



ERCIM "ALAIN BENSOUSSAN"
FELLOWSHIP PROGRAMME



Scientific Report

First name / Family name	Simon Hviid Del Pin
Nationality	Danish
Name of the <i>Host Organisation</i>	NTNU
First Name / family name of the <i>Scientific Coordinator</i>	Seyed Ali Amirshahi
Period of the fellowship	01/01/2022 to 31/12/2022

I – SCIENTIFIC ACTIVITY DURING YOUR FELLOWSHIP

My ERCIM fellowship allowed me to contribute to image and video quality research as a cognitive researcher. It allowed me to apply and further develop my research skills in measuring and quantifying human experience. In the year I have been in this field because of my ERCIM fellowship, I have identified pitfalls in current practices and worked on ways to address them. In the latter part of my stay, I have worked on how researchers can identify and statistically model individual differences in perceived quality. I have also shown that statistical models that include individual differences can better describe data than models which do not.

In the first part of my stay, I focused on learning more about the field by reviewing the literature to generate hypotheses I could experimentally test. Before starting the project, my aim was to generate a new scale for quality, but after reviewing the literature and discussing the topic with my scientific supervisor, Associate Professor Ali Amirshahi, I found more pressing questions to focus on first.

The topics of the matter were if I could statistically demonstrate a non-linear use of the

quality scale, if observers change their ratings throughout the experiment and whether there are individual observer differences in rating.

To investigate this, I designed and conducted several online experiments, which involved creating questionnaires, coding the results and then analysing the data. To do this, I first developed a web-based application for collecting image quality ratings. This involved designing the user interface, programming the application in python, and then testing and debugging it. I created a database of image quality ratings, which involved collecting and organising the data from the web-based application. I made this database publicly available at <https://www.ntnu.edu/web/colourlab/software>

I analysed the data using Bayesian mixed-effects modelling in the statistical software package and in the end, I could statistically demonstrate what I went out to investigate:

I demonstrated the non-linear use of the scale, changes in ratings throughout the experiment, and individual observer differences in ratings. These and a few other results were submitted as a conference paper and were accepted. I was very honoured by one reviewer's assessment: "*Very important topic and a great paper discussing the issue both theoretically and with experimental data. I believe and hope the paper may have a large impact in the field.*" I additionally presented the results at the 11th Colour and Visual Computing Symposium.

For the next step, I wanted to conduct an analysis of user preferences for different image content. This involved creating a new dataset of images. I created one consisting of 4 categories: Sceneries, Objects, People, and Animals. I did this by collecting hundreds of images for each category, manually screening them and analysing them with a so-called objective metric of quality. In the end, I had a set of reference images that had 64 images from each category. Each category should have the same quality according to the metric.

After creating the image set, I ran new experiments. The data from these experiments are still being analysed and written up. However, some of the main conclusions from my main experiments involve testing 16 people and then testing them one way later. I found that there are statistically meaningful individual differences in how they rate content and these are robust enough to perceive when the observers are tested later again.

Integration in the lab

Besides my main research, I advised and discussed ongoing projects with postdocs, PhD-students and master students. This involved providing guidance and advice on various projects and helping to identify potential issues and solutions. I also attended the weekly "Colourlunch" events and presented twice. These events involved listening to

presentations from people associated with the lab and discussing different topics.

In the end, I was so fortunate that the lab leader, Professor Marius Pedersen, hired me to spend the next years investigating image quality in more detail.

Summarizing my scientific year, I conducted experiments to measure and understand how people perceive the quality of digital images. This research supports an ongoing aim to improve the user experience by informing quality optimization for users. To evaluate the quality of an image, most standards use the Mean Opinion Score (MOS), which is the average rating given to an image by all observers. However, the MOS can be controversial because it assumes that all observers are the same, which is not true. I believe researchers should consider individual differences to better understand how people evaluate image quality. By grouping people with similar characteristics, it may be possible to provide tailored media experiences for users.

II – PUBLICATION(S) DURING YOUR FELLOWSHIP

Accepted paper:

Del Pin, S. H., & Amirshahi, S. A. (2022). Subjective quality evaluation: what can be learnt from cognitive science?. CEUR Workshop Proceedings.

Abstract

Subjective ratings given by observers are a critical part of research in image and video quality assessment. Like any other field of science, with subjective data collection, researchers may lack the expertise needed to address the different issues they face. In this study, we review different approaches and find potential pitfalls that generally seem overlooked in quality research. To address these issues, we found six relevant pitfalls relating to recruitment, instructions, experimental design, and data analysis that could be addressed by studies done in the field of cognitive science. Combining accessed datasets from quality research with newly collected data, we statistically demonstrated four of the six pitfalls: observers used the scale non-linearly; ratings can change throughout the experiment; features can influence individual observers differently; and allowing observers to decide how many ratings they give can lead to biases. We need additional data to investigate the two pitfalls related to instructions and recruitment. Our findings suggest that pitfalls which might not be initially clear to researchers in the field of image and video processing can still have an empirically demonstrable influence on the data. While this article will not solve every issue, it will try to suggest improvements that researchers can readily employ.

In preparation:

Del Pin, S. H., & Amirshahi, S. A.

Image Quality Experiments: Are Individual Differences a Problem or Opportunity?

Abstract

Current standards for measuring quality use mean opinion scores which average out individual differences, but those differences may be meaningful. Borrowing from cognitive science, we propose an experimental and statistical method to quantify individual differences. We created a new image set consisting of 256 images from the categories: Scenery, Objects, People and Animals. All reference images were shown to 16 observers along with 3 types of colour distortion that were equally distributed amongst the content categories. One week after completion, we asked the observers to rate the same reference images again. Using bayesian mixed-effects modelling, we created statistical models that can account for individual differences. The model including individual differences with an interaction of content and distortions had higher predictive power than models omitting them. Moreover, all but one observer had medium or large correlations in their individual effects when re-measured one week later. Differences in quality ratings thus seem meaningful and stable between individuals. As an exploratory step, we clustered the observers and found that $\frac{2}{3}$ of them fit into 3 groups. We believe individual, robust differences represent a problem for how data is typically analysed today but also an opportunity for individually tailored content delivery in the future.

III – ATTENDED SEMINARS, WORKHOPS, CONFERENCES

Gave a scientific talk at The 11th Colