 (9 - 12)
MELD-Na score		26 (21 - 29)
CLIF-C AD score		56 (49 - 63)
Development of ACLF (n, %)		15 (33%)
Length of hospital stay (days)		17 (8 - 34)
Need for intensive care (n, %)		10 (22%)
In-hospital mortality (n, %)		7 (15%)
Death or orthotopic liver transplant (OLT) within 28 days (n,%)		5 (11%)
Death or OLT within 90 days (n,%)		17 (37%)
Overall mortality (n, %)		19 (41%)
Cirrhosis aetiology	Alcohol	34 (74%)
	Nonalcoholic steatohepatitis (NASH)	4 (9%)
	Hepatitis C (HCV)	2 (4.3%)
	Autoimmune hepatitis (AIH)	2 (4.3%)
	Primary sclerosing cholangitis (PSC)	2 (4.3%)
	Secondary biliary cirrhosis	1 (2%)
	Unknown	1 (2%)
Decompensating event (n, %)	Alcoholic hepatitis	26 (57%)
	Infection	8 (17%)
	AIH flare	1 (2%)
	Drug-induced liver injury (DILI)	1 (2%)
	Dehydration	1 (2%)
	Bleeding	1 (2%)
	Unknown	8 (17%)

Table 4.1: Characteristics of the patients included in the study, not including those in the pilot study. Ranges presented are one standard deviation above and below the presented mean value.

referred to from here on as the Nexus. The Nexus used the front-facing camera for image capture, with the illumination for subtraction coming from the screen backlight being switched on to a solid white illumination. For self-capture, use of the front-facing camera is much simpler.

The second phone was the Samsung Galaxy S8, referred to from here on as the S8. In contrast to the Nexus, the S8 used the rear-facing camera for image capture. The rear-facing camera is much simpler to use when a third party is collecting the images.

(a) (b)

Figure 4.2: Images showing image capture at the Royal Free Hospital by a clinical fellow. Patient permission was obtained before acquiring the images. (a) S8 phone using the rear facing camera (b) Nexus phone using the front facing camera, where the patient's face has been blocked for confidentiality.

The camera flash was used for the illumination, however given the required proximity to the patients' eyes the pure flash would have been too bright for both comfort and image quality. A simple diffuser was developed using white diffusing acrylic which attaches to a phone case and covers the phone flash, shown in Figure 4.3. This diffuser lowers the intensity of the flash and makes its spatial distribution more uniform. The sides of the diffuser were covered with blackout tape to reduce the intensity of light coming from through the sides. The material used for the sides should entirely block the light, otherwise the colour of the light coming from the flash can become spatially non-uniform and negatively affect the results.

(a) (b)

Figure 4.3: The diffuser for the S8 phone used in this research. The diffuser was custom designed to dim the intensity of the flash and make its distribution more spatially uniform. (a) The diffuser before being attached to the phone case - the edges of the acrylic were covered using blackout tape to avoid a halo effect in the captured images. (b) The rear of the S8 phone shown with the case and custom designed diffuser in place.

For both phones, a custom app was used for image capture. When the app was opened, the flash of the phone (screen or actual flash) was turned on to allow the patient to acclimatise to the light. When capture was requested, the app automatically selected an ISO and exposure time based on the total illumination. This same combination of ISO and exposure time was used for both the flash and no-flash raw images which were then automatically captured in quick succession. The use of the same settings based on the flash image meant that ambient subtraction was valid, according to the linearity calibration detailed in Appendix C, and that pixel saturation was minimised. Note that for smartphones available at the time of this research, the camera aperture was fixed meaning that only ISO and exposure time varied.

The one-time calibration of each phone and processing of captured patient images was carried out according to the method outlined in Chapter 3.3. The need for more detailed regions of interest on the images, and subsequent filtering algorithm development, are discussed in Section 4.2. The availability of sclera image sets at different stages of processing for the two phones is shown in Table 4.2. For all cases when image capture was carried out, at least one image pair of acceptable quality was obtained. The vast majority of these had an SSNR level above the threshold, with a higher attrition rate for the Nexus owing to its dimmer flash illumination. Finally, the developed filtering algorithm performs well on the images. This data highlights that it was possible to obtain good quality image data for this patient cohort. After discussing and presenting the filtering algorithm, results for linking the extracted colour to TSB are presented in Sections 4.3-4.5. Finally, preliminary work considering haemoglobin assessment is presented in Section 4.6.

Stage of processing	Number of image sets remaining	
	S8	Nexus
Images captured	99	99
Useable images	99	99
Sufficient SSNR	97	89
Acceptable filtering	93	82

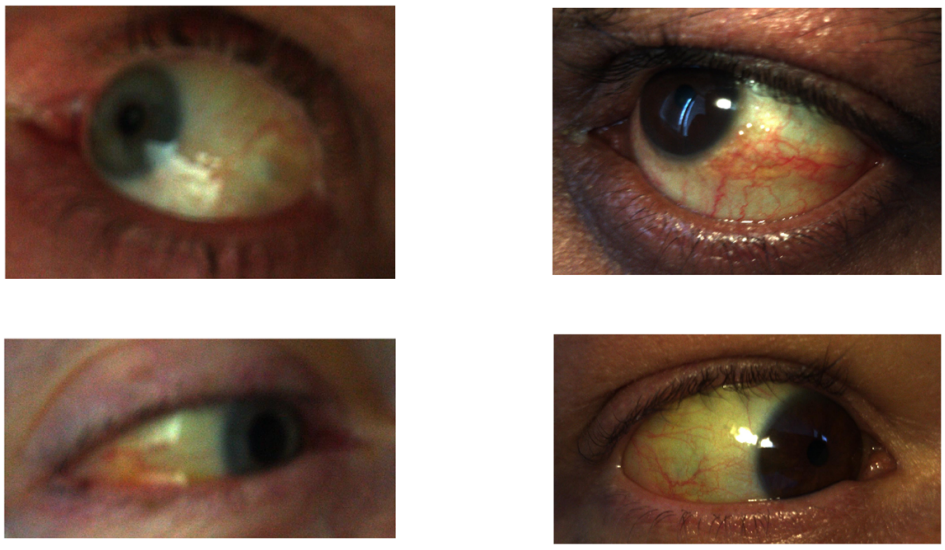
Table 4.2: Availability of sclera image sets at each stage of processing for the two study phones. Note that a total of 112 sets of images were captured - ideally all images would have been available for both phones, however in some instances only one phone was used.

4.2. Sclera filtering

4.2.1. Need for filtering

As shown in the data collection processing pipeline in Figure 3.11, after the region of interest selection, median RGB values are calculated for flash and no-flash images and then subtracted. Images of the sclera generally contain features such as blood vessels and specular reflection, referred to from now on as non-sclera pixels. The median is reasonably unaffected by outliers, hence it was selected as the averaging method to find the representative colour of pure sclera. Upon inspection of patient images, it became clear that the majority of these images contained large quantities of these non-sclera pixels leading to concern that these pixels may be affecting the end result. An illustrative set of images for both phones are shown in Figure 4.4, cropped to maintain patient privacy.

Theoretically, it may be possible to manually select regions which seem to contain less specular reflection and blood vessels. This approach is not ideal for two reasons. Firstly, it is not reasonable to expect patients or caregivers to reliably do this, especially since the goal is that users would require little training. To reduce user error or bias in region of interest selection, the long term aim is to introduce an automatic



(a) (b)

Figure 4.4: Two different example cropped patient flash images are shown for the Nexus (a) and S8 phone (b), demonstrating the large proportion of blood vessel and specular reflection pixels typically found in these images.

or semi-automatic segmentation algorithm to select the sclera region rather than the current manual selection. Secondly, even if this targeted region selection were possible, it is unlikely to yield optimal results. Smaller regions of interest means fewer pixels included and so higher noise levels. Additionally, due to the inherent variation in colour of the sclera, selecting a small localised region could lead to the introduction of a colour bias.

An alternative approach is to start with a full sclera mask, obtained through manual or automatic segmentation, and filter out non-sclera pixels. In other words, rather than selecting a regional mask, the aim is to remove problematic pixels from a full mask. Once these pixels have been removed from the mask, the median can be calculated as before. The filtering does not have to be perfect, since the aim is simply to make sure that there are more sclera pixels than non-sclera pixels and to have a reasonable number of pixels left in total.

4.2.2. Filtering algorithm

Figure 4.5 shows the overall algorithm developed to filter out non-sclera pixels. sRGB images generated from the raw Bayer images using a standard metadata conversion [82, 115] implemented in MATLAB were used for the pixel filtering. Clearly, for reasons discussed in Chapter 2.2, whilst the exact values of these converted images should not be used for analysis, they are useful for relative colours. Pixel filtering for both specular reflection and blood vessels is carried out for the flash image, followed by an attempt to register the flash and no-flash images. If the registration is successful, the filtered flash mask is transferred to the no-flash image. If not, then a second round of filtering is carried out on the no-flash image.

Not every sclera region contains specular reflection, as seen in Figure 4.4, but when it is present it is problematic. Often, the specular reflection presents as fully saturated pixels which are easy to remove. However, there are often also regions of specular reflection which are visible to the human eye but do not saturate the camera sensors. It is therefore not enough to simply remove saturated pixels. Carrying out this filtering operation on images where there is not specular reflection present can cause over-filtering of sclera pixels. Therefore it was necessary to do a basic test for the presence of specular reflection in the region of interest before carrying out the filtering. If any pixels met one of the following three criteria, then there was deemed to be specular reflection present:

- One or more of the RGB channels has a value above 0.95 (range: 0 - 1)
- Two or more of the RGB channels have a value above 0.9

- Any pixel intensity is more than five standard deviations away from the median intensity in the ROI

The third criterion was included to tackle specular reflection in images where the overall sclera intensity was low and so neither of the first criteria were met. For images where specular reflection was detected, a simplified method based on the work of Crihalmeanu et al [116] was used. A conversion from sRGB to HSV space was first carried out and a power law transform applied to the V channel to accentuate the specular reflection - V^{γ}. The original method involved testing a range of values for γ, however here it was found heuristically that a fixed value of 4 performed adequately. Diverging from the original method, Otsu thresholding was then used to separate specular from background pixels [117]. Otsu thresholding works by iterating through a range of possible threshold values for a greyscale image, splitting

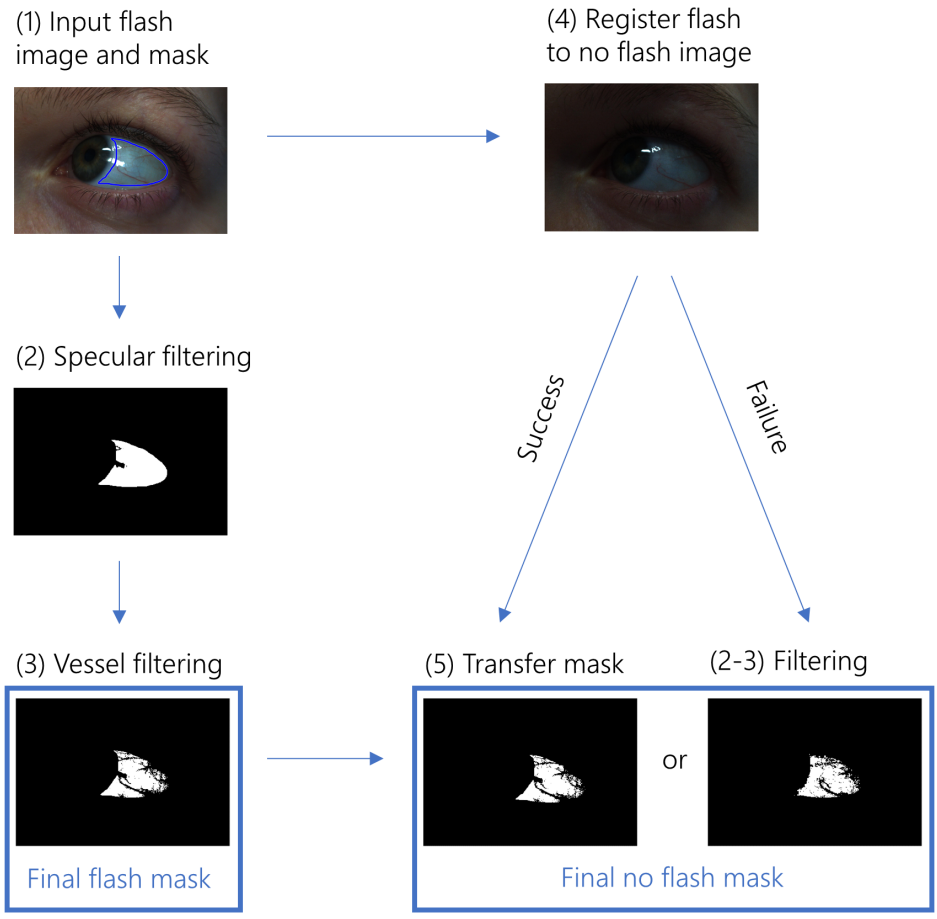


Figure 4.5: The algorithm for filtering out specular reflection and blood vessels from the sclera region of interest. Filtering for specular reflection and vessels is carried out on the flash image, before registering the flash to the no-flash image. If the registration succeeds then the filtered flash mask is transferred, else filtering is carried out on the no-flash image. Further details of the algorithm are discussed in the text.

the image into specular and background, and automatically selecting the threshold which minimises the intra-class variance. In other words it selects a threshold which minimises the width of the intensity clusters of the two classes. Here, the inbuilt MATLAB function graythresh was used.

The next stage was to remove blood vessels from the image. There can be varying amounts of blood vessels from person to person, but every sclera contains some. It is therefore possible to skip the step of testing for their presence, and simply apply the filtering to every image. Detection of blood vessels is a common task across many areas within medical imaging. A huge range of techniques exist for vessel extraction [118], including the well-known Frangi filtering approach [119]. There are many approaches developed specifically for sclera images, usually related to biometrics, for example involving contrast and line-based enhancements [116]. These approaches are able to return the vessel mask to a high degree of accuracy, however they typically involve many tunable parameters and can be quite slow.

Since our images are captured in different lighting environments and with different devices, high numbers of parameters which require tuning for different situations is undesirable. We note that here we do not require a high quality vessel segmentation, rather we would simply like to remove the majority of vessel pixels. With this in mind, a simpler alternative has been developed which was based on the basic observation that vessels are much more red than the surrounding sclera. The red chromaticity was calculated for the whole image, to mitigate the effects of shadows and to flatten the image, and then Otsu thresholding was applied to remove the redder vessel pixels [117]. Initially, the vessel filtering was carried out before the specular filtering. However it was found that the presence of specular reflection could cause the vessel filtering performance to deteriorate and significantly overfilter, so the order of these two filtering steps was reversed.

Due to the inevitable shifts between hand-held captured images, it is not advisable to use the filtered mask directly on the no flash image. Instead, once the flash image mask has been finalised, the flash image was registered to the no-flash image. A rigid transformation, allowing translation and rotation of the image, was determined using the built-in MATLAB function imregtform with the multimodal setting. The transformation was then applied to the filtered flash mask to use it with the no-flash image. In some cases for the test and patient images, a very large translation was suggested between the images. If any translation term in the mapping was larger than a tenth of the maximum image dimension, the registration was classed as having failed. The registration fails in two main cases: a) when there are

geometric changes between the images, in which case often the images are not as suitable for the subtraction method, or b) when the no-flash image is very dark. To distinguish between these cases, the intensity of the no flash image was calculated. If the intensity was above a heuristically determined limit, then the unfiltered flash mask was used on the no-flash image and the two stages of filtering were carried out as for the flash image. If the image was very dark then over-filtering would likely occur, so instead the unfiltered flash mask is simply used for the no-flash image. Where a separate manual mask for the no-flash image exists this can be used in place of the flash mask, and the same threshold used to determine whether to apply filtering.

4.2.3. Filtering validation experiment

It is challenging to validate whether the filtering algorithm is helping to improve the results from patient data, as there is no ground truth for the colour of the sclera or even which pixels are or are not pure sclera. One option would be to consider repeat images of the same patient taken at the same time, and see if carrying out filtering brings the colour values into closer agreement. Unfortunately, there are limited repeats available for each patient and they are not constrained enough to make a judgement.

Instead, a lab-based experiment was designed and carried out. Four samples were created, depicted in Figure 4.6, which were designed to be pseudo-representative of the sclera. The samples were created by adding red ink and clear adhesive tape, to represent blood vessels and specular reflection respectively, to four identical pale yellow paint swatches. Swatches were used as they are spot colours - the ink is pre-mixed and applied rather than printed in the typical CMYK manner. This makes swatch colours higher resolution and more consistent. The ground truth XYZ value for each sample was measured using the X-Rite ColorMunki spectrophotometer before any modifications were carried out. The samples were so close in value that it was possible to deem them the same, so an average of the results was used as the ground truth value. The first sample was kept unmodified, as a control, and the other samples were modified to include varying levels of blood vessels and specular reflection as described in the caption of Figure 4.6.

The samples were then imaged with the S8 phone with no ambient light to enable the filtering algorithm alone to be tested. Four repeats of each sample, rotating the sample between each repeat, were captured in three different setups. The first set was captured with a 45° angle between the phone and sample, designed to minimise

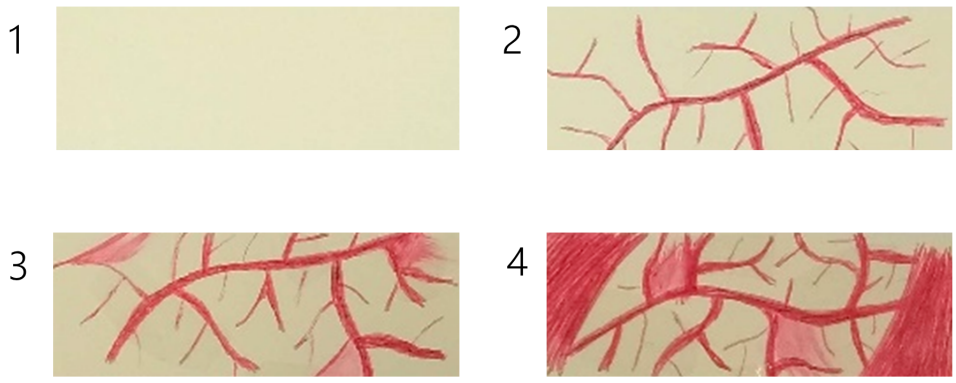


Figure 4.6: The four samples produced to test the filtering algorithm. Vessels and specular reflection were simulated using red ink and clear adhesive tape respectively. 1: plain uniform colour, the same background colour as all other samples. 2: vessels added. 3: vessels and specular reflection added (note that the specular reflection is not visible in these images due to camera angle). 4: more vessels and specular reflection

specular reflection. The second set was captured with the phone and sample parallel to one another, specifically aiming to generate specular reflection. The final set was captured with the phone and samples parallel, but with the samples bent into a curve to introduce geometric shading as observed for eye images. Manual regions of interest covering the whole sample were selected for each image, and the RGB values were extracted using the masks pre and post filtering. The device-specific mapping was then applied to map to xy space, and the results compared to the ground truth.

Images of the three modified samples and their masks for one setup each are shown in Figure 4.7. Visual inspection of all image data was performed to ensure that vessels and specular were being well removed. Note that the filtering performed well on these test samples, despite varying distance and angle from the card. The resulting xy values for the samples before and after filtering are shown in Figure 4.8, along with the ground truth result obtained by spectrophotometer measurement.

When inspecting the results before filtering, in Figure 4.8 (a) it can be seen that, as expected, the control sample already has a good agreement with the ground truth in all imaging setups considered. The other cards match the ground truth far less well, with the distance increasing for increasing modifications. The distance also increases as the setup moves from the ideal imaging setup to the more realistic setup - larger distances from the ground truth were observed when the phone was parallel to the curved samples. This means that if a patient image set has a large amount of visible blood vessels and specular reflection then it is likely that the extracted

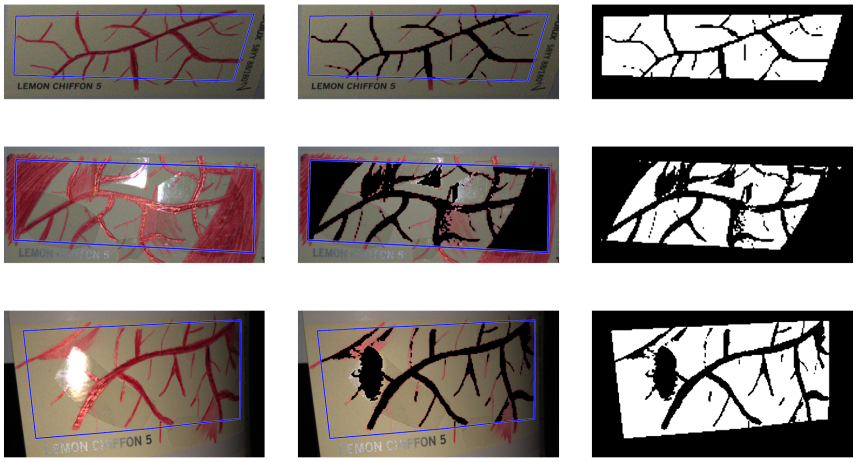
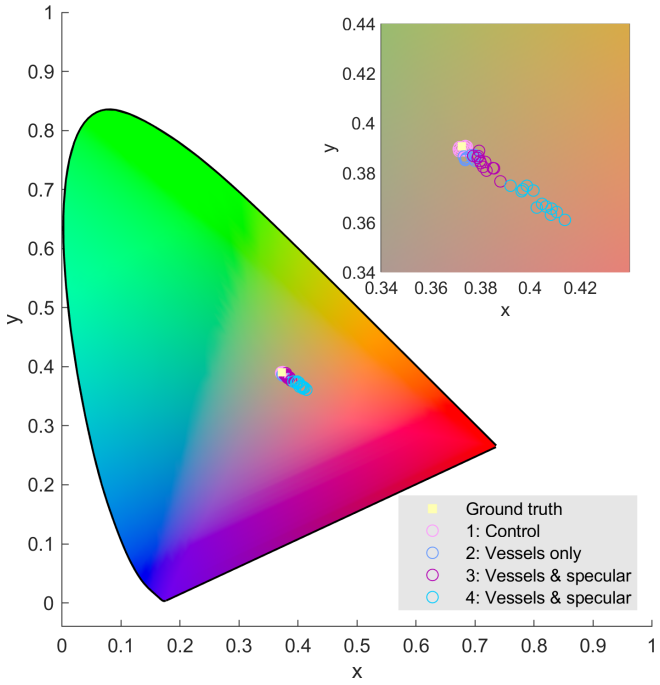


Figure 4.7: An example image of each of the modified samples is shown, before and after the filtering is applied. The manual mask outline is shown in blue. The first column shows the original image, the second shows the filtered image, and the third shows the resulting mask. The rows correspond to the three imaging setups - flat with no specular, flat with purposeful specular, curved with purposeful specular - shown for samples two, four and three respectively. Note how specular and blood vessels are accurately removed, whilst avoiding over-filtering.

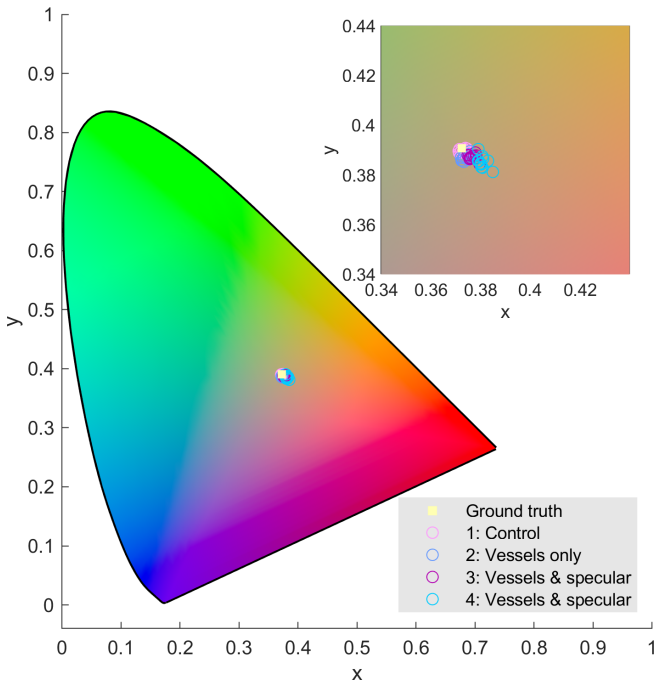
colour will depend on these features.

The xy values after filtering, in Figure 4.8 (b), display a much greater agreement with the ground truth. The results from sample two were close to the ground truth before filtering and were not greatly affected by the filtering. This confirms that the median is indeed robust to some outliers. Filtering makes a huge difference to samples three and four, with a significant reduction in xy distance from the ground truth (t test, $p<0.05$). The results after filtering are still not as good as the less modified samples. This is potentially due to some noise due to the overall smaller number of pixels remaining or the influence of some non-sclera pixels which were not entirely removed.

Overall the experiment demonstrates that the median alone is unlikely to be able to overcome the amount of non-sclera pixels encountered in real-world images. It also shows that a significant improvement is found by utilising the filtering algorithm described here.



(a)



(b)

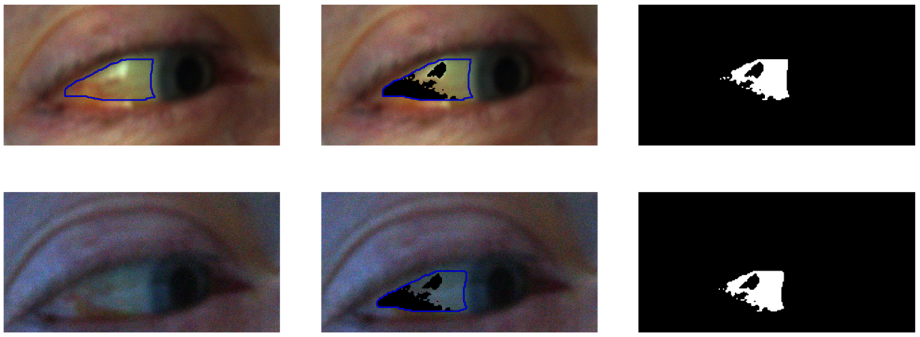
Figure 4.8: The xy values of the samples are shown before (a) and after filtering (b). The ground truth xy value for the samples is shown as a pale yellow square. The results for the numbered samples shown in Figure 4.6 are shown in pink, blue, magenta and turquoise respectively. Before filtering, the resulting xy values are highly spread from the ground truth with much closer clustering achieved after filtering.

4.2.4. Testing on patient data

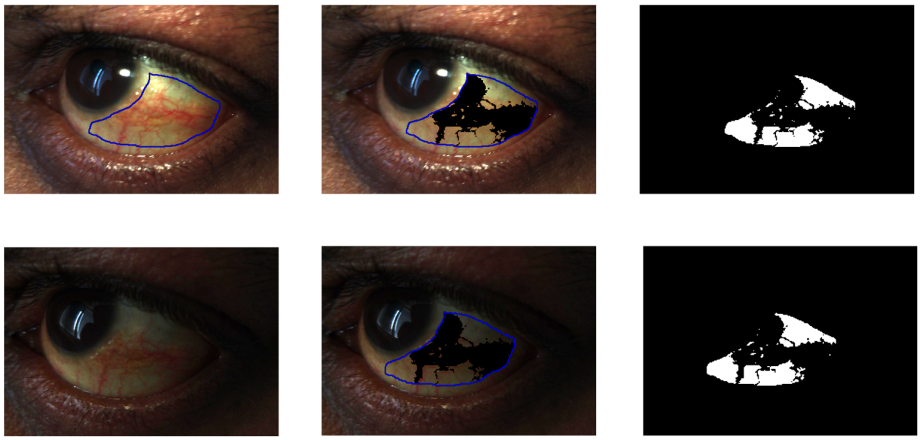
During the development of the filtering algorithm, testing was carried out on the patient data to ensure that it was fit for purpose, and to help hone the details. A key step was setting the intensity threshold for carrying out filtering on the no-flash image in case of registration failure. It was found that filtering on the no-flash image for the S8 phone worked to a lower intensity value than the Nexus phone, so different thresholds were set. It should be noted that this is the only tunable parameter in the algorithm, with all other stages being carried out automatically.

A necessary testing step for the algorithm in its final form was to assess how well it performed on the patient data. An example set of flash and no-flash cropped patient images and masks are shown in Figure 4.9 for each phone. Visually, the results look fit for purpose - the majority of blood vessels and specular reflection are removed. Histograms of RGB distributions for these example images before and after filtering are shown in Figure 4.10. Each distribution narrows upon filtering, saturated values are removed, and the red channel in particular is reduced in relative intensity in line with removal of blood vessels. These findings support the intended functioning of the algorithm. As an inherent result of applying the algorithm there is a clear loss in the number of pixels available for use. Figure 4.11 shows the percentage of sclera mask pixels available for analysis after applying the filtering algorithm. The mean available percentages for the S8 and Nexus phones respectively were 55% and 49%. This is a significant loss of data. However given that these pixels contain confounding information and that the absolute mean available number of pixels remains around 100,000 and 25,000 for the S8 and Nexus respectively (difference owing to their different image dimensions) the loss was deemed acceptable.

A visual inspection of the filtered masks was carried out for all available sets of patient data and repeat images. Image filtering was categorised as 'good' for 261/279 S8 image pairs (94%) and 238/257 Nexus image pairs (93%). The most common causes of issues in the algorithm were due to failure to register the images correctly and, specifically for the Nexus, not identifying specular reflection in all cases. Overall, the filtering algorithm performed well over a large set of images. Based on these qualitative good results and the findings of the lab experiment, it was decided that filtering should be applied as part of the processing pipeline with a visual inspection to check for any issues.



(a)



(b)

Figure 4.9: An example cropped image pair and corresponding masks before and after filtering for the Nexus and S8 phones are shown in (a) and (b) respectively. No-flash images have been brightened for visibility here. In each case, the top row shows the flash results, and the bottom row shows the no-flash results. The first column shows the flash image and mask (blue outline) before filtering, and the no flash image (no manual mask), the second column shows the images after filtering, and the third column shows the resulting masks. Note the large correct shift in mask location for the no-flash image with the Nexus phone, and the good level of vessel and specular removal for both phones.

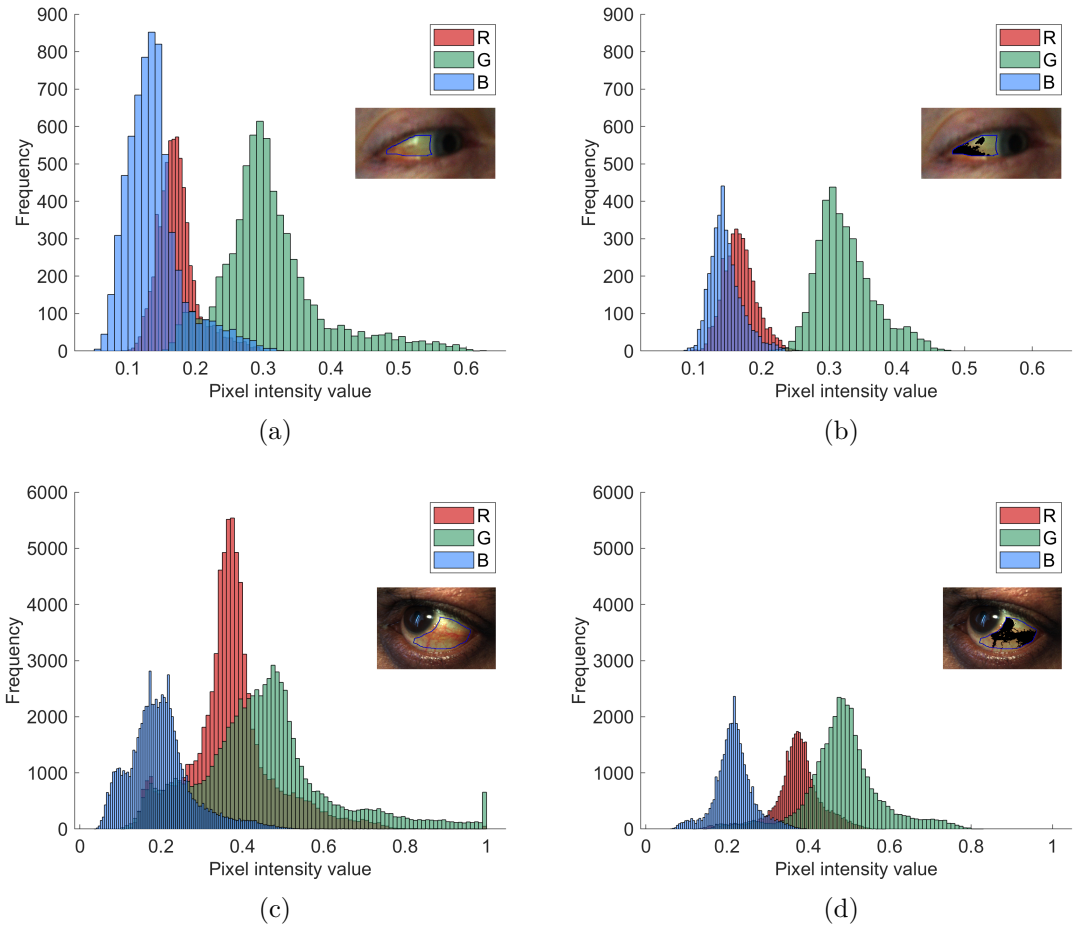


Figure 4.10: Histograms of RGB distributions before and after the filtering algorithm is applied. Results for the flash images of the example image pairs shown in Figure 4.9 are presented for the Nexus in (a) and (b) and the S8 in (c) and (d). The lefthand column shows the distributions before filtering and the righthand column shows them after filtering. Inset images show the mask regions.

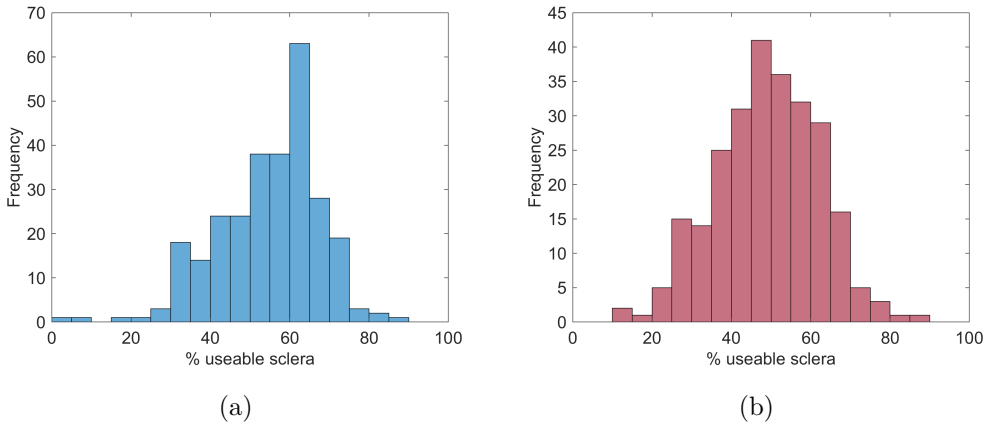


Figure 4.11: Histograms of the percentage of masked sclera area remaining for analysis after application of the filtering algorithm. Results for all available image sets for the S8 and Nexus phones are presented in (a) and (b) respectively.

4.3. Linking colour to bilirubin

4.3.1. Yellowness metrics

Before presenting any results, the metric to capture the 'yellowness' of the sclera must be considered. Since images were captured at different distances from the patient, under different lighting conditions, and with different exposure times, the subtracted RGB values and resulting XYZ values do not contain brightness information. This means that we cannot base a metric directly on individual channels, but should use chromaticity instead. Chromaticity, introduced in Chapter 2.4, involves a ratio of all the channels, so is independent of the image capture factors previously mentioned and has the additional advantage that it removes the effects of geometric shading.

For presenting results in device-specific space, we opt for a simple metric based on blue chromaticity. Since a lower level of blue denotes a higher level of yellow, the device-specific metric is 1-b to produce a positive correlation with bilirubin level. Figure 4.12 plots the absorption spectrum of bilirubin with the XYZ colour matching functions. The strong extinction of light in the blue wavelength range for bilirubin explains the yellowing of the sclera observed when bilirubin builds up. Of the three tristimulus values, the Z channel has the greatest sensitivity over the

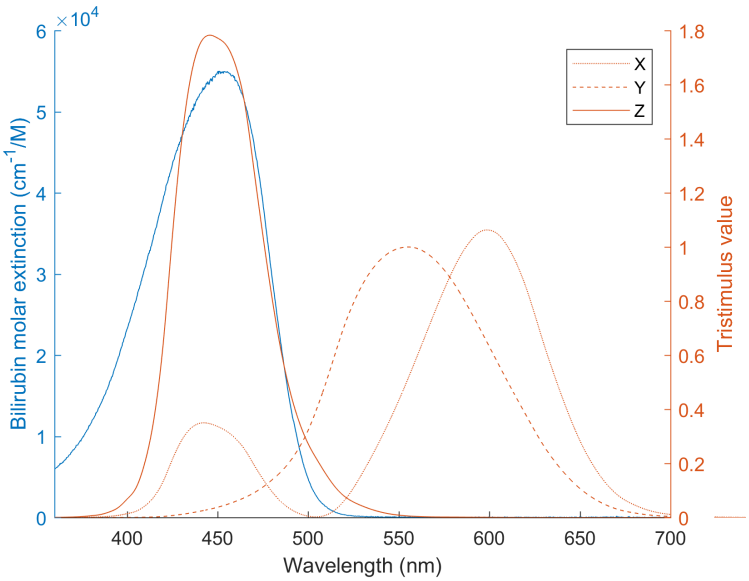


Figure 4.12: Bilirubin extinction is plotted as a function of wavelength in blue using the left hand vertical axis using data from [120]. The XYZ tristimulus values as a function of wavelength are plotted in red using the right hand vertical axis. Note the overlap between bilirubin absorption and Z.

relevant wavelength range. An intuitive metric, following the same form as for the device-specific metric, would therefore be 1-z. Section 4.3.3 considers this metric along with higher order prediction models based on multiple linear regressions of different combinations of chromaticities.

4.3.2. Device-specific results

Before applying the mapping to XYZ space, we first consider the data in native space. This enables a test of the first stages of processing before introducing the inherent slight error from the mapping. Figure 4.13 shows the correlations from the Nexus phone at three stages of processing. The general trends observed for the S8 phone were similar so have not been included. Before the subtraction step is applied there is a correlation, however the data is quite noisy. Applying subtraction greatly improves results, increasing the correlation coefficient from 0.56 to 0.82 ($p<0.001$ in both cases). This finding further demonstrates the power of ambient subtraction to reduce the impact of ambient light in a simple way. Finally, the correlation again increases for data which has had both filtering and ambient subtraction applied up to 0.87 ($p<0.001$). This demonstrates, in a practical way, how carrying out filtering for blood vessels and specular reflection can improve results.

For the Nexus phone, ten timepoints had to be excluded based on low SSNR - these are marked in pink in Figure 4.13. For the S8 phone, just two timepoints had to be excluded. This is due to the more challenging image capture with the Nexus. Since the front facing camera is used with the 'flash' provided by the phone screen, the overall brightness is lower, meaning that the phone must be held closer to the patient. Additionally, it was found that the patients were too ill to capture the images themselves using this phone, as had originally been intended, so the clinical fellows had to capture the images. It was therefore very hard to get the phone close enough to the patient whilst also being able to see the preview screen to check that the desired region of interest was being captured. The number of failures at the pixel filtering stage is also higher for the Nexus than the S8 - nine versus four. This is likely due to the lower image resolution, meaning that registration and filtering is more challenging. Additionally, the filtering algorithm has a higher rate of failure to detect the specular or partial specular pixels for the Nexus. The filtering algorithm could be improved in the future, however in its current form it already serves as a demonstration of the utility of filtering out vessels and specular reflection.

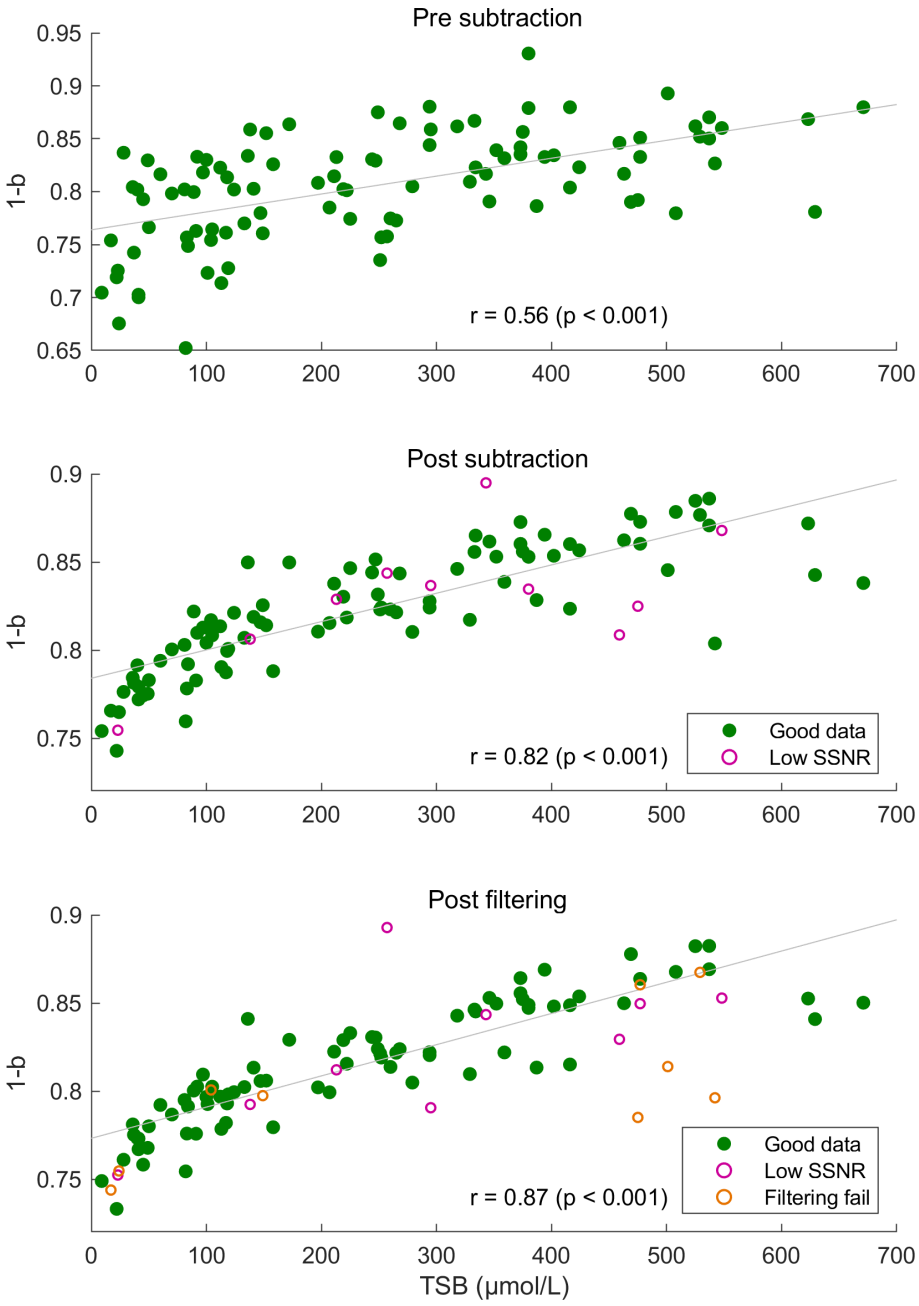


Figure 4.13: The yellowness metric, 1 - blue chromaticity, extracted from scleral images, taking a mean over measures from available repeat images, is shown as a function of the blood test result for bilirubin at three stages of processing for the Nexus phone. The top panel shows the data from the no flash images, before subtraction, where a linear best fit line has been included (grey). The middle panel shows the data after ambient subtraction has been performed, where image sets which failed to meet the subtracted signal to noise (SSNR) threshold have been excluded from further analysis (pink). Finally the bottom panel shows data after both ambient subtraction and pixel filtering have been applied, where additional image sets which failed a visual inspection of the filtering are also excluded (orange). The respective Pearson correlation coefficient (r) is shown on each panel - note the improvement after each stage of processing. The improvement over the course of the processing was similar for the S8 phone.

Having confirmed the utility of both pixel filtering and ambient subtraction, we turn our attention to handling multiple repeat image pairs. The processing described thus far could be applied to a single image pair at each timepoint to produce a yellowness value. However, the variability introduced by noise is typically reduced by using repeats. Table 4.3 presents the linear correlation coefficients obtained for the two study phones using three methods of handling repeat image pairs. The first method, labelled as random repeat, was to randomly select an image pair from those available and use the yellowness value produced. This process was repeated for all timepoints. The overall correlations produced varied, so the mean and range are presented in Table 4.3. The second method, labelled subjective best ranked, was again to use a result from a single image pair for each timepoint but this time to choose the pair. Image pairs were visually assessed and the best pair for each timepoint was chosen based on criteria including alignment, subjective SSNR level and presence of specular reflection. The final method was to simply take a mean over the results obtained from the available repeat image pairs.

When using the random repeat method, there is quite a bit of variability in the results for both phones. This confirms that using a single image pair may not produce reliable results. Using the subjective best ranked pair gives a slightly higher correlation than the average for the random repeat pair. However, this method relies on having an experienced user decide which image pair is best, and is, as the name suggests, somewhat subjective. The mean combining method produces the highest correlations of the three methods and does not rely on the input of a skilled user. Whilst collecting repeat images involves increasing the time required for im-

Combining method	S8 (n=93)		Nexus (n=82)	
	r	Range of r	r	Range of r
Random repeat	0.86	0.83 - 0.90	0.85	0.81 - 0.88
Subjective best ranked	0.88	-	0.85	-
Mean	0.88	-	0.87	-

Table 4.3: Linear correlation coefficients (r, $p<0.001$ in every case) of the extracted yellowness metric against TSB for the two study phones using different methods of combining data from repeat image pairs. Up to three repeats were available for each timepoint; the total number of timepoints for each phone is listed after the phone name. Random repeat: for each timepoint, one repeat was randomly selected from those available. This process was repeated 100 times and the mean result presented along with the minimum and maximum obtained correlations. Subjective best ranked: available repeat pairs were visually inspected and the result from the subjectively chosen best pair was selected. Mean: the mean over the yellowness metrics obtained from the available repeats was taken for each timepoint.

age capture, the total image capture time remains low and in this study patients were typically able to comply for three repeat images. The mean combining method has therefore been used for all data presented in this chapter, including the results presented in Figure 4.13. Note that for all methods a basic screening of images captured at each timepoint is initially carried out to select up to three repeats. This involves removing failed image pairs, for example where the patient blinked, and keeping suitable pairs. The process could be automated in the future to avoid any subjectivity.

4.3.3. Device-independent results

Having demonstrated the quality of the initial processing steps, we now move from device-specific space to device-independent XYZ space. To do this, as described in Chapter 3.3.2, the device-specific mapping is applied to the filtered and subtracted data. After the mapping is applied, the XYZ values are converted to chromaticity values. Note that since by definition $x + y + z = 1$, just two out of three values are needed. Having obtained these chromaticity values, we consider how to link colour to bilirubin level. The simplest approach would be to use just one of x,y or z. Multiple linear regression models could also be used.

A variety of models were constructed, with results presented in Table 4.4 for the S8 phone. Trends were similar for the Nexus phone so have not been presented. Two metrics were used to assess the performance of the different models. First is the overall linear correlation of the predicted bilirubin against the measured bilirubin. Second is the mean absolute error (MAE) of individual predicted bilirubin levels compared to measured bilirubin. The MAE was calculated using a 10-fold cross-validation to avoid training and testing on the same data. To further reduce variability in the results, this process was repeated 50 times and the average presented.

The data presented in Table 4.4 shows that there is a slight improvement in the overall correlation and MAE for models with higher numbers of terms. However, the results remain at a similar level for all models with three or more colour terms. The model with the best results and lowest number of terms is based on x,y,xy. Figure 4.14 shows a 3D visualisation of this model and the best single colour term model based on z. Inspection of the fits shows that the higher order model does not, for example, improve the fit to the data for higher bilirubin levels. Despite using cross-validation to calculate the MAE, there is still a risk of overfitting given the size of the dataset when using higher order models.

Given the modest improvement afforded by extra terms and the size of the dataset, the simpler z-based model for linking colour to bilirubin was selected. Figure 4.15 shows results for this metric against bilirubin level. The linear trend is well maintained even for very high bilirubin levels, as evidenced by the high correlation.

Colour terms	Correlation with TSB ($p<0.001$ for all)	Mean absolute error (μmol/L)
x	0.84	72
y	0.88	61
z	0.89	60
x, y	0.89	60
z, z^2	0.90	57
x, y, xy	**0.91**	**54**
x, y, x^2	**0.91**	55
x, y, y^2	**0.91**	55
x, y, xy, x^2	**0.91**	55
x, y, xy, y^2	**0.91**	**54**
x, y, x^2, y^2	**0.91**	**54**
x, y, xy, x^2, y^2	**0.91**	**54**

Table 4.4: Performance of multiple linear regression models for predicting bilirubin using different colour terms for data from the S8 phone. Performance is assessed using the overall linear correlation coefficient with TSB and the mean absolute error (MAE) in predicted bilirubin compared to measured bilirubin. The MAE was calculated for each model using a 10-fold cross-validation. The highest correlation and lowest MAE are shown in bold.

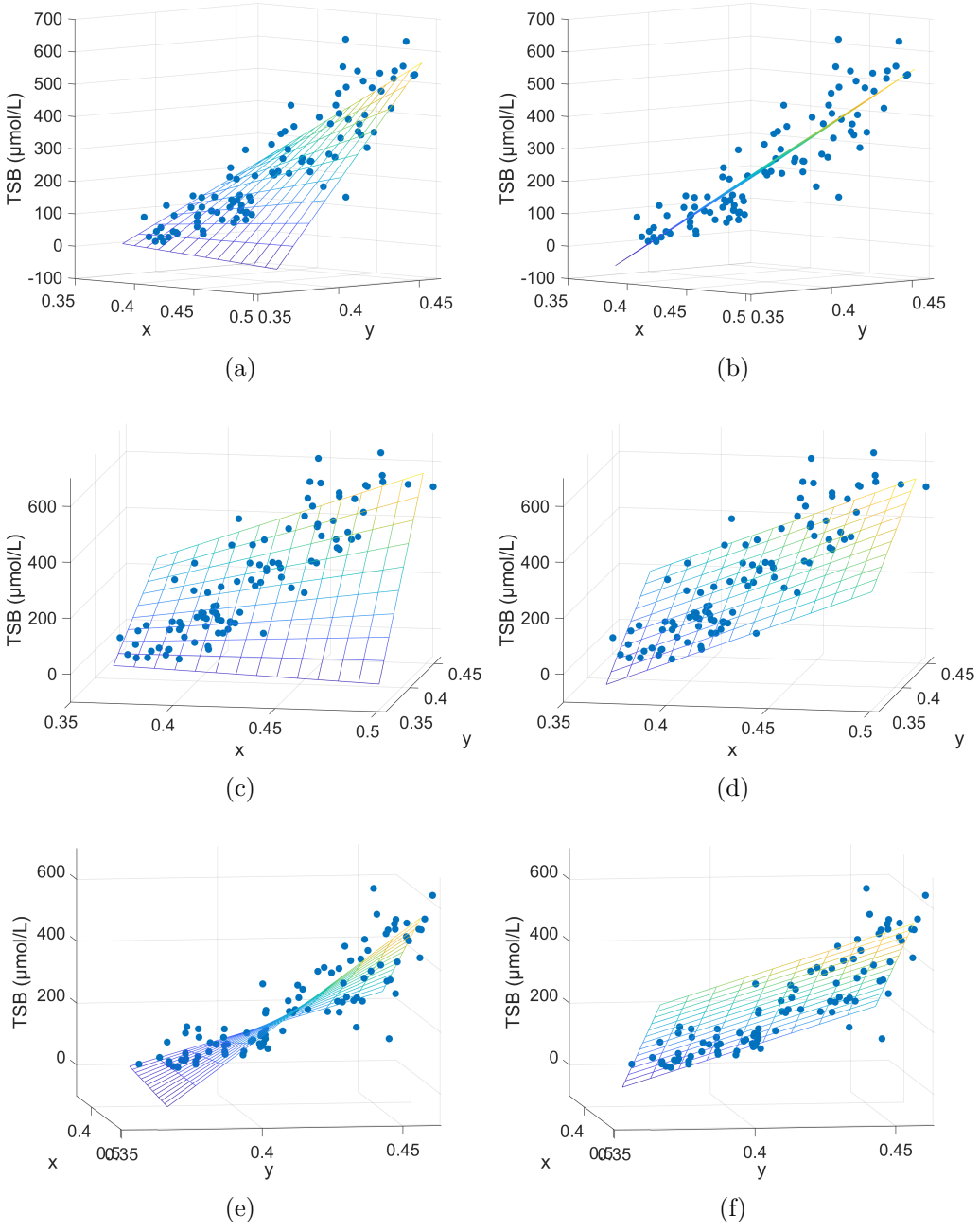


Figure 4.14: 3D visualisation of two models for linking colour to bilirubin, shown from three different angles. Colour values extracted from sclera images are scattered in blue, with the model fits shown as a coloured grid. The model based on x,y,xy is shown in (a,c,e) and the model based on z is in (b,d,f).

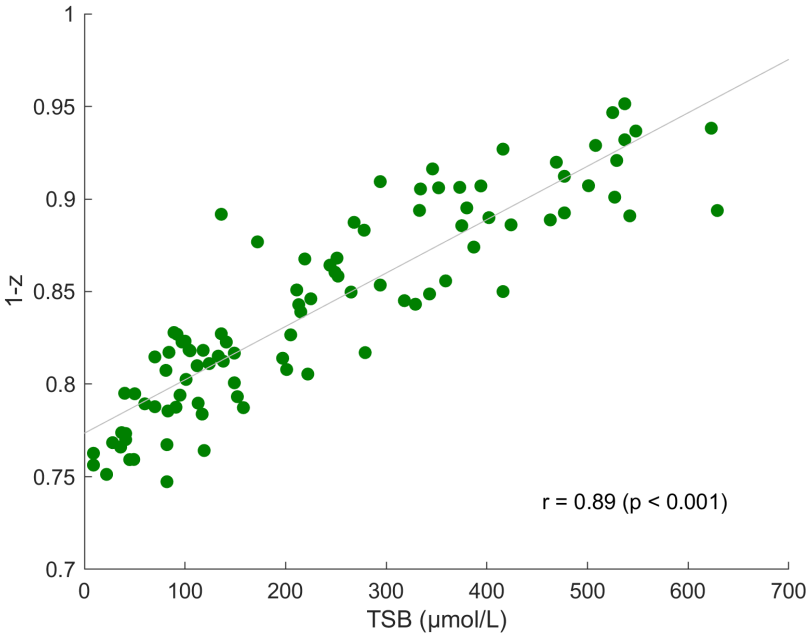


Figure 4.15: The selected device-independent yellowness metric values are scattered against the measured bilirubin for the S8 phone. The linear regression line is plotted in grey with the linear correlation also displayed.

4.3.4. Comparison against alternative processing

An alternative to the two step process proposed and demonstrated thus far is to determine a per-image mapping using a colour chart which must be included near to the region of interest in each image. This alternative was described in Chapter 3.4, where the per-image mapping accounts for both lighting and device dependence. The images required for this alternative method were captured along with the sclera-only images required for the proposed method to enable a comparison of the two processing approaches.

Figure 4.16 shows the resulting 1-z yellowness values in XYZ space for the two methods, plotted as a function of the measured TSB. The subset of data for which both types of images were available is presented for the S8 phone, and for both methods the mean yellowness was calculated across up to three available repeats. Results were similar in form for the Nexus phone so have not been included here. For both methods, the linear correlations are very strong, and visually the results are compatible despite their very different imaging and processing methods. This comparison shows that our proposed two step process produces similar results to the in-image colour chart method whilst maintaining a simpler image capture process.

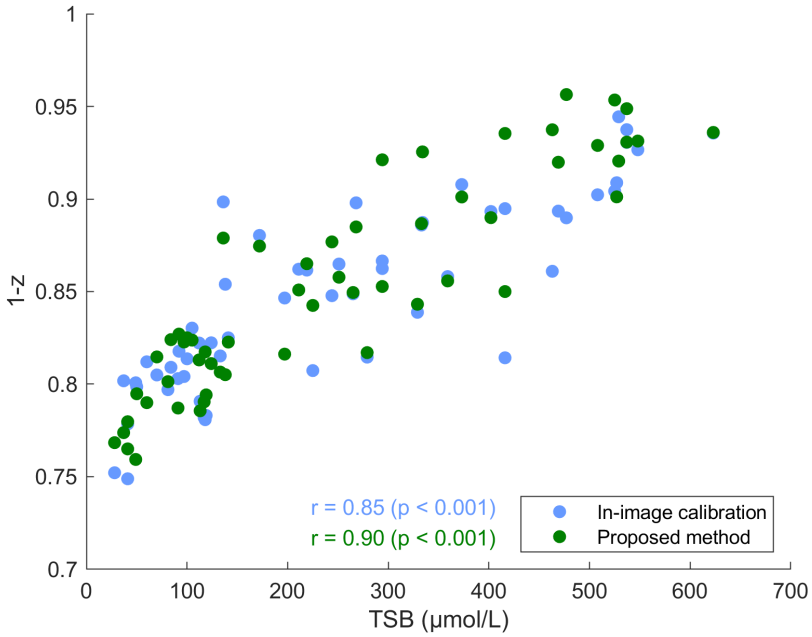


Figure 4.16: The extracted yellowness in device-independent space is plotted as a function of measured bilirubin when using the proposed method with the S8 phone (green) and a standard alternative in-image calibration method (blue). Linear correlation coefficients are presented in the respective plot colours.

4.4. Region of interest comparison

As discussed in Chapter 1.4.1, there are different available regions of interest to use in quantifying discolouration due to elevated bilirubin. The skin, specifically the forehead, is a common choice as it is flat, readily available and is relatively matte so does not suffer from issues of reflection [56]. However, the presence of melanin in the skin means that every person will have a different baseline skin colour. The sclera is another option as it is free from melanin - the human sclera is always white in healthy individuals [59,61]. There are drawbacks of using the sclera which include a small available area and the high occurrence of reflections owing to the layer of tear covering the eye. Another option, thus far not considered for assessing bilirubin, is to use the inside of the lower eyelid (palpebral conjunctiva), which is far less affected by melanin than the skin. Pulling down the eyelid to view the inside of the lower eyelid is however slightly uncomfortable. The lower eyelid is also prone to reflection, for a similar reason to the sclera. Here we present data comparing results from the sclera, our original proposed region of interest, to these two alternatives.

4.4.1. Sclera vs forehead

The image dataset gathered to compare our proposed processing method to the standard in-image calibration method was utilised for this analysis. Figure 4.17 shows an example image and region of interest selection. Where available, the region outlined in green between the eyebrows was used as it is typically smooth and less prone to shininess. Where there were significant blemishes, wrinkles or hair over this region, an available region within the large orange box was selected instead. The colour correction for both the skin and sclera regions was performed identically using the ColorChecker included in the image, as in Section 4.3.4 with theory presented in Chapter 3.4.

Figure 4.18 presents data from the S8 phone - the Nexus phone produced results with similar trends so has not been included here. Data using the sclera and skin are presented in (a) and (b) respectively, and scattered data has been colour coded by ethnicity in both cases. The sclera data yields a higher overall correlation ($r=0.85$ compared to 0.79, both $p<0.001$), with a wider range of chromaticities. The difference is not as great as originally expected, given the presence of melanin in the skin. However, as the colour coding demonstrates, the data set is highly biased towards Caucasian patients. In the data presented here, 80% of the image sets were of Caucasian patients, 15% of Asian patients and just 5% of Black patients. This dataset is not sufficient to draw significant conclusions on the issue of using the skin as a region of interest - with a more balanced patient cohort the confounding

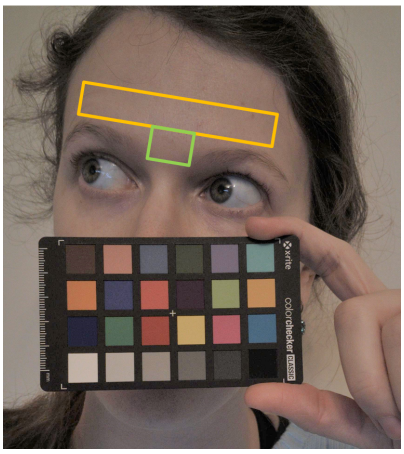
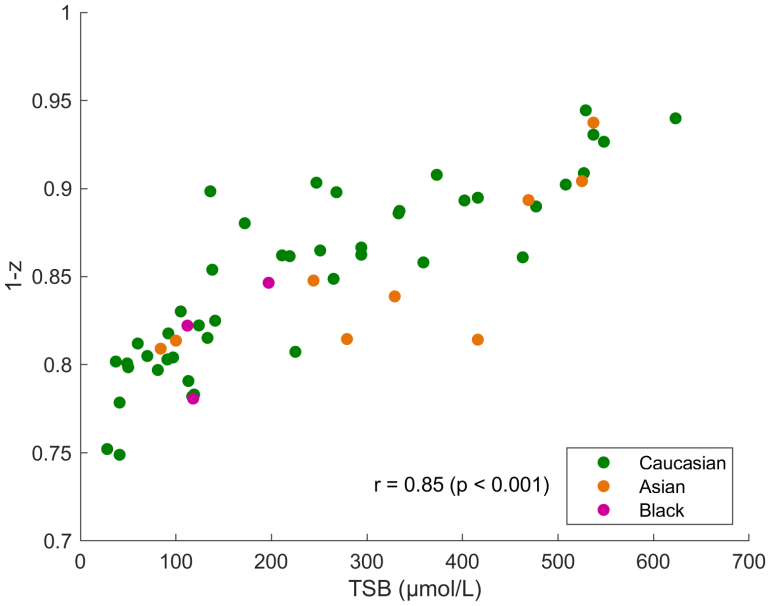
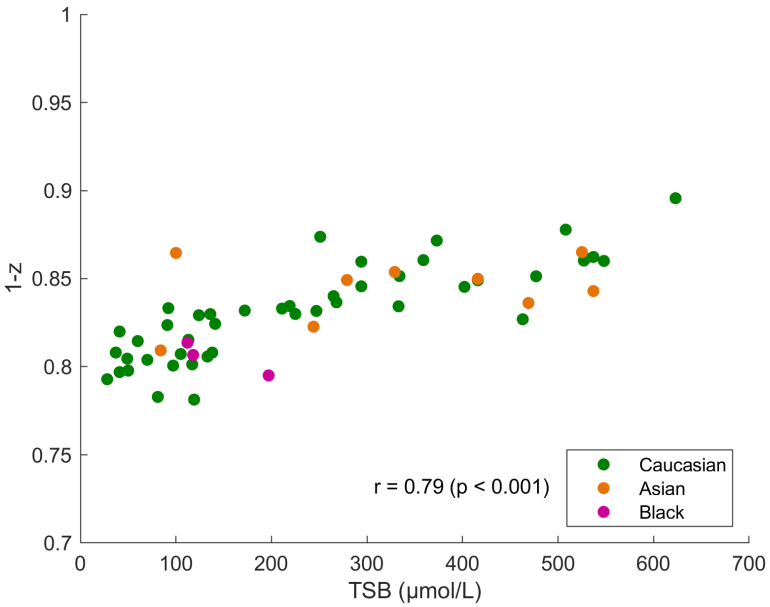


Figure 4.17: An example image of the format collected for the in-image calibration comparison, here used for the comparison to using skin as the region of interest rather than the sclera. Where available, the green boxed region was selected and when not viable a sub-region of the larger orange box was selected. More details are given in the text.

factor of melanin may become more apparent. The comparison of using the skin and sclera as regions of interest does confirm that as expected the sclera yields a higher correlation.



(a)



(b)

Figure 4.18: The extracted yellowness in device-independent space is plotted as a function of measured bilirubin using the in-image calibration method described in 4.3.4 for the S8 phone. (a) and (b) present data using the sclera and forehead skin respectively, with scattered data colour-coded by ethnicity and overall linear correlation coefficients presented.

4.4.2. Sclera vs lower eyelid

Image capture of the lower eyelid was incorporated into the image capture protocol, described in Section 4.1 as a standalone image pair enabling use of the proposed method for processing images of subtraction and a one-time calibration. In a study aiming to assess anaemia, two regions within the lower eyelid were compared: the inner or forniceal conjunctiva and the outer or palpebral conjunctiva [54]. The two regions are depicted in Figure 4.19. Both regions of interest were considered and correlations along with those from the sclera are presented in Table 4.5.

The work on anaemia found that the outer region was superior for assessing haemoglobin [54]. For the S8 phone, the correlation with bilirubin slightly decreased when using the inner region whereas there was a significant increase in correlation for the Nexus phone. Figure 4.20 shows scatter plots for the inner and outer regions of interest for both phones. The increase in correlation for the inner region for the Nexus phone seems to be caused by a skewing in result for low TSB levels. Given that the images captured using these two phones were of the same patients, with fewer images available on the Nexus phone, the cause of this is not clear. The main differences between the S8 and Nexus images are the lack of focus for the Nexus images and the larger region of possible partial specular reflection coming from the screen illumination. It is possible that the inner region is less affected by specular reflection with this phone.

In general, the correlations observed here for the lower eyelid are only slightly lower than for the sclera which is surprising given that the yellowing was far less noticeable to the human eye. The ability to use the lower eyelid for bilirubin estimation is interesting and may also be useful in other contexts.

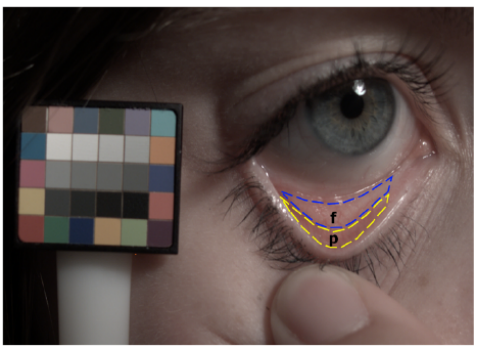


Figure 4.19: Lower eyelid image with the two standard lower eyelid regions of interest labelled - the palpebral (outer) and forniceal (inner) conjunctiva regions are labelled with a p and f respectively. Figure taken from [54].

ROI	S8 (n = 68)		Nexus (n = 44)	
	r	p	r	p
Outer	0.86	<0.001	0.68	<0.001
Inner	0.83	<0.001	0.80	<0.001
Sclera	**0.89**	<0.001	**0.87**	<0.001

Table 4.5: Pearson linear correlation coefficients of yellowness metrics using two different lower eyelid regions of interest against measured bilirubin levels. Correlations from the corresponding subset of sclera images are included for comparison.

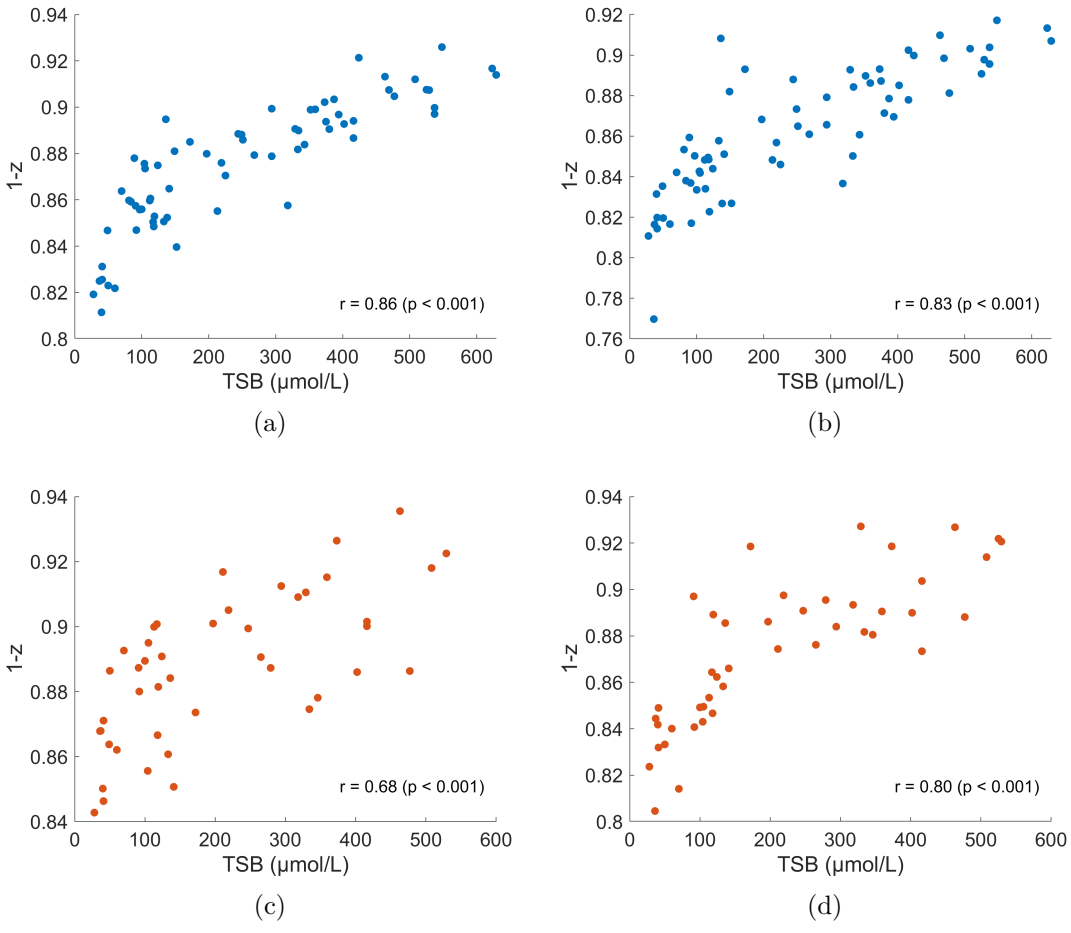


Figure 4.20: Correlations of yellowness with bilirubin for the S8 and Nexus phones are shown in (a-b) and (c-d) respectively. (a) and (c) show results using the outer region of interest, and (b) and (d) show those from the inner region of interest, with these regions defined in Figure 4.19.

4.4.3. Selection of sclera as region of interest

Table 4.6 presents a summary of the correlations achieved when using the three different regions of interest. Due to the reduced number of images available of the forehead and lower eyelid compared to the sclera, the result for the equivalent subset of sclera images is shown. A multivariate approach combining data from the sclera and lower eyelid is also included [121] - this combination was selected since it would be possible to extract data from both ROIs using a single image and so not disrupt the simple image capture.

As a single ROI, the sclera outperforms both the forehead and lower eyelid. The multivariate approach including the lower eyelid, however, provides a slight improvement. In general, for assessing bilirubin in this context the sclera imaging site is preferable to the lower eyelid for several reasons. Firstly, it is simpler and non-contact to take an image of the sclera since the patient simply needs to look to one side compared to needing to pull down the lower eyelid. Secondly, looking to the future, the sclera is a much more clearly defined region to automatically segment with its clear colour boundary. The minor improvement achieved by combining the data is outweighed by the added complexity and discomfort required to obtain lower eyelid images. We therefore select the sclera alone as the region of interest for all further analysis.

	S8			**Nexus**		
ROI	n	r	p	n	r	p
Forehead	57	0.79	<0.001	52	0.68	<0.001
Sclera - subset		**0.83**	<0.001		**0.78**	<0.001
Lower eyelid (best)	68	0.86	<0.001	44	0.80	<0.001
Sclera - subset		0.89	<0.001		0.88	<0.001
Sclera & lower eyelid		**0.91**	<0.001		**0.91**	<0.001
Sclera - total	93	0.89	<0.001	82	0.87	<0.001

Table 4.6: Comparison of correlations between yellowness and bilirubin level using different regions of interest for the two phones. In each case, the number of available images (n), the Pearson linear correlation coefficient (r), and significance (p) is presented. For the forehead and lower eyelid, correlations for the equivalent subset of sclera images are presented. The results for a multivariate approach combining data from the sclera and lower eyelid are also presented.

4.5. Clinical utility

4.5.1. Compatibility across devices

Returning to results generated using the proposed method with the sclera ROI, we will now focus on compatibility of results from different devices. Without compatibility, an entirely new set of paired images and blood bilirubin levels would need to be collected for each new phone being used for image capture. This is very clearly not feasible - care must therefore be taken to ensure that the results from different phones are compatible. In this case, a model linking extracted colour to bilirubin developed using one phone can be applied to data from a new phone. As described in Section 4.1, image capture was carried out in parallel using two different phones. This enables testing of the compatibility of results from two different phones, using the two modes of subtraction.

As an example, let us develop a model converting extracted yellowness to a predicted bilirubin level (referred to as SCB, Scleral-Conjunctival Bilirubin) using data from the S8 phone at two stages of processing. In both cases, a linear regression of yellowness to bilirubin is used. The first model is developed using the data presented in Section 4.3.2 using the 1-b metric, where ambient subtraction and filtering have been applied but without application of the device-specific mapping. This situation often occurs in pilot studies, where care is not taken to remove device dependence. The second model is developed using the 1-z metric after applying the device-specific mapping, at which point the data should be less affected by both ambient light and the device itself. The results for the S8 phone are presented in blue in Figure 4.21 (a) and (c). The difference in accuracy with and without the colour mapping is negligible for this phone.

We then attempt to apply these two models to data collected using the Nexus phone, with processing appropriate to each model applied. If the first model is used, SCB levels for the Nexus are extremely different from both the S8 SCB and ground truth TSB. This is highlighted by the large number of negative predicted TSB values shown in red in Figure 4.21 (a), and the strong trend observed in the Bland Altman plot in (b). However, when the second model is used, results from the two phones are extremely similar, as shown in Figure 4.21 (c). The mean squared error of the SCB compared to TSB is no longer significantly different between the two phones (t-test, p=0.34, df=71). The Bland Altman plot in (d) is also greatly improved, with a minimal trend observed and a 50% reduction in the 95% confidence intervals of -98 to 96 compared to (b) with -218 to 231.

This data demonstrates that when using the full proposed method, including the device-specific mapping, results from different phones are compatible and hence a model developed using data from a single phone can be applied to other phones in the future. The other extremely positive finding shown in Figure 4.21 (c) is the good match between SCB and TSB for both phones across a very wide range of TSB levels, with root mean squared errors of 62 and 97µmol/L for the S8 and Nexus phones respectively. Healthy levels of bilirubin are between 0 and just 17µmol/L, so to maintain a good correlation beyond 500µmol/L is very promising.

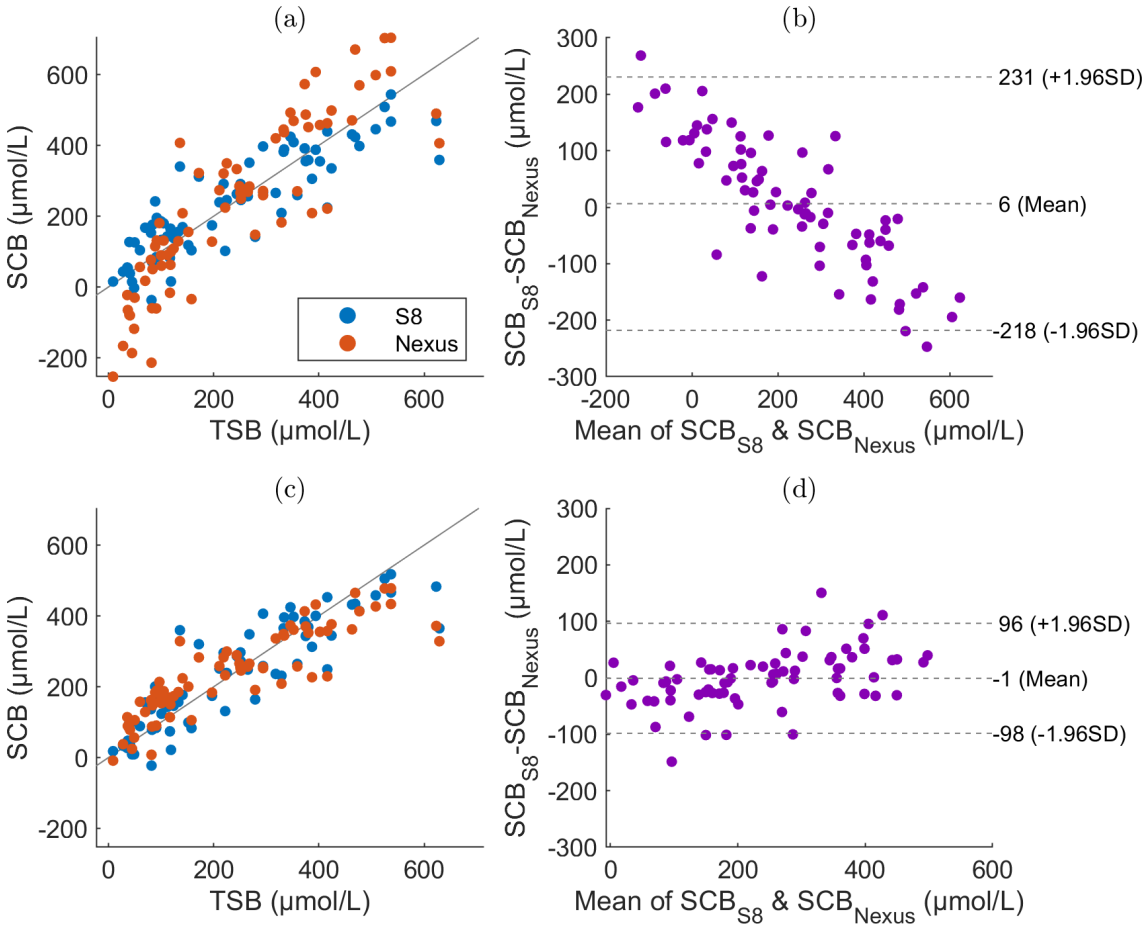


Figure 4.21: Scleral Conjunctival Bilirubin (SCB) predictions are plotted as a function of blood bilirubin levels (TSB) for models developed at two stages of processing. (a) S8 (blue) and Nexus (red) results for a model developed using S8 data after ambient subtraction. Data from both phones had the same level of processing applied as for model development. (c) S8 (blue) and Nexus (red) results for a model developed using S8 data after ambient subtraction and application of the device-specific mapping. Data from both phones was processed accordingly before application of the model. (b) and (d) Bland Altman comparisons of the data from the two phones in (a) and (c) respectively.

4.5.2. Longitudinal trends

As well as cross-sectional data, additional image sets were collected to allow longitudinal analysis. Where possible, repeat sets of images were captured for each patient, leaving a gap of at least three days between captures to allow the bilirubin level to change. As depicted in Figure 4.1, after processing there were 29 patients with repeat image sets available for the S8 phone and 22 patients for the Nexus. Overlapping data means displaying it all at once in this form is hard to understand so an illustrative subset is presented in Figure 4.22 for the S8 phone. A plot of all the longitudinal data for both phones is shown in Figure 4.23, with changes in TSB plotted as a function of changes in yellowness. Where more than two longitudinal points existed, each consecutive pair has been treated as an independent change.

In the previous section, compatibility of results from different phones was demonstrated. Here, longitudinal data pushes results one step further. Visual inspection of Figure 4.23 (a) and (b) highlights that the two phones produce slightly different longitudinal results. This is confirmed by the linear correlation coefficients for the two phones - 0.72 and 0.53 for the S8 and Nexus respectively (both with $p<0.001$). It is not immediately clear why the longitudinal data from the two phones should be different. It is possible that the lower image quality for the Nexus phone becomes relevant when trying to move beyond cross-sectional data. Images had lower resolu-

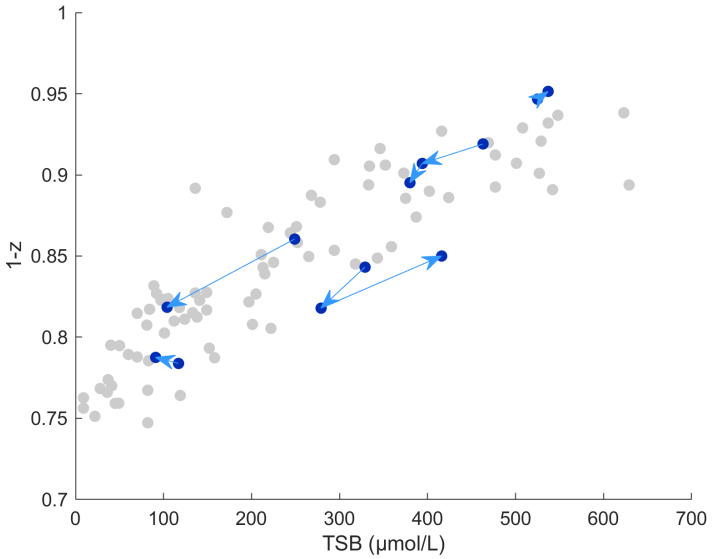
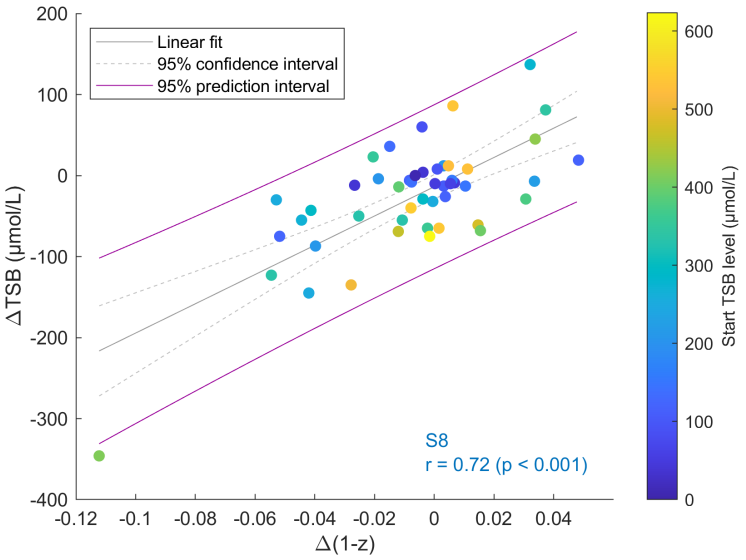
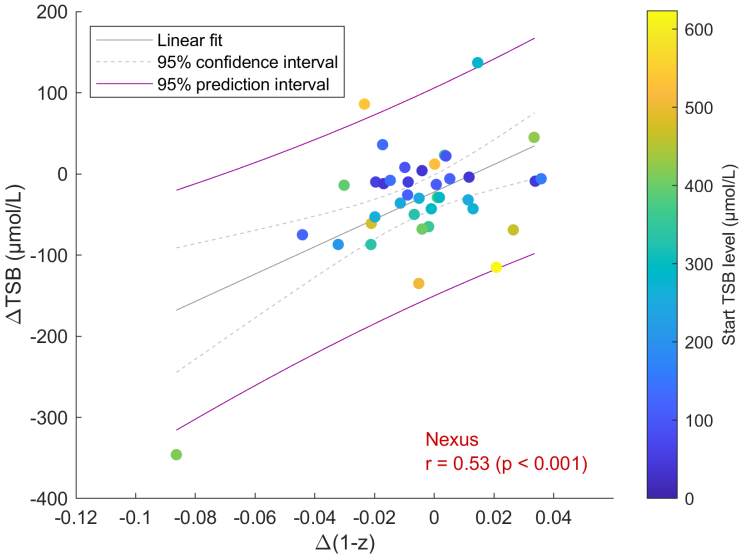


Figure 4.22: Extracted yellowness for all image sets for the S8 phone are plotted in grey as a function of the measured bilirubin. Longitudinal data for five representative patients are included in navy, with arrows connecting the points to show whether the bilirubin level increased or decreased.



(a)



(b)

Figure 4.23: Longitudinal change in measured bilirubin is shown as a function of longitudinal change in yellowness for images captured using the S8 and Nexus phones in (a) and (b) respectively. The colours of scattered points are based on the initial TSB level, with a colourbar to the right of each plot, highlighting that larger changes typically occurred for patients with higher TSB levels. The linear correlation coefficient and p-value for each plot is shown in the top right corner, with the regression line plotted in dark grey. The 95% confidence and prediction intervals are shown with dashed grey and solid purple lines respectively.

tion and were typically blurred due to a fixed focus front facing camera. The use of the screen as the flash illumination led to larger areas of partial specular reflection which were not always removed at the filtering step.

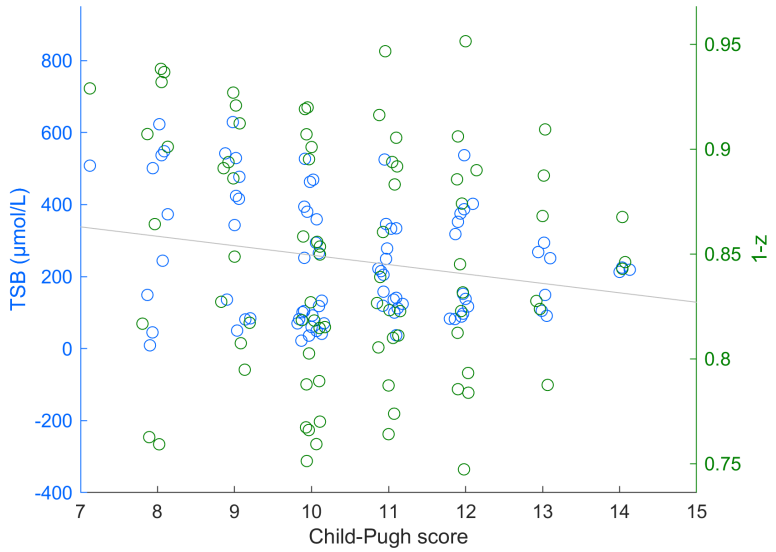
When we consider the percentages of longitudinal changes which follow the general trend - positive yellowness change for positive TSB change and vice versa - the two phones perform similarly. For the S8 phone, 31/48 (65%) follow the trend, with 24/39 (62%) on trend for the Nexus. When only changes in bilirubin above 50μmol/L are considered, 15/19 (79%) are on trend for the S8 and 11/14 (79%) for the Nexus.

In order to consider longitudinal accuracy more generally, 95% confidence and prediction intervals were calculated and are shown in Figure 4.23. The confidence interval gives the range into which 95% of the mean values for predicted changes in bilirubin will fall. This interval is not appropriate for considering individual measurement accuracy and is therefore less relevant. The prediction interval, on the other hand, gives the range that 95% of new predicted changes in bilirubin will fall into. For the data presented in Figure 4.23, the average prediction bound widths are 210 and 267μmol/L for the S8 and Nexus respectively. This means that for a future measurement of change in yellowness using the S8 phone, the change in bilirubin would lie in a 210μmol/L range. The overall correlation of longitudinal data demonstrates that tracking longitudinal changes should be possible, however these wide prediction bounds highlight that there are still improvements to be made.

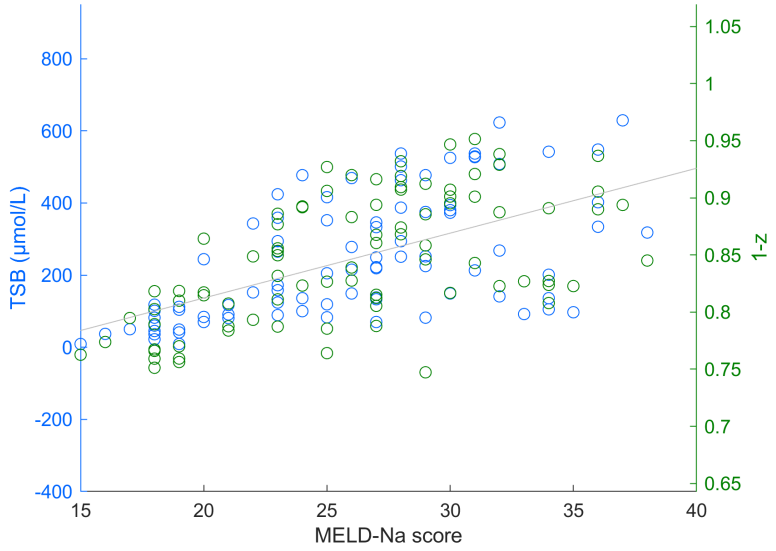
4.5.3. Association with clinical scores and outcome

We have demonstrated a strong correspondence between the yellowness and TSB both cross-sectionally and longitudinally. In clinical practice, various different clinical scores are used to assess patients' condition and predict mortality over different time periods. The four scores commonly used for this group of patients were introduced in Chapter 1.2. A key measure of success of a non-invasive bilirubin surrogate measure is whether it behaves similarly to TSB in comparison to these scores as well as directly compared to TSB.

Figure 4.24 shows scatter plots of the TSB and yellowness with each of the clinical scores for the S8 phone. Both phones gave similar results so the S8 phone, with a larger amount of data available, was chosen. For each score, separate vertical axes are used to display the TSB and yellowness with axis limits chosen such that the data overlaps, based on aligning the linear best fit line for each metric.

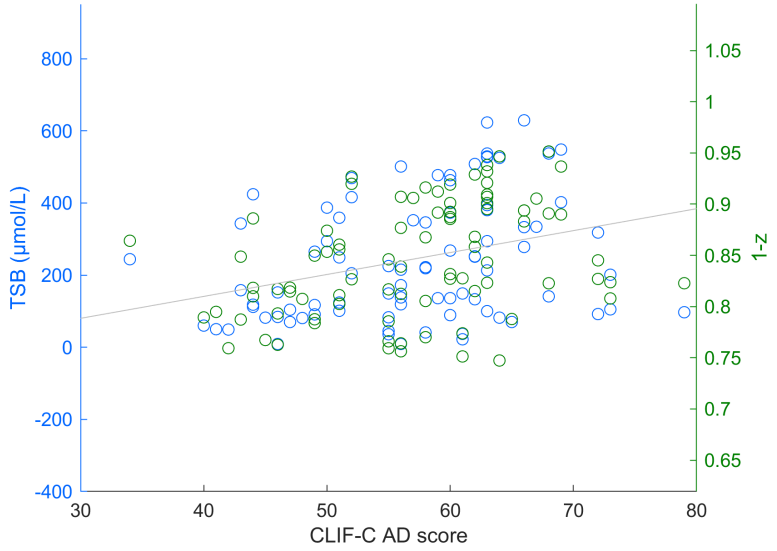


(a)

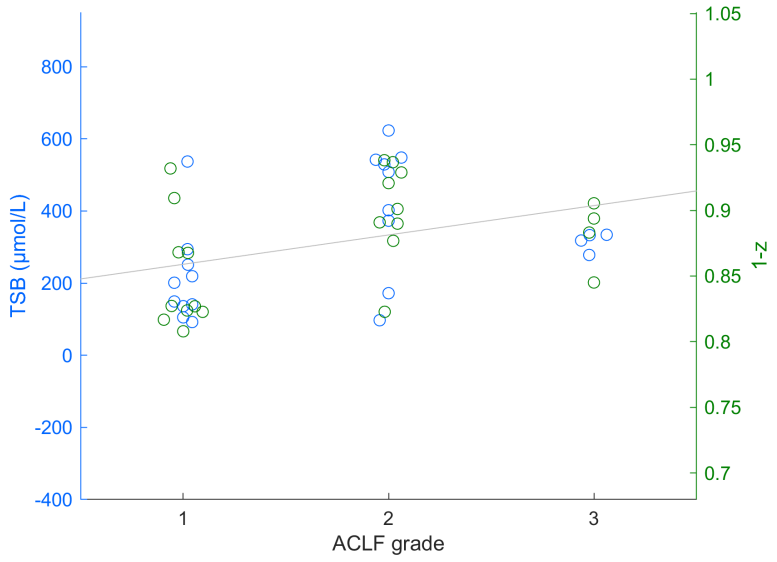


(b)

Figure 4.24: Associations of blood bilirubin level (TSB) and smartphone-extracted yellowness (1-z metric) with four common clinical scores. Continued on next page with further detail in the caption.



(c)



(d)

Figure 4.24: Continued from previous page: Associations of blood bilirubin level (TSB) and smartphone-extracted yellowness (1-z metric) with four common clinical scores. In each case, TSB data is shown in blue and uses the left-hand vertical axis, and the yellowness is shown in green and uses the right-hand vertical axis. The yellowness vertical axis has been set in each case such that the linear regression line (solid grey line) for both metrics is perfectly overlapping. To aid visibility of the data in (a) and (d) where the range of score values was low, jitter has been added.

The Spearman correlation coefficient was used to assess associations, rather than the Pearson coefficient. This was due to the categorical nature of, for example, the ACLF grade, and the desire to include correlations that were not specifically linear. The Spearman coefficient is also appropriate when data is not normally distributed. The correlation coefficients for the data in Figure 4.24 and associated p-values are shown in Table 4.7.

The correlations shown in Table 4.7 for each score are very similar for TSB and yellowness, with no discrepancy over significance and similar coefficient values. Inspection of Figure 4.24 confirms the similarity, and for MELD-Na and CLIF-C AD shows that these highly significant correlations are due to a real trend and not isolated outliers. The correlation of both TSB and yellowness with the AD score is interesting since the AD score does not explicitly incorporate bilirubin. In a reverse situation, the lack of correlation with the Child-Pugh score is interesting since it does depend directly on bilirubin. In this case, the lack of association is likely due to the ceiling effect mentioned in Chapter 1.2 - increasing bilirubin level above 50 μmol/L does not change the score. For all scores, however, the key finding here is not the significant correlation, or lack thereof, with TSB or yellowness. Rather it is that, as hoped, the non-invasive measure of bilirubin maintains the same associations with clinical scores as the blood test measured bilirubin.

As well as tracking associations with clinical scores, clinical outcome is also important. Here, this is assessed based on length of hospital stay, in-hospital death, and overall mortality. For all three measures of outcome, since they are binary, longitudinal data was excluded and only the first measure of TSB or yellowness was used for each patient. The analysis was restricted to those patients who had their first set of images collected within three days of admission, to avoid introducing the influence of in-hospital treatment. The Spearman correlations for the length of hospital stay for TSB and yellowness were 0.39 and 0.44 with p-values of 0.08 and

Score	TSB		Yellowness	
	ρ	p	ρ	p
Child-Pugh	−0.14	0.20	−0.13	0.26
MELD-Na	0.63	<0.001	0.59	<0.001
CLIF-C AD	0.35	<0.001	0.40	<0.001
ACLF grade	0.46	0.03	0.41	0.05

Table 4.7: Spearman correlations (ρ) and significance levels (p) for the ground truth TSB and smartphone-measured yellowness from the S8 phone, after subtraction and device-specific mapping, with four clinical scores.

0.05 respectively. There was no significant difference in either admission yellowness or TSB between those who died or received a liver transplant within 28 (p=0.90 and 0.62 respectively) or 90 days (p=0.92 in both cases). As with the clinical scores, these similar findings for TSB and yellowness demonstrate consistency.

The ability of TSB, yellowness and clinical scores to predict in-hospital death was tested by using the area under a ROC curve (AUROC). In this context, the AUROC gives the probability that a randomly selected patient who died in hospital had a higher admission TSB/yellowness/clinical score than a randomly selected patient who survived the hospital stay. The respective ROC curves are shown in Figure 4.25. The AUROCs are presented in Table 4.8, along with the 95% confidence interval ranges. Also shown in Table 4.8 is the p-value for the DeLong test of equality of areas under ROC curves for the yellowness compared to each of the other metrics [122]. Despite the variation in areas, the test shows no significant difference between the predictive power of the yellowness and any of the other metrics.

The results presented in this section demonstrate that not only does our smartphone measure of yellowness track bilirubin directly extremely well, but associations between yellowness and clinical outcomes and scores are the same as for bilirubin levels. This suggests that the smartphone measure may have clinical utility.

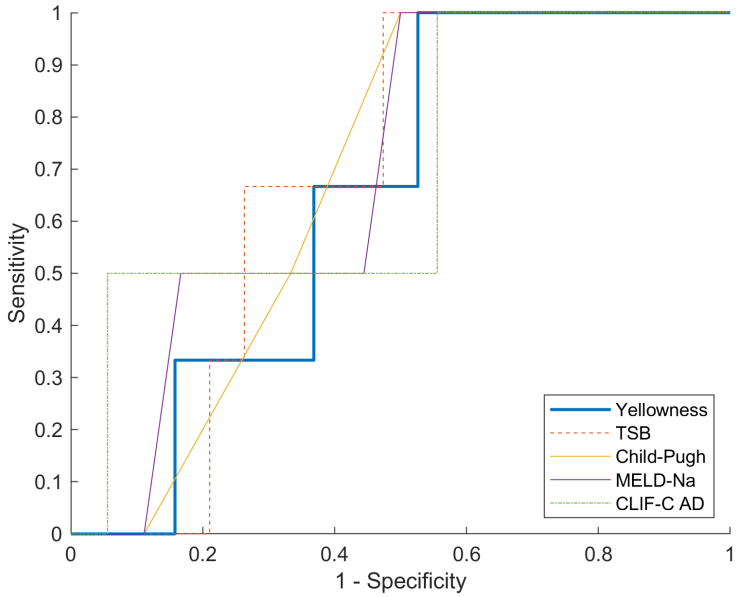


Figure 4.25: Receiver Operating Characteristics (ROC) curves for the prediction of in-hospital mortality for yellowness, TSB, and three common clinical scores. The areas under these ROC curves are presented in Table 4.8.

Metric	AUROC (with 95% confidence intervals)	p-value
Yellowness	0.65 (0.29 - 1.00)	-
TSB	0.68 (0.33 - 1.00)	0.65
Child-Pugh score	0.68 (0.25 - 1.00)	0.78
MELD-NA score	0.69 (0.26 - 1.00)	0.88
CLIF-C AD	0.69 (0.26 - 1.00)	0.90

Table 4.8: Area Under the Receiver Operating Characteristic curve (AUROC) for yellowness, TSB and three common clinical scores for the prediction of in-hospital death. The p-value for the DeLong test of equality of areas under ROC curves for yellowness compared to each other metric is also presented [122].

4.6. Assessing haemoglobin?

4.6.1. Motivation

As discussed in Chapter 1.3, clinical scores designed to assess a patient's condition are based on multiple factors. For a home monitoring system to be clinically useful it should therefore be based on multiple metrics, whilst maintaining a low level of complexity and equipment required. Possible contenders for such metrics, alongside the bilirubin level, are monitoring of heart rate variability [123] or variation in weight for ascites [27]. Another option could be the haemoglobin level.

Anaemia is very common in patients with chronic liver disease, due to a variety of factors, and has been found to be associated with higher MELD scores (an indicator of severity of liver disease) [124]. The presence of anaemia in patients with cirrhosis has also been found to be a predictor for the development of ACLF and for increased mortality [125]. Separately, work using smartphone imaging to assess haemoglobin level in other patient groups exists. These studies have used a variety of imaging sites, including the finger [126], nailbed [127], and lower eyelid [54, 128]. We therefore decided to try to look at assessing haemoglobin level for this patient cohort via smartphone images.

4.6.2. Redness metric

In general, haemoglobin absorbs in the green region of the spectrum and reflects in the red [129]. Two natural options for redness metrics, following the same logic as with bilirubin, would therefore be red and green chromaticity. As a reminder these are defined as $R/(R+G+B)$ and $G/(R+G+B)$ respectively. According to the absorption properties of haemoglobin, we would expect a positive correlation of

r with haemoglobin and a negative trend of g with haemoglobin. A third redness metric option is the erythema index (EI), proposed in [129] as

$$\text{EI} = ln\left(\frac{R}{G}\right) \tag{4.1}$$

where erythema is a term referring to reddening of the skin. Note that the fraction means that similarly to chromaticity this metric also removes dependence on brightness, and like r we would expect a positive correlation of EI with haemoglobin. For this preliminary work, we will remain in the device specific colour space and consider all three redness metrics.

4.6.3. Lower eyelid vs sclera

As discussed in Section 4.4.2, previous work looked at quantifying haemoglobin from images of the lower eyelid [54]. Two regions within the lower eyelid were compared: the inner or forniceal conjunctiva and the outer or palpebral conjunctiva. The two regions are depicted in Figure 4.19. The outer region was found to yield higher correlations, so this region of interest was considered here [54]. As described in Section 4.1, both lower eyelid and sclera image pairs were captured. Since the sclera images were also available, the sclera was considered as an alternative image site for assessing haemoglobin.

Results from the two phones showed similar trends, so here we present data from the S8 phone which had more image pairs available for analysis. For both sclera and lower eyelid results, data presented were obtained by taking a mean over up to three image pairs. The linear correlation coefficients (denoted by r) and p-values for the three redness metrics and two imaging sites are presented in Table 4.9. These results are after applying subtraction - before subtraction, in all cases no significant correlations were found. The direction of the correlations in Table 4.9 are as expected, however it is interesting that the correlations are slightly stronger for the

S8 **Redness metric**	**Lower eyelid** (n = 68)		**Sclera** (n = 97)	
	r	p	r	p
r	0.28	0.02	0.29	0.004
g	−0.32	0.007	−0.37	<0.001
EI	0.30	0.01	0.35	<0.001

Table 4.9: Correlations of redness metrics using lower eyelid and sclera regions of interest with measured haemoglobin levels. A correlation (denoted by the Pearson linear correlation coefficient r) was deemed significant if $p<0.05$

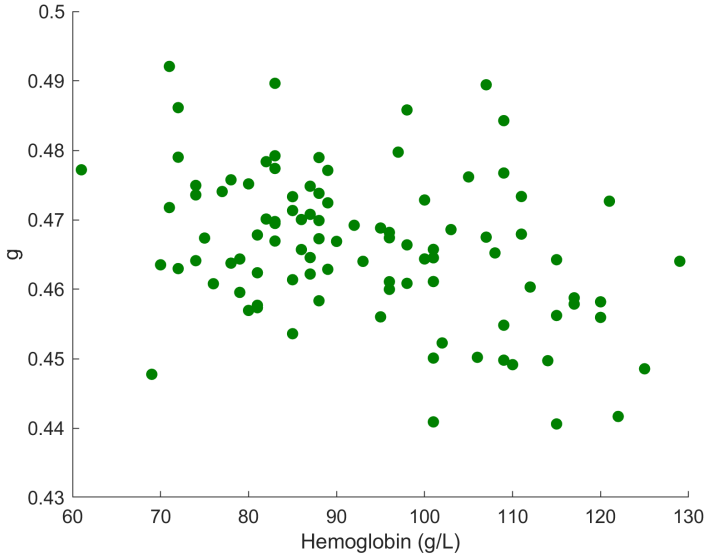


Figure 4.26: The mean green chromaticity (g) extracted from images of the sclera, plotted against the blood test haemoglobin level. The linear correlation coefficient for this data is -0.37 ($p<0.001$).

sclera images than the lower eyelid ones. For both regions of interest, the green chromaticity metric yielded the highest correlation.

As an example, the green chromaticity from the sclera image set is presented in Figure 4.26 as a function of haemoglobin level. It is known that for this patient cohort, there are two competing colour changes occurring: a variation in yellowness due to bilirubin level and a variation in redness due to haemoglobin level. Therefore it is important to verify if the trend shown in Figure 4.26 is truly due to variation in hemogobin level, or if the bilirubin levels are interfering. A slight significant correlation is observed between the blood test measured TSB and haemoglobin ($r=0.21$, $p=0.04$), however for the same images of the sclera no significant correlation is observed between the green chromaticity and bilirubin level ($r=0.18$, $p=0.08$). These results suggest that the trend observed is indeed due to varying haemoglobin level.

Overall the data presented in Table 4.9 and Figure 4.26 have much weaker correlations than those observed for yellowness and bilirubin. This is unsurprising since the slight variation in redness is much less than the change of white to quite saturated yellow caused by increasing bilirubin level. The results presented here are also weaker than presented by Collings et al [54]. This may be due to a reduced range of haemoglobin values observed for this patient cohort compared to the Collings paper. The presence of bilirubin may also be reducing correlations. For example, when the correlation of green chromaticity with TSB is calculated when using fil-

tering (specifically aiming to remove the blood vessels), a statistically significant positive correlation is observed (r=0.28, p=0.006). This positive correlation makes sense - for increasing bilirubin the blue channel value proportionally decreases and hence the green chromaticity value increases. Since this competing trend could be worsening results, a reverse filtering approach was tested where specular reflections were still removed but then the vessels were kept and the remaining sclera excluded. This did not improve results, likely as the precision required is high to maintain consistency between the flash and no flash images.

Rather than looking at assessing haemoglobin level directly, some papers have simply tried to classify patients as anaemic or not based on redness. The World Health Organization categorises patients as being anaemic when they have a haemoglobin level below 130 g/L or 120 g/L for men and women respectively [130]. The range of haemoglobin levels for this cohort was 69 - 129 g/L for male patients and 61 - 115 g/L for female patients. All of these patients would therefore be classed as anaemic. An alternative would be to look at degree of anaemia (mild, moderate, severe), however as previously discussed the correlations found here are weak and so attempts to do this yielded very poor results.

We have found significant correlations for all three redness metrics and both imaging sites, with the strongest results being from the green chromaticity and the sclera. These initial results suggest that it may be possible to assess haemoglobin level using this simple imaging technique, however the current low levels of correlation mean that it is not a useful addition to the system at this time.

5. Insight from sclera spectral information

5.1. Why move beyond RGB?

Based on the high quality results produced by the smartphone imaging methodology presented in this work, the rationale for trying to obtain full spectral data may not seem clear. The main aim is still to develop a cheap and portable tool, so permanently changing the methodology to obtain spectral data would be counterproductive. However, there could be valuable insight gained by gathering some spectral information that could help improve the current system. For example, spectral data could help to determine the cause of the plateau in measured colour observed for high bilirubin levels - whether this is due to the sensitivity of the camera sensors or a physical saturation of the sclera. The current yellowness metric has been shown to yield a very high correlation with bilirubin, but perhaps the realities of sclera colour changes could suggest an even better metric or colour space for this application. The information could be used to help design a custom colour chart with relevant yellows included for the one-time calibration step. Finally, work has been done to design pre-filters making cameras more colorimetric - to shift spectral sensitivities closer to the XYZ colour matching functions [131, 132]. Longer term, a similar principle could be applied to help a camera be more sensitive to changes in bilirubin, based on knowledge of the spectral change observed for increasing bilirubin.

5.1.1. Sclera reflectance factor

The spectral property of the sclera which we are interested in in this context is the reflectance. This is the property which determines how much light is reflected back from a surface, here the sclera, when exposed to spectrally uniform light. As described in Chapter 2.6.1, the surface reflectance is a key component, along with the spectrum of the light and the colour matching functions, in determining the resulting XYZ values. When considering something like a printed colour chart, it is simplest to measure the reflectance directly using a spectrophotometer. This contact method is clearly not possible for the sclera, meaning that more complex approaches must be taken to obtain spectral information.

To our knowledge, there has been just one *in vivo* bulk sclera reflectance measurement published as a small part of a different piece of research [133]. The resulting spectrum for a healthy adult is reproduced in Figure 5.1. Based on the general per-

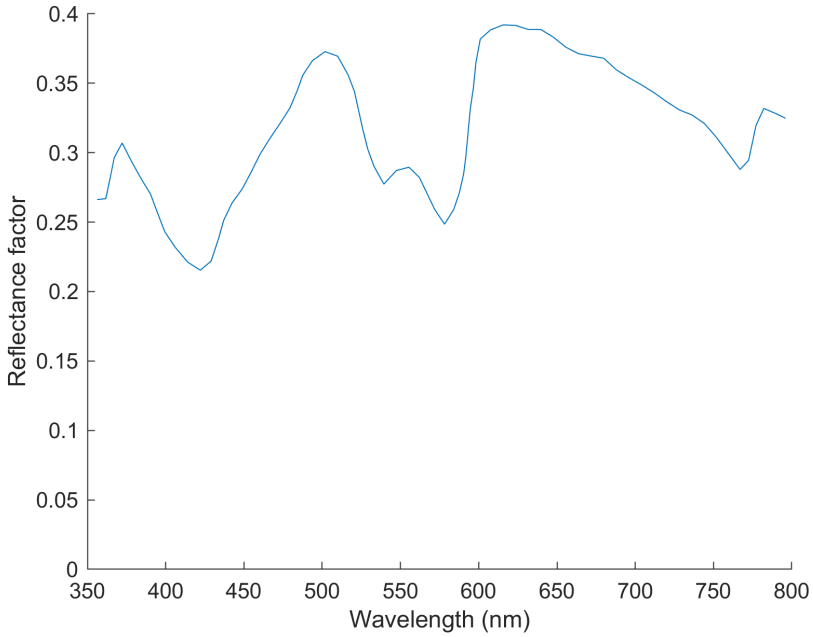


Figure 5.1: The reflectance of the sclera for a healthy adult. Data reproduced from Figure 4 of [133], extracted using [134].

ception that the healthy human sclera is white, it was expected that the spectrum would be relatively flat. However, the spectrum has prominent dips at around 540 and 575 nm. These correspond to absorption features of oxy-haemoglobin, due to the blood vessels on the sclera. The shape of the reflectance will be discussed in further detail in later sections. As far as we are aware, there have been no studies focussing on measuring the sclera reflectance for larger groups of healthy volunteers and particularly no measurements on individuals with elevated bilirubin.

5.2. Modelling

Modelling could be a good option to generate data since it would allow the time-consuming clinical data collection process to be side-stepped. Once a model has been set up, the sclera reflectance can be predicted for a range of different bilirubin levels. The majority of prior reflectance modelling work has been focussed on fitting models to data obtained via diffuse reflectance spectroscopy (DRS) and thus obtain estimates for different physical parameters in the tissue. This is a method which uses optical fibers to provide light to the area of interest and then uses more fibers to measure the light some distance away. Whilst we do not intend to carry out DRS measurements of the sclera, the aim is to leverage the models to give insight.

In general, the reflectance R is produced via some function combining the op-

tical absorption, μ_a, and reduced scattering, μ_s', coefficients of the tissue. There are different ways to produce and combine these coefficients. The reduced scattering coefficient depends directly on wavelength, whereas the absorption coefficient depends on the optical properties of chromophores found in the tissue, which in turn depend on wavelength. Chromophores are substances which have varied absorption in the visible range, and hence produce colour changes when present at different levels. The extinction coefficients for some typical chromophores found in skin and sclera are shown in Figure 5.2. Several relevant previously proposed analytical models for R will now be discussed in the context of modelling sclera reflectance.

To our knowledge, only one paper attempting to analytically model the sclera reflectance exists [135]. Other authors have questioned the validity of the model owing to its inclusion of neuroglobin in the expression for the absorption coefficient, a chromophore which has not been detected outside of neural tissue [136]. Despite very careful implementation of the model using the provided coefficients, it was not possible to replicate the results presented in the paper. The authors were contacted but did not respond. For this reason, the model was not considered further despite its relevance.

Owing to the lack of further sclera models, skin models were considered instead. These are popular in DRS so are more readily available. The intent was to start

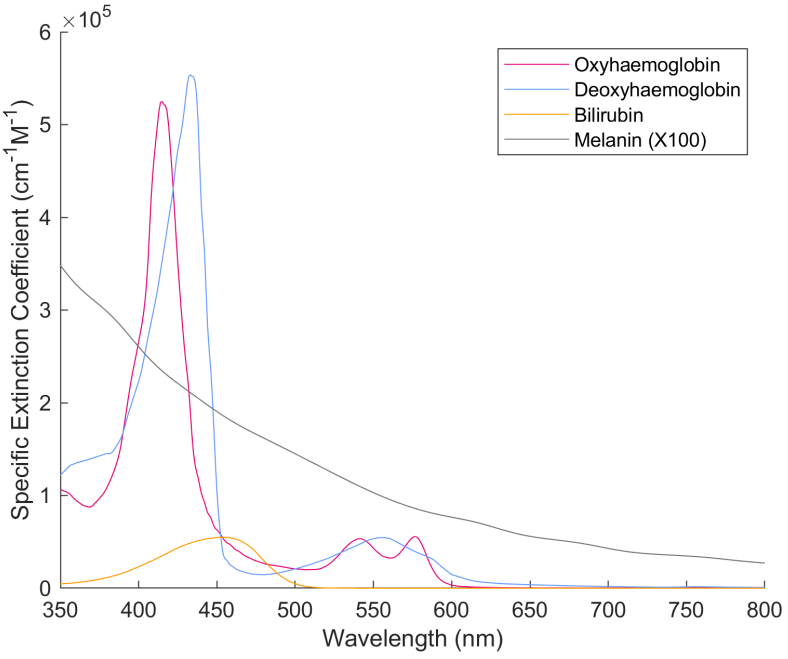


Figure 5.2: Extinction coefficients for oxy- and deoxyhaemoglobin, melanin and bilirubin. Data from [120]

with a skin model and modify it as appropriate for the sclera, for example by removing melanin and including bilirubin. We selected a popular, analytical model proposed by Zonios et al [137, 138], having tested several other analytical models [139, 140]. In this model, based on an integral assuming point delivery of light and collection over a uniform spot, reflectance is given by

$$R = R_0 \frac{\mu_s'}{\mu_s' + \mu_a} \Bigg(exp(-\mu z_0) + exp(-[1 + (4A/3)]\mu z_0) \\ - z_0 \frac{exp(-\mu r_1')}{r_1'} - [1 + (4A/3)] z_0 \frac{exp(-\mu r_2')}{r_2'} \Bigg) \tag{5.1}$$

with

$$\begin{aligned} r_1' &= (z_0^2 + r_c^2)^{1/2} \\ r_2' &= (z_0^2[1 + (4A/3)]^2 + r_c^2)^{1/2} \\ \mu &= (3\mu_a(\mu_a + \mu_s'))^{1/2} \\ z_0 &= \frac{1}{\mu_s'} \end{aligned}$$

where $R_0 = 0.7$ and $r_c = 4.0$mm based on experiments, and $A = 2.8$ based on the refractive index of tissue [137].

The reduced scattering coefficient is approximated as having a linear dependence on wavelength, based on Mie theory, and is given by

$$\mu_s'(\lambda) = \left(1 - \sqrt{\frac{d_0}{d_s}} \frac{\lambda - \lambda_{min}}{\lambda_{max} - \lambda_{min}} \right) \mu_s'(\lambda_{min}) \tag{5.2}$$

where d_s represents the effective scatterer size, $d_0 = 0.0625\mu$m is a constant, λ_{min} and λ_{max} are 460nm and 820nm respectively, and $\mu_s'(\lambda_{min})$ is another model fitting parameter [137]. As is typical, the absorption coefficient is expressed as a weighted sum of extinction coefficients

$$\mu_a(\lambda) = ln(10)[c_{Hb}\,(\alpha \epsilon_{HbO_2}(\lambda) + (1 - \alpha)\epsilon_{Hhb}(\lambda)) + c_{mel}\epsilon_{mel}(\lambda)] \tag{5.3}$$

where c_{Hb} is the haemoglobin concentration, α is the haemoglobin oxygen saturation, ϵ_{HbO_2} and ϵ_{Hhb} are the extinction coefficients of oxy- and deoxyhaemoglobin respectively, and c_{mel} and ϵ_{mel} are the melanin concentration and extinction coefficients respectively. Note the prefactor of $ln(10)$ to account for differences between absorbance and attenuation, which is typically implicitly assumed in DRS publications.

The Zonios model considered here focusses on wavelengths above 460nm, as with most DRS skin models. Since the main absorption feature for bilirubin is centred around 450nm, shown in Figure 5.2, care must be taken that the model is appropriate for use at shorter wavelengths before trying to include the impact of bilirubin. The main adjustment is the need to account for the fact that haemoglobin is found in blood vessels, not homogeneously distributed. This affects the absorption coefficient particularly for shorter wavelengths [141]. Rather than having to model the blood vessels explicitly, it is possible to apply a mathematically derived wavelength-dependent correction factor to a homogeneous model to account for the packaging of blood in blood vessels. Several correction terms have been developed. Here the simple analytical correction term originally proposed by Svaasand et al has been used to modify the expression for μ_a from Equation 5.3 to

$$\mu_a(\lambda) = ln(10)[c_{diff}\nu\left(\alpha\epsilon_{HbO_2}(\lambda) + (1-\alpha)\epsilon_{Hb}(\lambda)\right) + c_{mel}\epsilon_{mel}(\lambda)] \tag{5.4}$$

using the correction factor

$$c_{diff} = \frac{1 - exp(-2\mu_{a,bl}r_{vess})}{2\mu_{a,bl}r_{vess}} \tag{5.5}$$

where $\mu_{a,bl}$ is the absorption coefficient of whole blood, r_{vess} is the effective vessel radius here set to 0.5mm, and ν is the blood volume fraction [142–144]. Note that the correction factor is not applied to the homogeneously distributed absorber melanin.

Figure 5.3 shows example reflectance outputs of the model with and without the vessel correction. Note again the characteristic haemoglobin features, and the shallower more physical dip around 420nm when the vessel correction is included.

It is now possible to adjust the model to be appropriate for the sclera rather than the skin. We achieve this by modifying the expression for the absorption coefficient to include the effect of bilirubin rather than melanin

$$\mu_a(\lambda) = ln(10)[c_{diff}\nu\left(\alpha\epsilon_{HbO_2}(\lambda) + (1-\alpha)\epsilon_{Hb}(\lambda)\right) + c_{bili}\epsilon_{bili}(\lambda)] \tag{5.6}$$

Figure 5.4 presents model results incorporating bilirubin alongside real clinical data. Figure 5.4 (a) shows example reflectances for the adjusted model with varying levels of bilirubin. The changing shape of the reflectance data with increasing bilirubin makes sense, with the minimum appearing around the known peak absorption of bilirubin. Figure 5.4 (b) shows the xy values produced from this reflectance data,

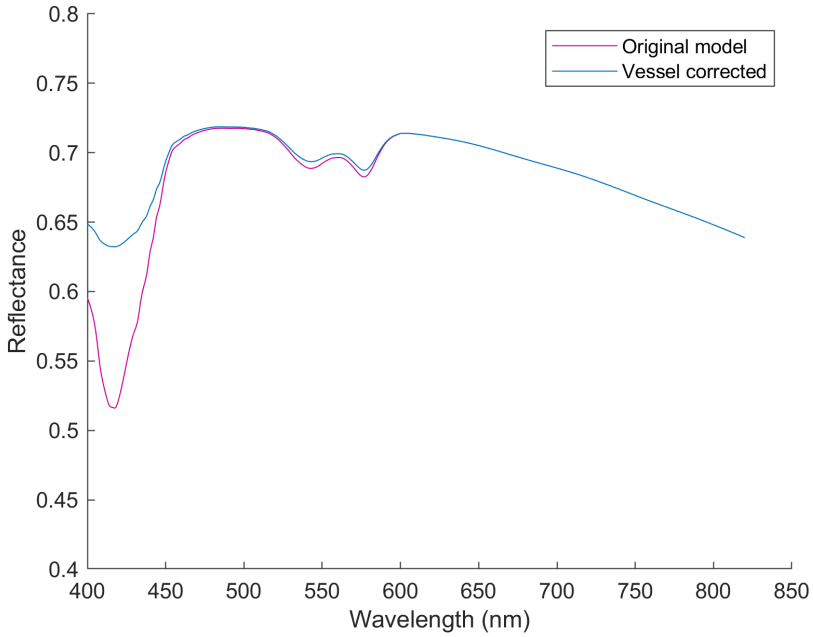
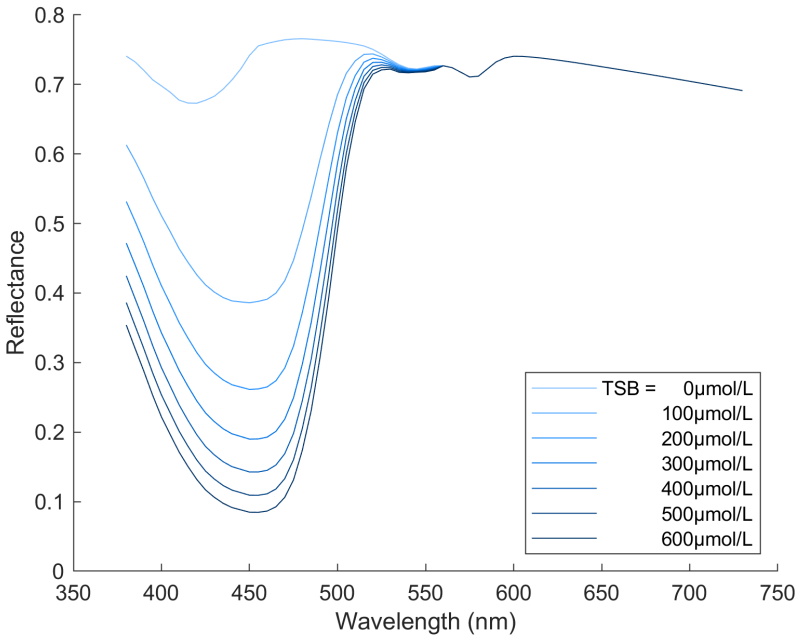
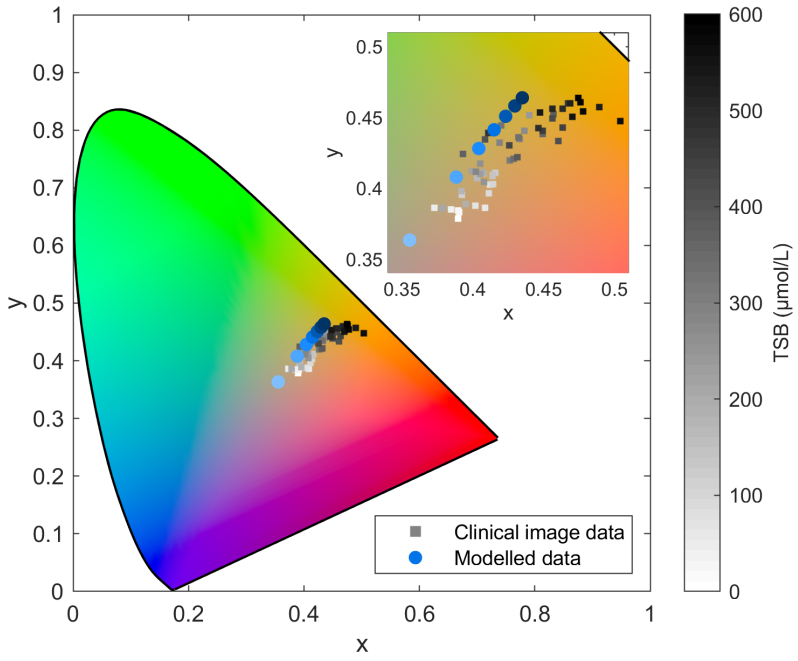


Figure 5.3: Example reflectance of healthy adult skin, using model parameters yielded for volunteer 6 from Table 1 in [137]

along with the clinical image data results. The simulated xy values appear in the same general area of xy space as the clinical image data results. The major issue with the model outputs is how non-linearly spaced the results are in xy space, visible too in (a). Whilst there is a question as to whether the image data remains linear for high bilirubin levels, such significant non-linearity for low bilirubin is certainly not observed in practice. Figure 5.4 (c) shows the standard yellowness metric (1-z) as a function of bilirubin for both modeled and clinical data. The non-linearity of the modeled data is apparent in this form too. This non-linearity, in combination with the imperfect overlap with the clinical image data, suggests that there are subtleties that this model is not capturing. To improve results, and so be in a position to be able to answer the original questions about the clinical image data, would require significant model development. New model development was deemed to be beyond the scope of this work.

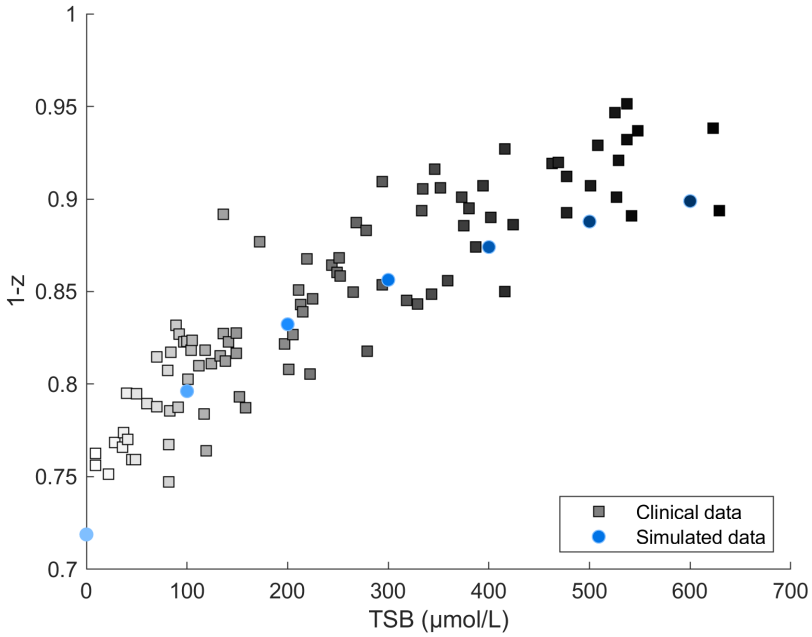


(a)



(b)

Figure 5.4: Modeled data compared to measured clinical data. Continued on next page with further detail in the caption.



(c)

Figure 5.4: Continued from previous page. (a) Example reflectance data produced by the model adjusted to include the contribution of bilirubin. (b) Simulated xy values for this data (circle markers with the same blue scale as in (a)), plotted alongside the xy values resulting from clinical sclera smartphone image data of patients with varying levels of bilirubin (square markers with a white-black colour scale). (c) Yellowness values (1-z) plotted as a function of bilirubin for modeled data (circle markers with the same blue scale as in (a)) and clinical data (square markers with the same white-black colour scale of bilirubin as in (b)).

5.3. Spectroradiometric measurement

Since it was not possible to obtain the desired results by modelling, the focus turned to trying to measure the reflectance of the human sclera in patients *in vivo*. The simplest way to measure reflectance is using a spectrophotometer. However, this approach is not appropriate here since it is a contact measurement. An alternative is to use a spectroradiometer, also known as a telespectroradiometer. This device measures, from a distance, the spectral radiance coming from a sample r_s. The reflectance factor of the region of interest is given by

$$R_s(\lambda) = \frac{r_s(\lambda)}{r_{PRD}(\lambda)} \tag{5.7}$$

where r_{PRD} represents the radiance of an identically irradiated perfect reflecting diffuser (PRD). The division is necessary to remove the effect of ambient light. Since we cannot measure the radiance for a PRD in real life, we can use instead

a white tile. In advance, the reflectance of the tile R_{WT} can be measured using a spectrophotometer, and then the radiance measured to yield

$$R_{WT}(\lambda) = \frac{r_{WT}(\lambda)}{r_{PRD}(\lambda)} \tag{5.8}$$

The two previous equations can be combined to yield a practical way to obtain the reflectance of a sample using a spectroradiometer

$$R_s(\lambda) = \frac{r_s(\lambda)}{r_{WT}(\lambda)} R_{WT}(\lambda) \tag{5.9}$$

In the lab, a high quality white standard such as Spectralon could be used in place of a white tile. In this case, the reflectance of the standard is near 1 across the visible spectrum so the sample reflectance can be obtained by a simple division of radiances. In general, however, it is safer to use a cheaper white tile and Equation 5.9 to avoid damaging the expensive white standard.

5.3.1. Experimental conditions

An example image of data collection with a volunteer is shown in Figure 5.5. A SpectraScan PR-650 spectroradiometer (Photo Research Inc.) was used for data collection, controlled via a laptop. The PR-650 records radiance at 4nm intervals from 380-780nm. A lens was used to reduce the area of sclera that was included in the averaging region. Measurements were carried out in a dark room, with blue-filtered tungsten lighting. The bare lightbulbs were used as it was possible to avoid the small bright spots of specular reflection on the sclera, whereas the use of a diffuser led to more extended regions of reflection. Tungsten lighting was used since it has a continuous spectrum over the visible region, compared to fluorescent lighting which has very sharp spikes and regions of very low intensity - the relative spectral radiances of these two light sources are shown in Figure 5.6.

A head and chin rest was used to stabilise the subject's position, as the natural motion of the head is otherwise too great to allow a consistent alignment. Once the PR-650 was aligned to the eye region, the subject was requested to look to the side and the alignment was tweaked to ensure a specular-free region of sclera was selected. The subject was asked to refrain from blinking for the couple of seconds required to capture the radiance, and was then asked to hold the white tile directly in front of their eye for a second measurement (not pictured). Equation 5.9 was used to determine the reflectance factor of the sclera based on these measurements and a prior reflectance measurement of the tile.

Figure 5.5: Image showing data collection of the scleral reflectance using a telespectroradiometer. A head and chin rest was used to stabilise the subject's position, and they were requested to look to one side to expose the sclera. The measurements were carried out in a dark room with illumination provided by filtered tungsten lightbulbs. A subsequent measurement using a white tile enables calculation of the sclera reflectance.

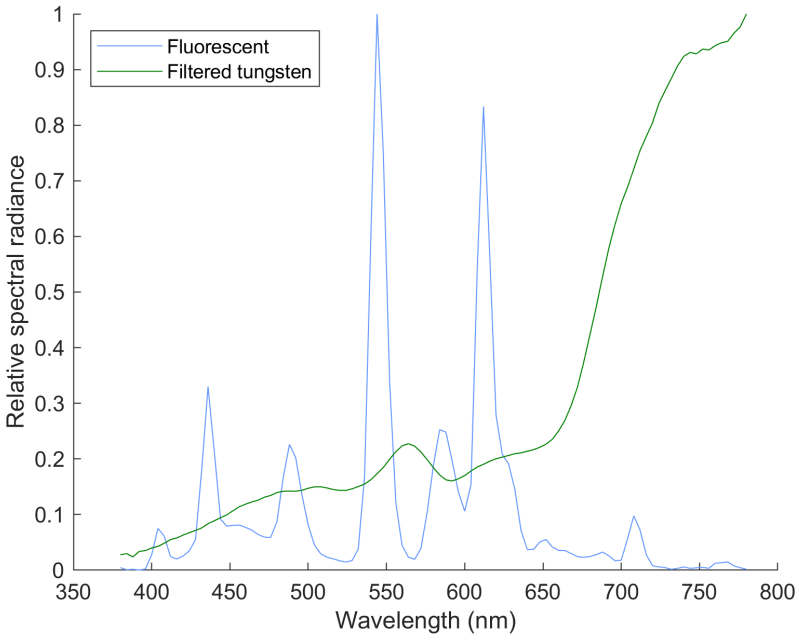


Figure 5.6: The relative spectral radiance of fluorescent (blue) and filtered tungsten lighting (green) are shown, highlighting the more appropriate smoother spectrum of tungsten.

5.3.2. Preliminary results

ColorChecker test

To ensure that the set-up was working correctly, test measurements were first taken using a ColorChecker in place of a human subject. The reflectance factor measured using a spectrophotometer was compared to the reflectance obtained via PR-650 measurement. Figure 5.7 shows a comparison of the two for patch 11 of the ColorChecker, a lime green colour. Relative results are presented since the absolute value of the PR-650 method is affected by the white tile being at a different distance from the device than the sample. The scaled reflectances have extremely similar shapes, giving confidence that the set-up is working correctly.

Healthy adult volunteers

Data was then collected on several healthy adult volunteers. An example result of the sclera reflectance is shown in Figure 5.8. The characteristic features of the oxyhaemoglobin spectrum are marked with vertical dashed lines, lining up with the features observed here. The form is similar to the one published spectrum presented by Palmer et al [133]. This preliminary result was very promising, however when repeat measurements were taken of the same volunteer's sclera during the same

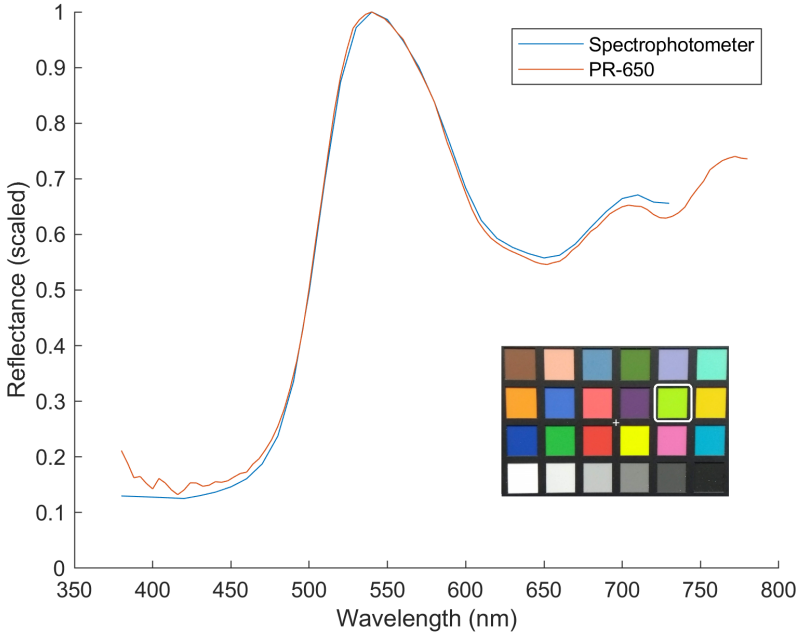


Figure 5.7: The relative reflectance of an example patch from the ColorChecker chart (inset, patch outlined in white) is shown for a standard spectrophotometer measurement and for the PR-650 method.

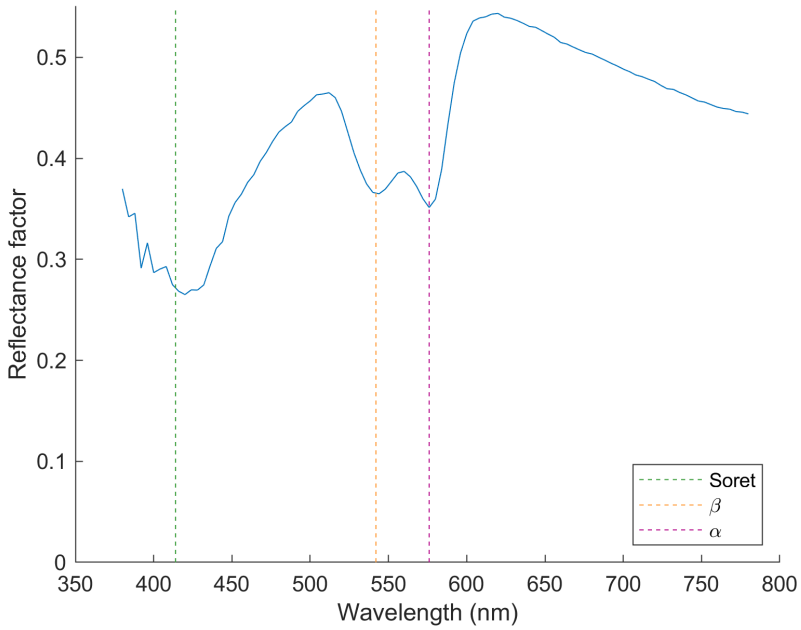


Figure 5.8: The measured sclera reflectance of a healthy adult volunteer is shown. Typical oxyhaemoglobin features are marked with dashed lines, coinciding with reflectance features.

measurement session significant variability was observed, with data shown in Figure 5.9. Relative values have been presented for the same reasons discussed for the ColorChecker test measurements. The cause of the variability is not immediately apparent, but possible reasons include different proportions of blood vessels included in the averaging area, different measurement angle to the sclera owing to its curvature, and variability introduced by operator positioning and coloured clothing.

Additional measurement sessions had been planned to collect more repeat data from volunteers. Using this data, it was hoped that a combination of averaging and careful region of interest selection would reduce the variability. The level of variation across a cohort of healthy volunteers would then be quantified, since we note that the sclera has been found to become slightly darker and more yellow with age [145, 146]. At this stage, once the measurement protocol had been further honed, data on patients with liver disease from the Royal Free Hospital with varying levels of bilirubin would have been collected. This would have allowed the impact of bilirubin on the sclera reflectance to be investigated. Unfortunately, due to the global Covid-19 pandemic no further data collection was possible. It is hoped that by sharing these preliminary results, others may be able to build on the work presented thus far.

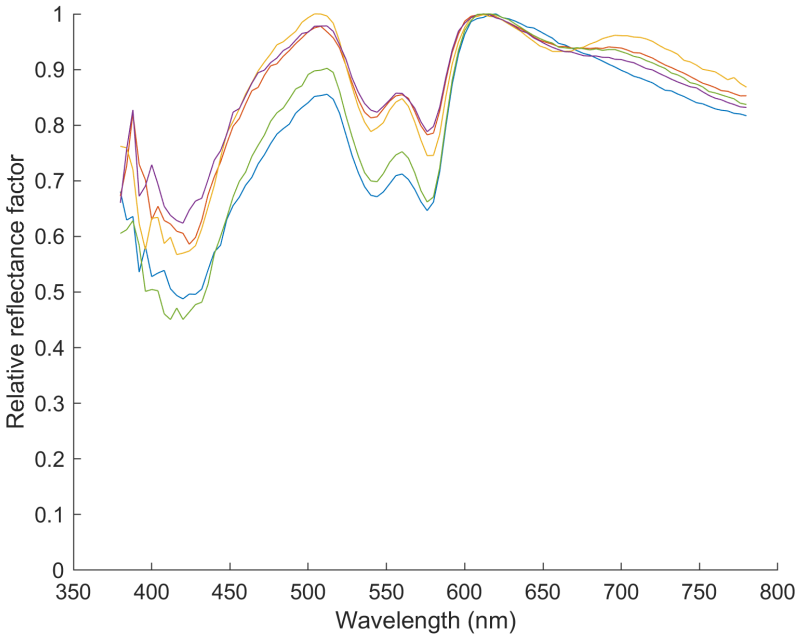


Figure 5.9: Five relative measurements of the reflectance of the same adult sclera taken during the same capture session are plotted, showing significant variability.

6. General conclusions

6.1. Summary

In this thesis, a system to enable non-invasive monitoring of chronic liver disease patients has been developed. As a starting point, the system focusses on assessment of the bilirubin level via the discolouration of the sclera. This biomarker was selected as a key indicator of liver function which also has the potential for non-invasive measurement. The goal of developing such a system is to target patients outside of hospital, where the conventional blood test approach for measuring bilirubin is not possible. For a home monitoring system to be viable it must be simple to use, cheap, and portable.

During the development of the system, significant care has been taken to maintain these viability criteria. The use of smartphone imaging to obtain quantitative measurements is a key factor. Smartphones are ubiquitous in society, and the cost of devices which have sufficient camera quality is ever decreasing. Smartphones are portable by design, and care has been taken to ensure that minimal additional equipment is required for the system to work.

Obtaining accurate colour measurements using smartphone images involves overcoming the influence of both ambient light and device dependence. When developing the processing, simplicity of use was an important consideration. Complex phone add-ons and challenging image capture protocols were therefore avoided. Use of ambient subtraction via flash/ no-flash image pairs, thus requiring no colour chart in shot, makes the sclera image capture step as simple as possible (see Chapter 2.3.1). Either the front or rear-facing camera can be used depending on the operator. The dependence on the device can be removed by carrying out a calibration step of imaging a colour chart. This step only needs to be carried out once per device, meaning that each user does not require a colour chart and that for day-to-day use the imaging is very straightforward.

6.2. Contributions and findings

6.2.1 Processing method

The use of an ambient subtraction technique to account for variation in lighting conditions was validated using a colour chart containing a wide range of colours, and for human volunteers under five different ambient conditions (see Chapter 3.1). To enable the collection of high quality data, a simple metric was developed based around an estimation of the subtracted signal to noise ratio (SSNR). The experimentally determined SSNR threshold of 3.4 allows an in-app recommendation that the images be re-captured if the ambient light is too bright or the phone is too far away.

The production of data which is independent of the device is key for generalisability. A model converting extracted colour to bilirubin must be device-independent to enable the use of other devices in the future. The process of mapping to a device-independent space was considered in detail (see Chapter 3.2). Approaches based on the information stored in the image metadata and on images of colour charts were compared. The classification accuracy across multiple phones was calculated for these approaches based on images of the DC colour chart. It was found that it is necessary to carry out a one-time device-specific calibration to obtain results of a practical accuracy level. The calibration can be one-time since after subtraction is applied, all data is under a fixed effective illumination. Traditionally, the phone would need to be held in a tripod to capture aligned images of the colour chart and a neutral grey chart to account for variation in illumination intensity across the chart. However, it was demonstrated that the use of an alternating least squares mapping approach enables just hand-held images of the colour chart to be used. This greatly simplifies the imaging process and reduces the likelihood of user error.

6.2.2 Direct scleral measurement

To add additional spectral insight, a methodology to measure the reflectance spectrum of the sclera was developed based on the use of a telespectroradiometer (see Chapter 5.3). Measurements were carried out on healthy adult volunteers producing results which were similar in form to modelling, demonstrating that the collection of data was possible. Further data collection was not possible due to the Covid-19 pandemic, however it is hoped that the methodology may be useful to others taking the work forwards.

6.2.3 Clinical database and filtering

A study was carried out at the Royal Free Hospital in London, collecting image data of patients with acute decompensation due to chronic liver disease using two different phones and recording detailed clinical information (see Chapter 4.1). It was found that for this adult population, there was a high proportion of blood vessels as well as different levels of specular reflection appearing on the sclera. Test samples demonstrated that these features could cause significant offsets in extracted colour values, hence a filtering algorithm was developed to remove affected pixels (see Chapter 4.2). The algorithm had a success rate of 93-94% on patient images. In device-specific space, an increased correlation between yellowness and bilirubin was observed when both subtraction and filtering were applied.

6.2.4 Processing validation

Different models of converting extracted device-independent sclera colour values to bilirubin levels were considered, ranging from single chromaticity metrics up to higher order multiple linear regressions (see Chapter 4.3). It was found that higher order models risked over-fitting based on the dataset size and did not yield significant increased accuracy. A simple metric based on Z chromaticity, in device-independent XYZ space, is therefore recommended for use with this population. Strong cross-sectional correlations of 0.87-0.89 ($p<0.001$) were obtained for the two phones when this metric was used (see Chapter 4.3.3). Additionally, results from the two phones were demonstrated to be compatible with one another using the proposed processing and yellowness metric (see Chapter 4.5.1).

Results from three regions of interest - the sclera, forehead, and lower eyelid - were compared (see Chapter 4.4). As expected, the sclera provided the best results owing to its lack of melanin. The lower eyelid has not been previously used to assess bilirubin level and produced a high correlation of yellowness and bilirubin. This region of interest may be of particular interest in other applications.

6.2.5 Clinical scores

Along with the key finding of inter-device compatibility with the proposed processing, the measured bilirubin and sclera-extracted yellowness metric were demonstrated to have the same associations with common clinical scores (see Chapter 4.5.3). The correlations and predictive ability of clinical outcomes such as hospital stay length, in-hospital death and overall mortality were also the same for both yellowness and bilirubin. For clinical utility, it is important that a bilirubin surrogate

measure behaves similarly to the clinical scores and outcomes used in treatment planning as well as to bilirubin directly. Therefore this is a positive finding despite the fact that the actual correlations of both bilirubin and the yellowness with the clinical scores and outcomes were low.

6.2.6 Longitudinal clinical data

As well as cross-sectional data, for the first time longitudinal patient image data was also collected and analysed (see Chapter 4.5.2). Changes in yellowness and bilirubin level were considered for individuals and correlations of 0.53 and 0.72 ($p< 0.001$) were determined for the two phones. This shows that it should be possible to track changes in the same patient over time. However, the prediction intervals for the expected range of new measurements were 210 and 267μmol/L for the two phones respectively. These would need to be improved for the system to have clinical utility.

6.2.7 Additional biomarkers

Finally, work was begun looking to the future of the system. A monitoring system for overall liver health must be based on more than one biomarker. Initial data and analysis were carried out to investigate the possibility of incorporating an image-based assessment of haemoglobin level, since reduced haemoglobin (anaemia) has been found to be associated with more severe liver disease (see Chapter 4.6). Redness measures extracted from images of the lower eyelid and sclera were compared to the measured haemoglobin levels. The strongest result was found using images of the sclera, yielding a correlation of 0.37 ($p<0.001$). This is not strong enough to warrant inclusion at present, but highlights the possibility.

6.3. Future work

Further patient imaging studies would be extremely valuable. Gathering additional longitudinal image data in particular would help to improve the methodology and processing, focussing in on the predictive power. The patient study carried out thus far was in a hospital-based setting, however the long term aim is that the system be used in the home. Gathering data in this context is very important for testing the true usability of the system, and suggesting any changes.

Alongside collection of further patient image data, several processing steps could be improved. There is potential for improvement in the filtering algorithm. The current form serves at least as a proof of concept that filtering is both useful and

possible, but is not completely optimised. The test for the presence of specular reflection should be improved, as images with partial specular reflection are often not flagged. The registration of flash to no-flash should also be improved, since reliably transferring the filtered flash mask helps to ensure consistency. An entirely different filtering algorithm which still removes blood vessels and specular reflection could also be developed.

Another processing step warranting improvement is the segmentation step. Manual segmentation of the sclera region was used for all results presented in this thesis. To improve usability and efficiency, an automatic or semi-automatic algorithm should be introduced. The biometrics community are interested in the use of the sclera as a unique identifier, and as a result there has been significant research into automatic sclera segmentation [147]. Existing approaches could therefore be leveraged for this context.

The planned studies measuring scleral reflectance in both volunteers and patients had to be cancelled due to the Covid-19 pandemic. Carrying out these studies would provide novel information about the properties of the human sclera. With the adult volunteer study, measuring the reflectance of the sclera across different ages would help to quantify natural variation between individuals as well as the slight darkening and yellowing of the sclera with age which has been previously observed. If significant, this information could be fed back into the imaging system to include an age-related correction. To our knowledge, the scleral reflectance of adults with different levels of bilirubin has not been measured. Obtaining this information via a patient study would be extremely valuable, again feeding back in to the imaging system. Quantitative measurements of scleral yellowness could enable the development of a custom colour chart for the one-time calibration step, with yellow patches chosen based on real measurements. The spectral trends could also inform the choice of colour metric or even colour space for the analysis. As well as the telespectroradiometric methodology tested and presented here, the possibility of using a hyperspectral camera should be considered. Spectra for a series of points across the sclera could be obtained simultaneously, and this may help to isolate the contribution of blood vessels.

6.4. Outlook and uptake

The desire for remote monitoring of patients with liver disease has been growing in recent years, and the Covid-19 pandemic has further highlighted the utility. There is definite interest from clinicians in having a tool to enable assessment of their patients while out of hospital. The image-based assessment method for bilirubin level presented here has the potential to help meet this desire and need.

The image capture process was shown to be both fast and straightforward during the in-patient study, with the vast majority of image capture sessions resulting in useable data. This is despite the fact that many of the patients in hospital were significantly more unwell than the target population of discharged patients. Whilst the image capture for this study was performed largely by clinical fellows, this success implies that with a small amount of training, a relative or caregiver, a visiting healthcare professional, or even the patient themselves should be able to capture the images. The ability to capture good quality data should therefore not present a barrier to adoption of the technology.

Full automation of the processing after the image capture step is possible, and results could then be presented to the user in real time as well as relayed to their clinician to aid in any decisions required based on the results. It is hoped that the low level of complexity of use and real time results may help the patient to feel empowered in tracking their own health and feel more in control of their condition.

Other biomarkers are known to be important in tracking a patient's health, and a multi-factor system is expected to provide greater insight than the sum of its parts. Since the proposed bilirubin tracking system is already app-based it could easily be incorporated into an overall monitoring system. The ability to combine the proposed approach with other monitoring methods should help to produce a tool which has the potential to make a real difference to patients with liver disease.

Bibliography

[1] R. Williams, R. Aspinall, M. Bellis, *et al.*, "Addressing liver disease in the UK: A blueprint for attaining excellence in health care and reducing premature mortality from lifestyle issues of excess consumption of alcohol, obesity, and viral hepatitis," *The Lancet*, vol. 384, no. 9958, pp. 1953–1997, 2014.

[2] D. A. Leon and J. McCambridge, "Liver cirrhosis mortality rates in Britain from 1950-2002: an analysis of routine data.," *The Lancet*, vol. 367, no. 9504, pp. 52–56, 2006.

[3] K. Berman, S. Tandra, K. Forssell, *et al.*, "Incidence and Predictors of 30-Day Readmission Among Patients Hospitalized for Advanced Liver Disease," *Clinical Gastroenterology and Hepatology*, vol. 9, no. 3, pp. 254–259, 2011.

[4] R. Jalan, M. Pavesi, F. Saliba, *et al.*, "The CLIF Consortium Acute Decompensation score (CLIF-C ADs) for prognosis of hospitalised cirrhotic patients without acute-on-chronic liver failure," *Journal of Hepatology*, vol. 62, no. 4, pp. 831–840, 2015.

[5] R. Moreau, R. Jalan, P. Gines, *et al.*, "Acute-on-chronic liver failure is a distinct syndrome that develops in patients with acute decompensation of cirrhosis," *Gastroenterology*, vol. 144, no. 7, pp. 1426–1437.e9, 2013.

[6] E. A. Tsochatzis, J. Bosch, and A. K. Burroughs, "Liver cirrhosis.," *The Lancet*, vol. 383, pp. 1749–61, may 2014.

[7] L. Castera, J. Foucher, P. H. Bernard, *et al.*, "Pitfalls of liver stiffness measurement: A 5-year prospective study of 13,369 examinations," *Hepatology*, vol. 51, pp. 828–835, mar 2010.

[8] A. Srivastava, R. Gailer, S. Tanwar, *et al.*, "Prospective evaluation of a primary care referral pathway for patients with non-alcoholic fatty liver disease," *Journal of Hepatology*, vol. 71, pp. 371–378, aug 2019.

[9] M. Nagel, C. Labenz, M. A. Wörns, *et al.*, "Impact of acute-on-chronic liver failure and decompensated liver cirrhosis on psychosocial burden and quality of life of patients and their close relatives," *Health and Quality of Life Outcomes*, vol. 18, jan 2020.

[10] E. B. Tapper and M. Volk, "Strategies to Reduce 30-Day Readmissions in Patients with Cirrhosis," *Current Gastroenterology Reports*, vol. 19, no. 1, pp. 1–7, 2017.

[11] J. S. Bajaj, K. R. Reddy, P. Tandon, *et al.*, "The 3-month readmission rate remains unacceptably high in a large North American cohort of patients with cirrhosis," *Hepatology*, vol. 64, no. 1, pp. 200–208, 2016.

[12] K. J. Fagan, E. Y. Zhao, L. U. Horsfall, *et al.*, "Burden of decompensated cirrhosis and ascites on hospital services in a tertiary care facility: Time for change?," *Internal Medicine Journal*, vol. 44, no. 9, pp. 865–872, 2014.

[13] E. B. Tapper, D. Finkelstein, M. A. Mittleman, *et al.*, "A Quality Improvement Initiative Reduces 30-Day Rate of Readmission for Patients With Cirrhosis," *Clinical Gastroenterology and Hepatology*, vol. 14, pp. 753–759, may 2016.

[14] F. Morando, G. Maresio, S. Piano, *et al.*, "How to improve care in outpatients with cirrhosis and ascites: A new model of care coordination by consultant hepatologists," *Journal of Hepatology*, vol. 59, no. 2, pp. 257–264, 2013.

[15] T. K. Nuckols, E. Keeler, S. Morton, *et al.*, "Economic evaluation of quality improvement interventions designed to prevent hospital readmission: A systematic review and meta-analysis," *JAMA Internal Medicine*, vol. 177, pp. 975–985, jul 2017.

[16] E. B. Tapper and S. K. Asrani, "The COVID-19 pandemic will have a long-lasting impact on the quality of cirrhosis care," *Journal of Hepatology*, vol. 73, pp. 441–445, aug 2020.

[17] *Mobile cellular subscriptions data: International Telecommunications Union - http://data.worldbank.org/indicator/IT.CEL.SETS/countries?display=default. Retrieved March 2021.* 2021.

[18] S. Majumder and M. J. Deen, "Smartphone sensors for health monitoring and diagnosis," may 2019.

[19] F. Sposaro, J. Danielson, and G. Tyson, "iWander: An Android application for dementia patients," *2010 Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBC'10*, pp. 3875–3878, 2010.

[20] D. Ben-Zeev, E. A. Scherer, R. Wang, H. Xie, and A. T. Campbell, "Next-generation psychiatric assessment: Using smartphone sensors to monitor behavior and mental health," *Psychiatric Rehabilitation Journal*, vol. 38, no. 3, pp. 218–226, 2015.

[21] E. C. Larson, M. Goel, G. Boriello, *et al.*, "SpiroSmart: Using a microphone to measure lung function on a mobile phone," in *UbiComp'12 - Proceedings of the 2012 ACM Conference on Ubiquitous Computing*, (New York, New York, USA), pp. 280–289, ACM Press, 2012.

[22] M. J. Stotts, J. A. Grischkan, and V. Khungar, "Improving cirrhosis care: The potential for telemedicine and mobile health technologies," aug 2019.

[23] T. Wu, D. A. Simonetto, J. D. Halamka, and V. H. Shah, "The digital transformation of hepatology: The patient is logged in," mar 2022.

[24] J. Sack, T. Reid, E. Schlossberg, and N. Hashemi, "A Smartphone App for Patients With End-Stage Liver Disease Can Detect Behavioral Changes That Predict Liver-Related Events," *Iproceedings*, vol. 5, p. e15229, oct 2019.

[25] J. S. Bajaj, L. R. Thacker, D. M. Heuman, *et al.*, "The stroop smartphone application is a short and valid method to screen for minimal hepatic encephalopathy," *Hepatology*, vol. 58, pp. 1122–1132, sep 2013.

[26] J. S. Bajaj, "Adventures in Developing an App for Covert Hepatic Encephalopathy," *Clinical and Translational Gastroenterology*, vol. 8, p. e85, apr 2017.

[27] P. Bloom, M. Marx, T. Wang, *et al.*, "A Smartphone App Is Feasible for Outpatient Cirrhotic Ascites Management," *Iproceedings*, vol. 5, p. e15130, oct 2019.

[28] P. P. Bloom, M. Marx, T. J. Wang, *et al.*, "Attitudes towards digital health tools for outpatient cirrhosis management in patients with decompensated cirrhosis," *BMJ Innovations*, vol. 6, pp. 18–25, jan 2020.

[29] K. Kazankov, S. Novelli, D. A. Chatterjee, *et al.*, "Evaluation of CirrhoCare® - A digital-health solution for home management of patients with cirrhosis," *Journal of Hepatology*, sep 2022.

[30] R. N. H. Pugh, I. M. Murray-Lyon, J. L. Dawson, M. C. Pietroni, and R. Williams, "Transection of the oesophagus for bleeding oesophageal varices," *British Journal of Surgery*, vol. 60, pp. 646–649, aug 1973.

[31] P. Ferenci, A. Lockwood, K. Mullen, *et al.*, "Hepatic encephalopathy - Definition, nomenclature, diagnosis, and quantification: Final report of the Working Party at the 11th World Congresses of Gastroenterology, Vienna, 1998," *Hepatology*, vol. 35, no. 3, pp. 716–721, 2002.

[32] F. Durand and D. Valla, "Assessment of the prognosis of cirrhosis: Child-Pugh versus MELD," *Journal of Hepatology*, vol. 42, pp. S100–S107, apr 2005.

[33] M. Malinchoc, P. S. Kamath, F. D. Gordon, *et al.*, "A model to predict poor survival in patients undergoing transjugular intrahepatic portosystemic shunts," *Hepatology*, vol. 31, pp. 864–871, apr 2000.

[34] P. S. Kamath and W. R. Kim, "The Model for End-stage Liver Disease (MELD)," *Hepatology*, vol. 45, pp. 797–805, mar 2007.

[35] W. R. Kim, S. W. Biggins, W. K. Kremers, *et al.*, "Hyponatremia and Mortality among Patients on the Liver-Transplant Waiting List," *New England Journal of Medicine*, vol. 359, pp. 1018–1026, sep 2008.

[36] R. J. Porte, T. Lisman, A. Tripodi, S. H. Caldwell, and J. F. Trotter, "The International Normalized Ratio (INR) in the MELD Score: Problems and Solutions," *American Journal of Transplantation*, vol. 10, pp. 1349–1353, mar 2010.

[37] C. E. Nwankire, M. Czugala, R. Burger, *et al.*, "A portable centrifugal analyser for liver function screening," *Biosensors and Bioelectronics*, vol. 56, pp. 352–358, 2014.

[38] B. L. Thompson, S. L. Wyckoff, D. M. Haverstick, and J. P. Landers, "Simple, Reagentless Quantification of Total Bilirubin in Blood Via Microfluidic Phototreatment and Image Analysis," *Analytical Chemistry*, vol. 89, no. 5, pp. 3228–3234, 2017.

[39] W. Tan, L. Zhang, J. C. Doery, and W. Shen, "Three-dimensional microfluidic tape-paper-based sensing device for blood total bilirubin measurement in jaundiced neonates," *Lab on a Chip*, vol. 20, pp. 394–404, jan 2020.

[40] W. D. Engle, G. L. Jackson, and N. G. Engle, "Transcutaneous bilirubinometry," *Seminars in Perinatology*, vol. 38, pp. 438–451, nov 2014.

[41] R. C. Amos, H. Jacob, and W. Leith, "Jaundice in newborn babies under 28 days: NICE guideline 2016 (CG98)," aug 2017.

[42] A. Mariakakis, M. A. Banks, L. Phillipi, *et al.*, "BiliScreen: Smartphone-Based Scleral Jaundice Monitoring for Liver and Pancreatic Disorders," *Proceedings of the ACM on Interactive, Mobile, Wearable and Ubiquitous Technologies*, vol. 1, no. 2, pp. 1–26, 2017.

[43] B. O. Olusanya, D. O. Imosemi, and A. A. Emokpae, "Differences between transcutaneous and serum bilirubin measurements in black African neonates," *Pediatrics*, vol. 138, sep 2016.

[44] M. A. Ruiz, L. S. Rickman, and S. Saab, "The clinical detection of scleral icterus: Observations of multiple examiners," *Military Medicine*, vol. 162, pp. 560–563, aug 1997.

[45] O. L. Hung, N. S. Kwon, A. E. Cole, *et al.*, "Evaluation of the physician's ability to recognize the presence or absence of anemia, fever, and jaundice," *Academic Emergency Medicine*, vol. 7, pp. 146–156, feb 2000.

[46] A. K. Yetisen, J. L. Martinez-Hurtado, A. Garcia-Melendrez, F. Da Cruz Vasconcellos, and C. R. Lowe, "A smartphone algorithm with inter-phone repeatability for the analysis of colorimetric tests," *Sensors and Actuators, B: Chemical*, vol. 196, pp. 156–160, jun 2014.

[47] L. Shen, J. A. Hagen, and I. Papautsky, "Point-of-care colorimetric detection with a smartphone," *Lab on a Chip*, vol. 12, no. 21, pp. 4240–4243, 2012.

[48] A. Y. Mutlu, V. Kiliç, G. K. Özdemir, *et al.*, "Smartphone-based colorimetric detection via machine learning," *Analyst*, vol. 142, no. 13, pp. 2434–2441, 2017.

[49] M. Y. Jia, Q. S. Wu, H. Li, *et al.*, "The calibration of cellphone camera-based colorimetric sensor array and its application in the determination of glucose in urine," *Biosensors and Bioelectronics*, vol. 74, pp. 1029–1037, 2015.

[50] Y. Jung, J. Kim, O. Awofeso, *et al.*, "Smartphone-based colorimetric analysis for detection of saliva alcohol concentration," *Applied Optics*, vol. 54, no. 31, p. 9183, 2015.

[51] V. Kiliç, G. Alankus, N. Horzum, *et al.*, "Single-Image-Referenced Colorimetric Water Quality Detection Using a Smartphone," *ACS Omega*, vol. 3, pp. 5531–5536, may 2018.

[52] D. Akkaynak, E. Chan, J. J. Allen, and R. T. Hanlon, "Using spectrometry and photography to study color underwater," *Oceans'11 Mts/Ieee Kona*, pp. 1–8, sep 2011.

[53] A. C. C. Fulgêncio, V. P. Araújo, H. V. Pereira, B. G. Botelho, and M. M. Sena, "Development of a Simple and Rapid Method for Color Determination in Beers Using Digital Images," *Food Analytical Methods*, vol. 13, pp. 303–312, jan 2020.

[54] S. Collings, O. Thompson, E. Hirst, *et al.*, "Non-invasive detection of anaemia using digital photographs of the conjunctiva," *PLoS ONE*, vol. 11, no. 4, 2016.

[55] V. Y. Bunya, D. H. Brainard, E. Daniel, *et al.*, "Assessment of signs of anterior blepharitis using standardized color photographs," *Cornea*, vol. 32, no. 11, pp. 1475–1482, 2013.

[56] L. de Greef, M. Goel, M. J. Seo, *et al.*, "Bilicam: Using Mobile Phones to Monitor Newborn Jaundice," in *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing - UbiComp '14 Adjunct*, pp. 331–342, 2014.

[57] J. A. Taylor, J. W. Stout, L. de Greef, *et al.*, "Use of a Smartphone App To Assess Neonatal Jaundice," *Pediatrics*, vol. 140, p. e20170312, sep 2017.

[58] A. Aune, G. Vartdal, H. Bergseng, L. L. Randeberg, and E. Darj, "Bilirubin estimates from smartphone images of newborn infants' skin correlated highly to serum bilirubin levels," *Acta Paediatrica*, vol. 00, pp. 1–7, apr 2020.

[59] A. N. Bashkatov, E. A. Genina, V. I. Kochubey, and V. V. Tuchin, "Optical properties of human sclera in spectral range 370-2500 nm," *Optics and Spectroscopy (English translation of Optika i Spektroskopiya)*, vol. 109, pp. 197–204, aug 2010.

[60] A. Laddi, S. Kumar, S. Sharma, and A. Kumar, "Non-invasive jaundice detection using machine vision," *IETE Journal of Research*, vol. 59, no. 5, pp. 591–596, 2013.

[61] T. S. Leung, K. Kapur, A. Guilliam, *et al.*, "Screening neonatal jaundice based on the sclera color of the eye using digital photography," *Biomedical Optics Express*, vol. 6, no. 11, p. 4529, 2015.

[62] M. R. Rizvi, F. M. Alaskar, R. S. Albaradie, N. F. Rizvi, and K. Al-Abdulwahab, "A novel non-invasive technique of measuring bilirubin levels using bilicapture," *Oman Medical Journal*, vol. 34, pp. 26–33, jan 2019.

[63] M. R. Sammir, K. M. Towhidul Alam, P. Saha, M. M. Rahaman, and Q. D. Hossain, "Design and implementation of a non-invasive jaundice detection and

total serum bilirubin measurement system," *ICECE 2018 - 10th International Conference on Electrical and Computer Engineering*, pp. 137–140, 2018.

[64] M. M. M. Miah, R. J. Tazim, F. T. Johora, *et al.*, "Non-Invasive Bilirubin Level Quantification and Jaundice Detection by Sclera Image Processing," *IEEE Global Humanitarian Technology Conference (GHTC)*, pp. 1–7, mar 2019.

[65] M. Nixon, F. Outlaw, and T. S. Leung, "Accurate device-independent colorimetric measurements using smartphones," *PLOS ONE*, vol. 15, p. e0230561, mar 2020.

[66] M. Nixon, F. Outlaw, L. W. MacDonald, and T. S. Leung, "The importance of a device specific calibration for smartphone colorimetry," *Color and Imaging Conference 27*, 2019.

[67] M. Nixon-Hill, F. Outlaw, L. W. Macdonald, R. Mookerjee, and T. S. Leung, "Minimising ambient illumination via ambient subtraction : smartphone assessment of jaundice in liver patients via sclera images," *Color and Imaging Conference 28*, pp. 307–312, 2020.

[68] T. Gevers, A. Gijsenij, J. van de Weijer, and J. Geusebroek, "Color Image Formation," in *Color in Computer Vision: Fundamentals and Applications*, pp. 26–45, Hoboken, NJ, USA: John Wiley & Sons, Inc., aug 2012.

[69] S. A. Shafer, "Using color to separate reflection components," *Color Research & Application*, vol. 10, no. 4, pp. 210–218, 1985.

[70] B. E. Bayer, "Color Imaging Array Patent US 3971065," mar 1976.

[71] O. Burggraaff, N. Schmidt, J. Zamorano, *et al.*, "Standardized spectral and radiometric calibration of consumer cameras," *Optics Express*, vol. 27, no. 14, p. 19075, 2019.

[72] S. D. Hordley, "Scene illuminant estimation: Past, present, and future," *Color Research and Application*, vol. 31, no. 4, pp. 303–314, 2006.

[73] A. Gijsenij, T. Gevers, and J. Van De Weijer, "Computational color constancy: Survey and experiments," *IEEE Transactions on Image Processing*, vol. 20, no. 9, pp. 2475–2489, 2011.

[74] M. Males, A. Hedi, and M. Grgic, "Colour balancing using sclera colour," *IET Image Processing*, vol. 12, pp. 416–421, mar 2018.

[75] D. Lee and K. N. Plataniotis, "A taxonomy of color constancy and invariance algorithm," *Advances in Low-Level Color Image Processing*, vol. 11, pp. 55–94, 2014.

[76] V. Agarwal, B. R. Abidi, A. Koschan, and M. A. Abidi, "An Overview of Color Constancy Algorithms," *Journal of Pattern Recognition Research*, vol. 1, no. 1, pp. 42–54, 2006.

[77] J. M. DiCarlo, F. Xiao, and B. a. Wandell, "Illuminating Illumination," *Ninth Color Imaging Conference*, no. January 2001, pp. 27–34, 2001.

[78] G. Petschnigg, R. Szeliski, M. Agrawala, *et al.*, "Digital photography with flash and no-flash image pairs," *ACM Transactions on Graphics*, vol. 23, no. 3, p. 664, 2004.

[79] C. Lu and M. S. Drew, "Practical Scene Illuminant Estimation via Flash / No-Flash Pairs," *Fourteenth Color and Imaging Conference*, no. January 2006, pp. 84–89, 2006.

[80] Z. Hui, A. C. Sankaranarayanan, K. Sunkavalli, and S. Hadap, "White balance under mixed illumination using flash photography," *2016 IEEE International Conference on Computational Photography, ICCP 2016 - Proceedings*, 2016.

[81] F. Outlaw, M. Nixon, O. Odeyemi, *et al.*, "Smartphone screening for neonatal jaundice via ambient-subtracted sclera chromaticity," *PLOS ONE*, vol. 15, mar 2020.

[82] A. Rowlands, S. D. J. Gwyn, H. Wang, K.-s. Lee, and J.-h. Ryu, *Chapter 4 Raw conversion*. IOP Publishing, 2017.

[83] S. Westland, *Computational Colour Science using MATLAB 2e (https://www.mathworks.com/matlabcentral/fileexchange/40640-computational-colour-science-using-matlab-2e), MATLAB Central File Exchange. Retrieved October 2018.*

[84] A. Clouet, D. Alleysson, and G. Alpes, "Physical noise propagation in color image construction : a geo- metrical interpretation," *Color and Imaging Conference*, pp. 375–380, 2019.

[85] R. Luther, "Aus dem Gebiet der Farbreizmetrik (On color stimulus metrics)," *Zeitschrift für technische Physik*, vol. 8, no. 1927, pp. 540–558, 1927.

[86] J. Nakamura, ed., *Image sensors and signal processing for digital still cameras*. Taylor & Francis, 2005.

[87] Image Engineering GmbH &Co. KG, "camSPECS express Measurement Device,"

[88] L. W. MacDonald, "Determining Camera Spectral Responsivity with Multispectral Transmission Filters," *Color and Imaging Conference*, pp. 12–17, 2015.

[89] D. S. Hawkins and P. Green, "Spectral Characterisation of a Digital Still Camera Through a Single Integrating Exposure .," *Proc. 4th IS&T Eur. Conf. on Colour in Graphics, Imaging and Vision (CGIV)*, pp. 477–480, 2008.

[90] J. Jiang, D. Liu, J. Gu, and S. Susstrunk, "What is the space of spectral sensitivity functions for digital color cameras?," *Proceedings of IEEE Workshop on Applications of Computer Vision*, pp. 168–179, 2013.

[91] D. Coffin, "Decoding raw digital photos in Linux."

[92] Adobe Systems Incorporated, "Digital Negative (DNG) specification. Version 1.4.0.0," 2012.

[93] S. Westland, C. Ripamonti, and V. Cheung, *Computational Colour Science using MATLAB®, Second Edition*. John Wiley & Sons, 2012.

[94] G. D. Finlayson and M. S. Drew, "Constrained least-squares regression in color spaces," *J Electron Imaging*, vol. 6, no. 4, p. 484, 1997.

[95] G. D. Finlayson, M. MacKiewicz, and A. Hurlbert, "Color Correction Using Root-Polynomial Regression," *IEEE Transactions on Image Processing*, vol. 24, pp. 1460–1470, may 2015.

[96] B. Funt and P. Bastani, "Irradiance-independent camera color calibration," *Color Research and Application*, vol. 39, pp. 540–548, dec 2014.

[97] P. Bastani and B. Funt, "Simplifying irradiance independent color calibration," in *Color Imaging XIX: Displaying, Processing, Hardcopy, and Applications* (R. Eschbach, G. G. Marcu, and A. Rizzi, eds.), vol. 9015, p. 90150N, International Society for Optics and Photonics, jan 2014.

[98] G. D. Finlayson, M. Mohammadzadeh Darrodi, and M. Mackiewicz, "The alternating least squares technique for nonuniform intensity color correction," *Color Research and Application*, vol. 40, no. 3, pp. 232–242, 2015.

[99] R. C. Gonzalez, R. E. Woods, and S. L. Eddins, *Digital Image Processing Using MATLAB*. Upper Saddle River NJ: Pearson/Prentice Hall, 2004.

[100] L. W. MacDonald, ed., *Digital Heritage: Applying Digital Imaging to Cultural Heritage*. Routledge Ltd, 2006.

[101] I. G. Hughes and T. P. A. Hase, *Measurements and their Uncertainties*. Oxford University Press, 2010.

[102] R. W. G. Hunt and M. R. Pointer, *Measuring Colour*. Chichester, UK: John Wiley & Sons, Ltd, sep 2011.

[103] D. Akkaynak, T. Treibitz, B. Xiao, *et al.*, "Use of commercial off-the-shelf digital cameras for scientific data acquisition and scene-specific color calibration," *Journal of the Optical Society of America A*, vol. 31, no. 2, p. 312, 2014.

[104] T. E. White, R. L. Dalrymple, D. W. Noble, *et al.*, "Reproducible research in the study of biological coloration," *Animal Behaviour*, vol. 106, pp. 51–57, aug 2015.

[105] S. Hannuna, A. N, S. Subramanian, *et al.*, "Agriculture Disease Mitigation System," *ICTACT Journal on Communication Technology*, vol. 02, pp. 363–369, jun 2011.

[106] K. Carpenter and S. Farnand, "Assessing the use of smartphones to determine crop ripeness," *Workshop on Food and Agriculture, Electronic Imaging*, 2020.

[107] S. Farnand and K. Parulski, "Color calibration of unmanned aerial system digital still cameras," in *IS and T International Symposium on Electronic Imaging Science and Technology*, Society for Imaging Science and Technology, 2018.

[108] M. S. Kurečić, D. Antonic, and I. Vranjkovic, "Custom Colour Reference Target for Chronic Wound Photography," *AIC Colour conference proceedings*, pp. 1353–1356, 2013.

[109] G. Trumpy, "Digital Reproduction of Small Gamut Objects: A Profiling Procedure based on Custom Color Targets," *CGIV*, 2010.

[110] M. S. Kurečić, D. Agić, and L. Mandić, "Developing a custom colour target for artwork imaging," *Imaging Science Journal*, vol. 59, no. 6, pp. 317–331, 2011.

[111] A. Olejnik-Krugly and P. Korytkowski, "Precise color capture using custom color targets," *Color Research & Application*, pp. 1–9, oct 2019.

[112] M. Nixon-Hill, *Semi-automatic ColorChecker segmentation (https://www.mathworks.com/matlabcentral/fileexchange/76715-semi-automatic-colorchecker-segmentation), MATLAB Central File Exchange. Retrieved June 16, 2021.*

[113] M. Nixon-Hill, R. P. Mookerjee, and T. S. Leung, "Assessment of bilirubin levels in patients with cirrhosis via forehead, sclera and lower eyelid smartphone images," *PLOS Digital Health*, no. Under review, 2022.

[114] K. Kazankov, M. Nixon-Hill, R. Kumar, *et al.*, "A novel Smartphone scleral-image based tool for assessing jaundice in decompensated cirrhosis patients," *Journal of Gastroenterology and Hepatology*, p. Under review, 2022.

[115] R. Sumner, "Processing RAW Images in MATLAB," *Department of Electrical Engineering, University of California Sata Cruz*, 2014.

[116] S. Crihalmeanu, A. Ross, and R. Derakhshani, "Enhancement and registration schemes for matching conjunctival vasculature," in *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, vol. 5558 LNCS, pp. 1240–1249, Springer, Berlin, Heidelberg, 2009.

[117] N. Otsu, "Threshold Selection Method from Gray-Level Histograms," *IEEE Trans Syst Man Cybern*, vol. SMC-9, no. 1, pp. 62–66, 1979.

[118] C. Kirbas and F. Quek, "A review of vessel extraction techniques and algorithms," *ACM Computing Surveys*, vol. 36, no. 2, pp. 81–121, 2004.

[119] A. F. Frangi, W. J. Niessen, K. L. Vincken, and M. A. Viergever, "Multiscale vessel enhancement filtering," in *Medical Image Computing and Computer-Assisted Intervention — MICCAI'98*, vol. 1496, 1998.

[120] S. Prahl, *Optical extinction coefficients: https://omlc.org/spectra/.* 2018.

[121] T. Wemyss, M. Nixon-hill, A. Karsa, J. Meek, and T. S. Leung, "Feasibility of smartphone colorimetry of the face as an anaemia screening tool for infants and young children in Ghana," *PLoS ONE*, no. Under review, pp. 1–23, 2022.

[122] E. R. DeLong, D. M. DeLong, and D. L. Clarke-Pearson, "Comparing the Areas under Two or More Correlated Receiver Operating Characteristic Curves : A Nonparametric Approach," *Biometrics*, vol. 44, no. 3, pp. 837–845, 1988.

[123] C. Jansen, D. A. Chatterjee, K. L. Thomsen, *et al.*, "Significant reduction in heart rate variability is a feature of acute decompensation of cirrhosis and predicts 90-day mortality," *Alimentary Pharmacology and Therapeutics*, 2019.

[124] O. Collas, F. P. Robertson, B. J. Fuller, and B. R. Davidson, "Anaemia in patients with chronic liver disease and its association with morbidity and mortality following liver transplantation," *International Journal of Surgery*, vol. 53, pp. 48–52, may 2018.

[125] S. Piano, M. Tonon, E. Vettore, *et al.*, "Incidence, predictors and outcomes of acute-on-chronic liver failure in outpatients with cirrhosis," *Journal of Hepatology*, vol. 67, no. 6, pp. 1177–1184, 2017.

[126] E. J. Wang, W. Li, J. Zhu, R. Rana, and S. N. Patel, "Noninvasive hemoglobin measurement using unmodified smartphone camera and white flash," in *Proceedings of the Annual International Conference of the IEEE Engineering in Medicine and Biology Society, EMBS*, pp. 2333–2336, Institute of Electrical and Electronics Engineers Inc., sep 2017.

[127] R. G. Mannino, D. R. Myers, E. A. Tyburski, *et al.*, "Smartphone app for non-invasive detection of anemia using only patient-sourced photos," *Nature Communications*, vol. 9, p. 4924, dec 2018.

[128] G. Dimauro, A. Guarini, D. Caivano, *et al.*, "Detecting clinical signs of anaemia from digital images of the palpebral conjunctiva," *IEEE Access*, vol. 7, pp. 113488–113498, 2019.

[129] T. Yamamoto, H. Takiwaki, S. Arase, and H. Ohshima, "Derivation and clinical application of special imaging by means of digital cameras and Image J freeware for quantification of erythema and pigmentation," *Skin Research and Technology*, vol. 14, pp. 26–34, mar 2008.

[130] *World Health Organization: Haemoglobin concentrations for the diagnosis of anaemia and assessment of severity - https://www.who.int/vmnis/indicators/haemoglobin.pdf. Retrieved April 2021.* 2021.

[131] Y. Zhu and G. D. Finlayson, "Designing a Color Filter via Optimization of Vora-Value for Making a Camera more Colorimetric," *Color and Imaging Conference 28*, pp. 181–186, 2020.

[132] G. D. Finlayson, Y. Zhu, and H. Gong, "Using a simple colour pre-filter to make cameras more colorimetric," in *Final Program and Proceedings - IS and*

T/SID Color Imaging Conference, pp. 182–186, Society for Imaging Science and Technology, 2018.

[133] J. R. Palmer, C. G. Owen, A. M. Ford, R. E. Jacobson, and E. G. Woodward, "Optimal photographic imaging of the bulbar conjunctival vasculature," *Ophthalmic and Physiological Optics*, vol. 16, pp. 141–149, mar 1996.

[134] A. Rohatgi, *WebPlotDigitizer, version 4.4. Retrieved from: https://automeris.io/WebPlotDigitizer*. 2020.

[135] S. A. Lisenko, V. A. Firago, M. M. Kugeiko, and A. I. Kubarko, "Determination of Structural and Morphological Parameters of Human Bulbar Conjunctiva from Optical Diffuse Reflectance Spectra," *Journal of Applied Spectroscopy*, vol. 83, pp. 617–626, sep 2016.

[136] L. E. Mackenzie, T. R. Choudhary, A. I. McNaught, and A. R. Harvey, "Comment on the Influence of Episcleral Blood Vessels in Diffuse Reflectance Spectroscopy Measurements of the Bulbar Conjunctiva," *Article in Journal of Applied Spectroscopy*, 2017.

[137] G. Zonios, J. Bykowski, and N. Kollias, "Skin melanin, hemoglobin, and light scattering properties can be quantitatively assessed in vivo using diffuse reflectance spectroscopy," *Journal of Investigative Dermatology*, vol. 117, pp. 1452–1457, dec 2001.

[138] G. Zonios, L. T. Perelman, V. Backman, *et al.*, "Diffuse reflectance spectroscopy of human adenomatous colon polyps in vivo," *Applied Optics*, vol. 38, no. 31, p. 6628, 1999.

[139] G. Zonios and A. Dimou, "Modeling diffuse reflectance from semi-infinite turbid media: application to the study of skin optical properties," *Optics Express*, vol. 14, p. 8661, sep 2006.

[140] S. T. Flock, M. S. Patterson, B. C. Wilson, and D. R. Wyman, "Monte Carlo Modeling of Light Propagation in Highly Scattering Tissues—I: Model Predictions and Comparison with Diffusion Theory," *IEEE Transactions on Biomedical Engineering*, vol. 36, no. 12, pp. 1162–1168, 1989.

[141] W. Verkruysse, G. W. Lucassen, J. F. De Boer, *et al.*, "Modelling light distributions of homogeneous versus discrete absorbers in light irradiated turbid media," *Physics in Medicine and Biology*, vol. 42, no. 1, pp. 51–65, 1997.

[142] L. O. Svaasand, E. J. Fiskerstrand, G. Kopstad, *et al.*, "Therapeutic Response During Pulsed Laser Treatment of Port-wine Stains: Dependence on Vessel Diameter and Depth in Dermis," *Lasers in Medical Science*, vol. 10, pp. 235–243, 1995.

[143] R. L. P. van Veen, W. Verkruysse, and H. J. C. M. Sterenborg, "Diffuse-reflectance spectroscopy from 500 to 1060 nm by correction for inhomogeneously distributed absorbers," *Optics Letters*, vol. 27, pp. 246–248, feb 2002.

[144] *MIT Diffuse reflectance spectroscopy webpage - http://web.mit.edu/spectroscopy/research/biomedresearch/A1-22.html. Retrieved March 2021.* 2021.

[145] M. Gründl, S. Knoll, M. Eisenmann-Klein, and L. Prantl, "The blue-eyes stereotype: Do eye color, pupil diameter, and scleral color affect attractiveness?," *Aesthetic Plastic Surgery*, vol. 36, pp. 234–240, apr 2012.

[146] R. Russell, J. R. Sweda, A. Porcheron, and E. Mauger, "Sclera color changes with age and is a cue for perceiving age, health, and beauty," *Psychology and Aging*, vol. 29, pp. 626–635, sep 2014.

[147] M. Vitek, A. Das, Y. Pourcenoux, *et al.*, "SSBC 2020: Sclera segmentation benchmarking competition in the mobile environment," in *IJCB 2020 - IEEE/IAPR International Joint Conference on Biometrics*, Institute of Electrical and Electronics Engineers Inc., sep 2020.

[148] A. A. Paterlini, M. A. Nascimento, and C. Traina, "Using Pivots to Speed-Up k-Medoids Clustering," *JIDM*, vol. 2, pp. 221–236, 2011.

[149] S. Pratt, *Fast Medoid Selection using MFAMES (https://www.mathworks.com/matlabcentral/fileexchange/41757-fast-medoid-selection-using-mfames), MATLAB Central File Exchange. Retrieved June 5, 2020.* 2013.

A. Camera spectral sensitivity measurement

The camera spectral sensitivity (CSS) for each of the smartphones used in this study was measured using the setup shown in Figure A.1. A monochromator setup consisting of a Horiba Tunable PowerArc Illuminator equipped with 75 W xenon arc lamp and adjustable bandpass was used to provide continuous narrow bandwidth light from 380 - 780nm. The resulting illumination was directed onto a white reflectance standard (SphereOptics zenith; 99% reflectance). A fixed exposure time and ISO were selected for each smartphone such that the brightest region of the spectrum (here, green) did not saturate the sensors. The peak wavelength emitted from the monochromator was then stepped from 380 - 780nm in 5nm increments, and a PhotoResearch PR655 spectroradiometer was used to measure the precise spectrum of light at the white standard. A raw smartphone image was also captured with the pre-selected fixed exposure settings at each step, with a visualisation of the resulting images shown in Figure A.2. This measurement process was repeated in full for each smartphone used in this research.

The resulting measurements were then processed to obtain the CSS. Firstly, the mean RGB value from the central region of the white standard in each image was calculated. The power of light emitted at each wavelength was calculated by summing over the spectroradiometer measurement The RGB values were then divided by this power to normalise the results for wavelength-dependent illumination intensity. Finally, the resulting normalised RGB values were plotted as a function of the peak wavelength to obtain the CSS for each device. The CSS for the four phones measured are shown overlaid in Figure A.3. Note the minor differences observed for devices of the same model, and slightly larger differences seen between the models. These findings highlight the need for careful processing of image data to obtain meaningful results.

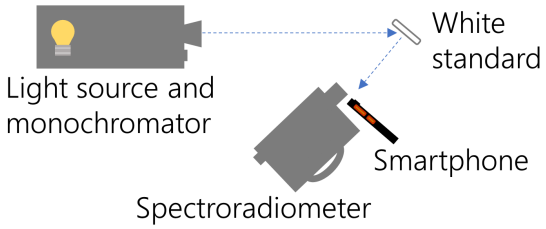


Figure A.1: Diagram showing the setup for CSS measurement - a 75 W xenon arc lamp was combined with a Horiba Tunable PowerArc Illuminator to provide narrow bandwidth light in 5nm steps, this light was shone onto a uniform reflectance SphereOptics zenith white tile, raw images at each wavelength were captured using the smartphone and the spectrum of light measured using a PhotoResearch PR655 spectroradiometer.

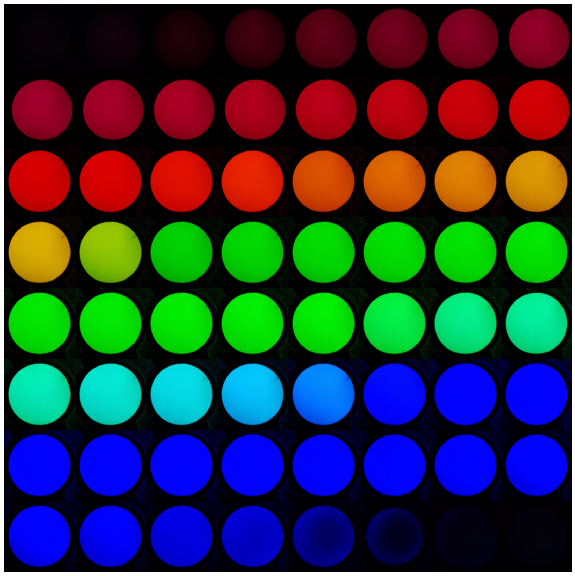


Figure A.2: Cropped example images captured by a Samsung Galaxy S8 phone during CSS measurement. Images were captured every 5nm from 380 - 780nm and those with visible in-image results shown here.

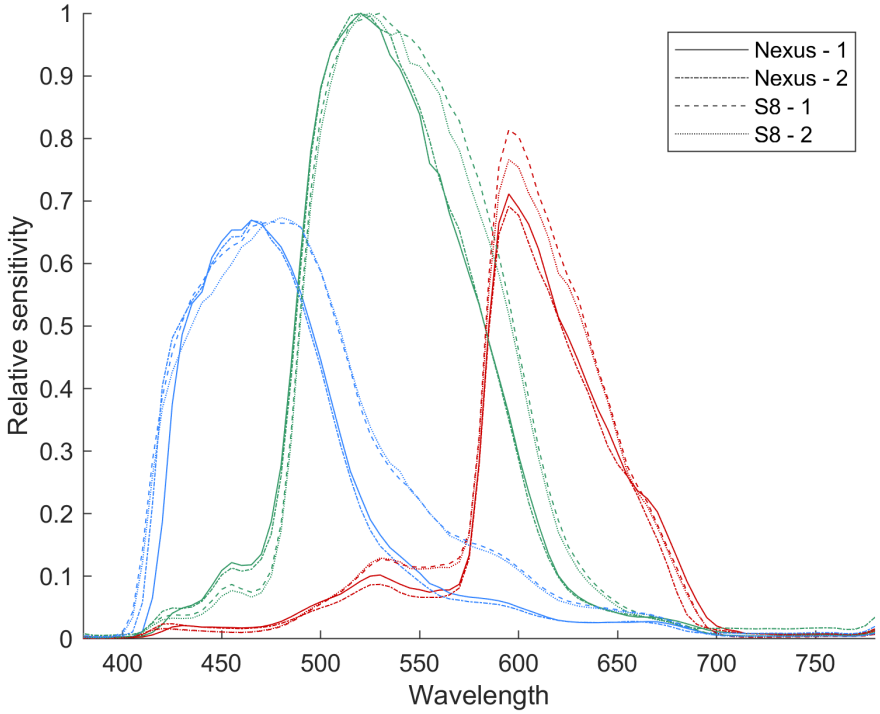


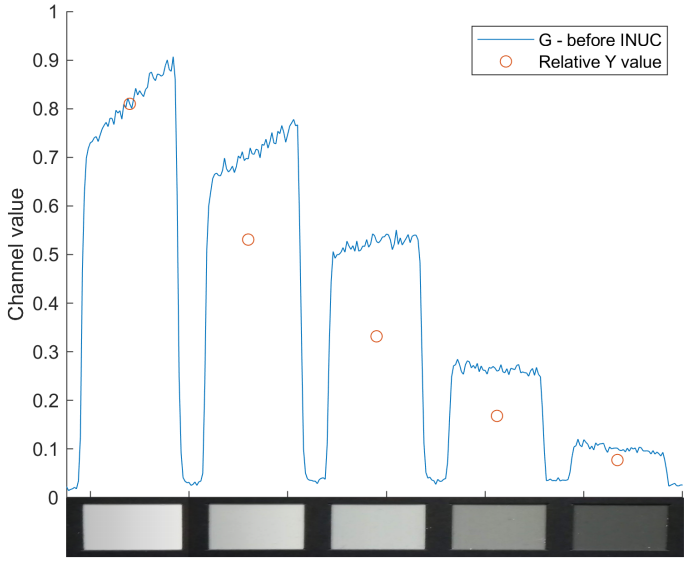
Figure A.3: Normalised camera spectral sensitivity for two LG Nexus 5X phones (solid and dot-dash) and two Samsung Galaxy S8 phones (dash and dot). Sensitivities for each colour channel from long to short wavelength are shown in red, green and blue respectively.

B. Intensity non-uniformity correction impact

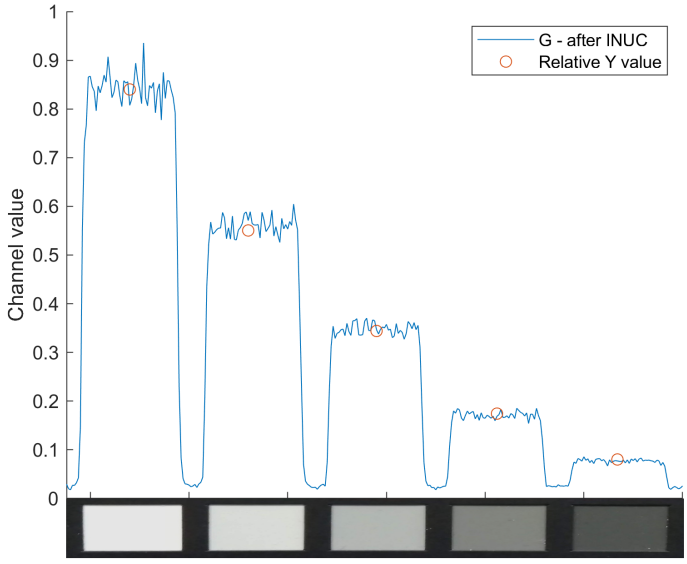
The process for carrying out an intensity non-uniformity correction (INUC) is described in Chapter 2.6.3. Here, data is presented to demonstrate the large impact the INUC has on the extracted pixel values.

The first five neutral patches of the Classic colour chart are considered here, with the black patch excluded owing to its low pixel values. Figure B.1 shows an example line profile across the centre of these Classic neutrals patches for a Nexus phone, before and after the INUC is applied. The line profiles are presented for the green channel, since it has the largest signal to noise ratio and so enables the clearest visualisation. We would expect to see a uniform pixel value across each patch, since they are a constant colour. The varying illumination intensity, however, results in a large variation in pixel values across the patches, especially for the first two, before the correction is applied. Images of the neutral patches are also shown in Figure B.1, where the shading before correction is clearly visible. After the non-uniformity correction is applied the patch values are far more uniform.

Since the phone sensors are linear, with validation data presented in Appendix C, there should be a linear relationship between the RGB channel neutral values and their corresponding XYZ Y channel value. Figure B.1 also shows the Y values for the patches, scaled to the white patch G value. Before the correction, there is a large mismatch between the average G value for each patch and the corresponding Y value. Clearly the variation in illumination intensity is causing a larger problem than just having a non-constant value across the patches. After the correction, there is a strong agreement between the average patch value and the corresponding Y value, suggesting that the combination of downsampling and non-uniformity correction has successfully removed the effects of varying illumination. Note that the increased noise visible in the line profile after INUC will be removed when the average patch values are found using a large area.



(a)



(b)

Figure B.1: Line profiles across the first five neutral patches of the Nexus image of the ColorChecker Classic chart. Data is presented for the green channel values (blue solid line) before (a) and after (b) the INUC, along with the corresponding XYZ Y channel values for the patches scaled to the white patch. Images of the neutral patches are given below the horizontal axes before and after correction. Intensity non-uniformity is evident in the top figure by the angled patch values, mismatch to the Y values, and visible shading in the image, whereas after correction these issues have been removed.

C. Sensor linearity

For quantitative use, it is important that the camera sensors' response to the scene radiance is linear - doubling the amount of light coming into the camera from the scene should double the pixel values (up to the saturation point). This means that the Lambertion model of Equation 2.2 holds within an image, and that a linear transformation to a device-independent space should be possible, as discussed in Chapter 2.5.

An experiment was carried out to investigate the linearity of the sensors of the phones used throughout this research. Images of the 13 neutral patches from the Macbeth ColorChecker DC chart were captured under a fixed ambient illumination for a range of exposure time and ISO values, using the setup shown in Figure C.1. The exposure times and ISOs were chosen to ensure that the full dynamic range of the cameras were tested and to check the common combinations of exposure time and ISO for typical room lighting levels. In order to check the linearity, the extracted RGB values were plotted against the XYZ Y channel values measured using an X-Rite ColorMunki spectrophotometer, which are related to the luminance of the patches [82]. The relationship between RGB and Y should be linear if the sensors are linear [93].

In order to compare the relative RGB values of the patches with any degree of precision, it is necessary to correct for the spatial variation in the intensity of the illumination of the chart as described in Chapter 2.6.3 and demonstrated in Appendix B. As for the usual INUC, grey chart images were captured along with the colour chart images. However, we would like to compare values not just within an image but between different images. A slightly different correction was therefore applied, modified from [93],

$$c_{c,corr}(\boldsymbol{x}) = \overline{n}_G(\boldsymbol{x}_{tot})\frac{c_c(\boldsymbol{x})}{n_G(\boldsymbol{x})} \tag{C.1}$$

where $c_{c,corr}(\boldsymbol{x})$ represents the corrected colour chart image values, with $c \in \{R, G, B\}$. The pixel values in the colour chart image, $c_c(\boldsymbol{x})$, are divided by the corresponding green channel values of the grey chart image, $n_G(\boldsymbol{x})$, and multiplied by the average green channel value of the grey chart image over the whole region of interest $\overline{n}_G(\boldsymbol{x}_{tot})$ to maintain relative pixel values between images with different exposure times. As with the standard INUC, the images were downsampled before carrying out the

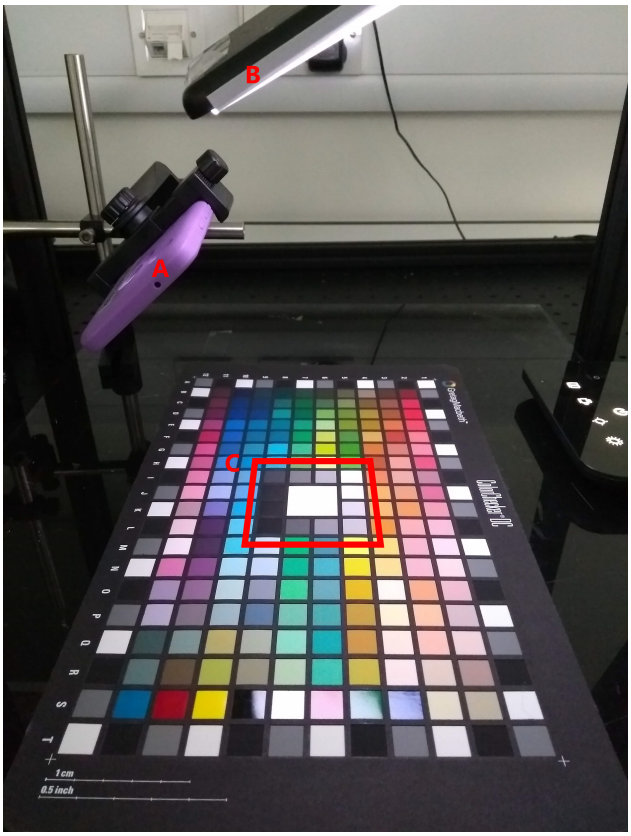


Figure C.1: An example image of the setup used for the linearity experiment, shown for the Nexus phone (A). A fixed ambient illumination was provided by an LED lamp (B), and images of the central neutral patches of Macbeth ColorChecker DC chart (C) were captured, outlined in red, maintaining a 45° angle between the phone and the chart. Corresponding grey chart images using the same setup were captured to perform an intensity non-uniformity correction.

correction to minimise the effects of any small physical shifts between images and noise in the images.

Over 50 combinations of exposure time and ISO were tested for each phone, and some example results are presented in Figure C.2 for the S8 phone. The results for the Nexus phone were very similar in form so have not been presented. The relative RGB values for the neutral patches are shown for two different exposure times, demonstrating the testing of the dynamic range and the resulting linear fits in each case. The RGB responses were linear for both phones for all combinations of exposure time and ISO tested in the experiment, although of course not with exactly the same gradient. This linearity means that it is acceptable to vary the exposure time and ISO depending on the image capture conditions and maintain the linear relationship, and also to carry out the ambient subtraction for pairs of images with the same exposure time and ISO. It is worth noting that if the responses were

not linear, it would be possible to apply a correction factor to linearise the results, but it is an additional complication and source of error that is much better simply avoided.

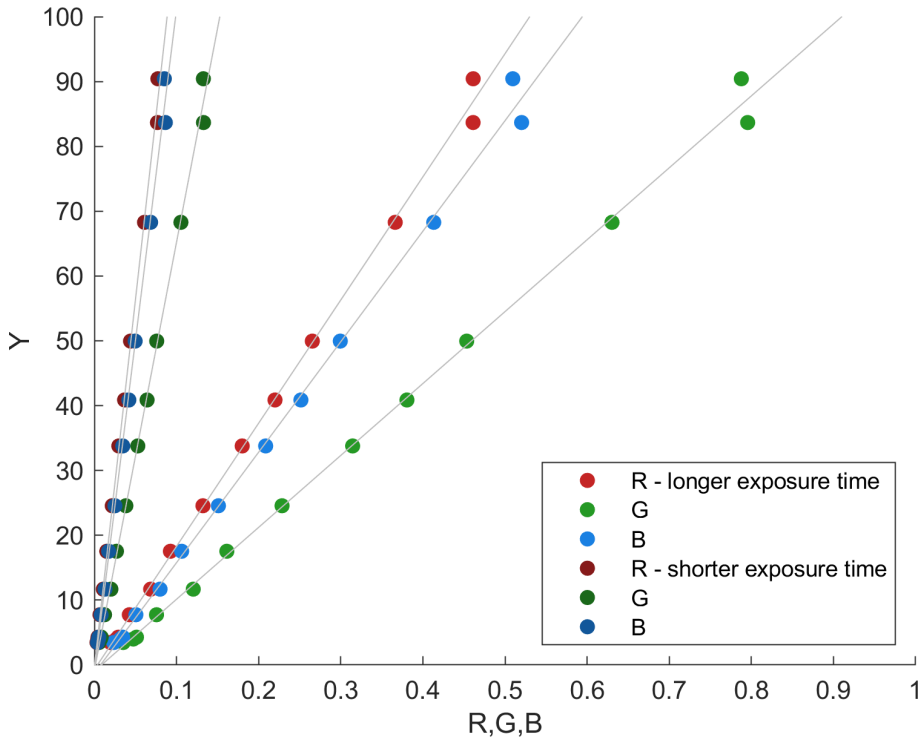


Figure C.2: Linearity results for the S8 phone. The ground truth Y channel values, related to the overall luminance of the patches, are plotted against the intensity-corrected RGB values for two example exposure times along with linear best fit lines for each. In both cases, the RGB responses are linear.

D. Median vs medoid extraction

The motivation to use a median average to help exclude outliers within a region of interest is described in Chapter 4.2. Here, a comparison of two common options for median calculation are compared.

The simplest approach for calculating the median value for a set of pixels is to calculate the median for each RGB channel separately, and use those median values together as the final average. The positive of this approach is that it is simple and computationally cheap. However, the median RGB value it produces is not necessarily a "real" pixel value in the ROI - that is, amongst the input pixels there is not one which has the exact value returned by this approach.

An alternative approach is to use a higher dimensional equivalent of the median, known as the medoid. For a region of interest with N RGB pixels, p_1, p_2,$\cdots$,p_N, the medoid is defined as

$$p_{medoid} = \text{argmin}_{x \in \{p_1, p_2, \cdots, p_N\}} \sum_{i=1}^{N} d(x, p_i) \tag{D.1}$$

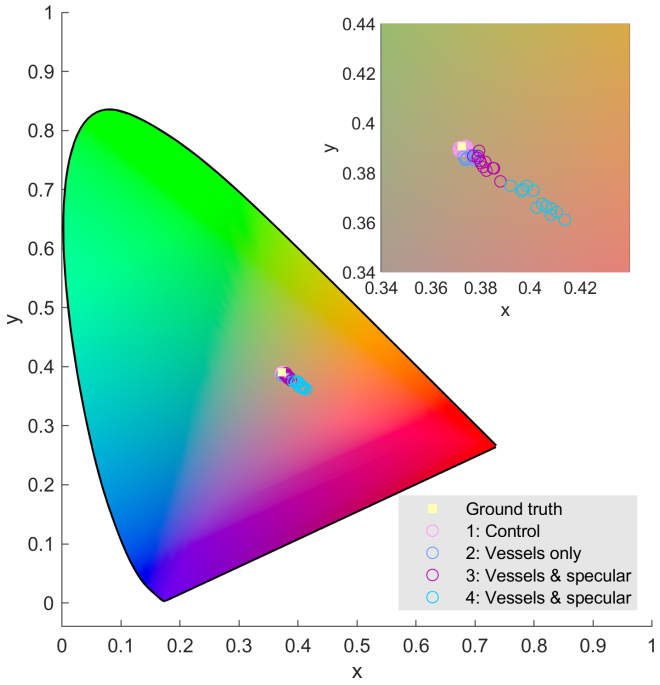
where d represents a distance metric between RGB pixel values, here the Euclidean distance. In other words, the medoid pixel is the pixel with the lowest combined distance to every other pixel. The important difference here from the per channel median is that the medoid pixel is always a member of the input pixels, it is a "real" pixel value. The medoid can be determined with a brute-force approach by simply calculating the pairwise distances for all pixels in the ROI and finding the minimum of the sum, however this approach becomes prohibitively slow for larger ROIs. One way to speed up the calculations would be to use the approach presented in [148] and implemented in [149], however it is non-deterministic so can provide variable results. An alternative approach is to subsample the region of interest with a stride size related to the size of the ROI, thus reducing the number of pixels to enable the use of the straightforward medoid calculation. Here, the latter approach has been utilised where necessary.

To compare the results obtained using a per-channel median and medoid value extraction process, two experiments were carried out. In the first experiment, data collected as part of the SSNR experiment presented in Chapter 3.1.3 was used. This data consists of a series of flash/ no-flash image pairs of a ColorChecker Classic

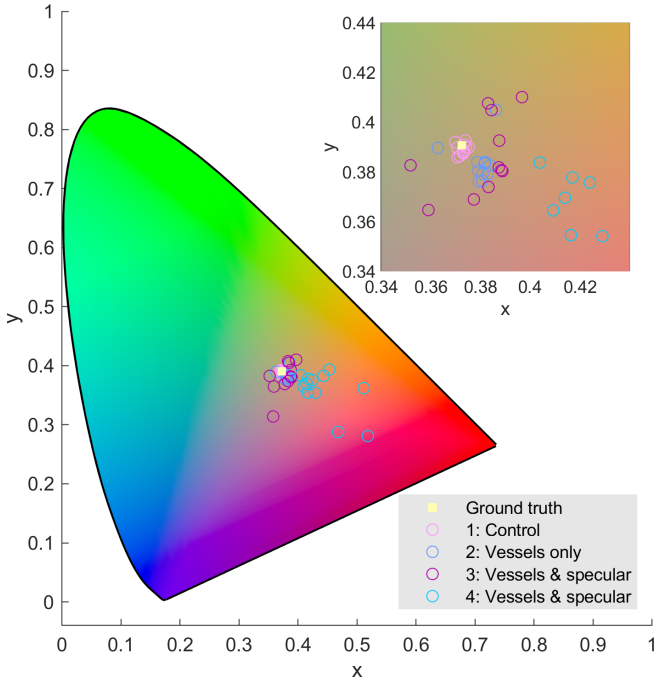
chart captured under varying intensities of ambient light. Since each Classic chart patch is extremely uniform in colour, this experiment enabled a comparison between the approaches for uniform surfaces under non-uniform intensity illumination. The post-subtraction rg values resulting from the two extraction approaches were compared. Across four phones and the 24 patches, it was found that on average the standard deviation of patch values increased by 19% when the medoid was used rather than the per-channel median. This suprisingly suggests that the results are more variable when using the medoid.

In the second experiment, data from the filtering validation experiment presented in Chapter 4.2.3 was used. This data consists of a series of samples designed to mimic a jaundiced sclera with varying levels of blood vessels and specular reflection. The samples are described and depicted in more detail in Figure 4.6. The second experiment enables a comparison between the approaches for non-uniform illumination and non-uniform surfaces. The images in this dataset had large regions of interest, so for medoid calculations a subsampling was carried out using a stride of 5 pixels. The extracted values post mapping to xy space are shown in Figure D.1. It can clearly be seen by comparing the results of the two approaches that the medoid results in (b) are far more scattered than the per-channel median values in (a). An additional note is that when a different stride size was used for the subsampling, the medoid results changed quite significantly, suggesting that for higher levels of outliers the subsampling is an important factor.

Overall, the results from these two experiments suggest that the per-channel median is able to produce more reliable and accurate RGB values for a given ROI and has the advantage of being far more computationally inexpensive. For these reasons, we have used per-channel median values throughout this work.



(a)



(b)

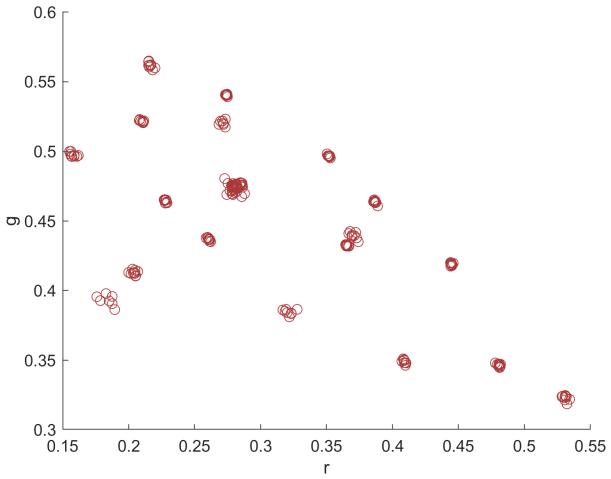
Figure D.1: The xy values of the samples are shown without filtering for a per-channel median in (a). The results for the numbered samples shown in Figure 4.6 are shown in pink, blue, magenta and turquoise respectively. The xy values for the samples also without filtering for a medoid value extraction are shown in (b). Note that not all points are included in the enlarged region in (b) since they are quite significantly spread out.

E. Flash and no-flash exposure settings

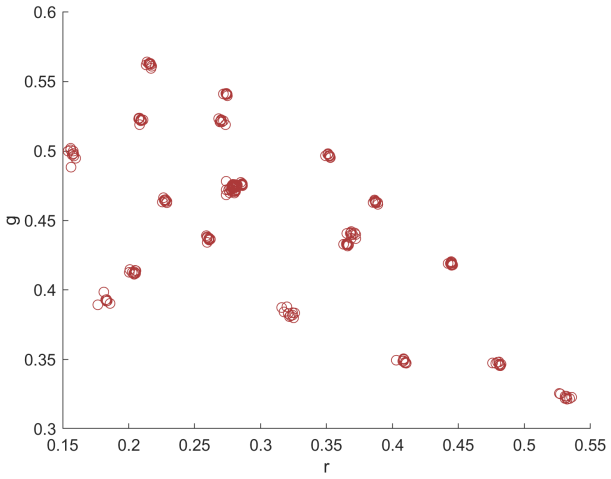
Throughout this work, while the exposure settings for image capture were adjusted based on ambient conditions, the no-flash image always had the same exposure settings as the flash image. This meant that, since the sensors are linear as demonstrated in Appendix C, a simple subtraction of pixel values enabled the effect of ambient light to be minimised.

Theoretically, it should be possible to choose different exposure settings for the no-flash image and then scale the resulting image values according to the ratios of exposure time and ISO. A longer exposure time for the no-flash image (which will inevitably be darker than the flash image) should reduce the noise in the image, hopefully producing more precise results overall as well as images which are easier to segment. However, scaling the no-flash image results may introduce error, for example due to slightly different linear sensor responses or fixed signal offsets at different exposure times.

The following experiment was devised to investigate whether there was a significant benefit to separately exposing flash and no-flash images. A ColorChecker Classic chart was imaged under 10 different ambient light intensities. At each intensity level, the exposure settings for the flash image were selected to produce appropriately exposed images. Four no-flash images were then captured, the first with the same settings as the flash and the next three with up to five times longer exposure times. The rg values for the patches under the varying conditions were inspected for the standard subtraction method and for the scaled exposure settings. Results are shown in Figure E.1 for the usual subtraction method and one example set of scaled results. The first thing to note is that results are very similar between the two approaches. This confirms that it is possible to allow image pairs to expose separately. In order to change approaches we had hoped to see a marked improvement in precision when allowing the no flash image to expose for longer, however the spread amongst data is similar for the two approaches. We therefore conclude that it is not worth allowing the no flash image to separately expose, especially since longer exposure times would increase the likelihood of motion artefacts within image pairs.



(a)



(b)

Figure E.1: rg chromaticity values for ColorChecker patches imaged at different ambient light intensity levels using a Samsung S8 phone after (a) standard subtraction - same exposure settings for pairs of flash and no-flash images, and (b) scaled subtraction - longer exposure times for no-flash images. Data is shown for one example set of scaled results.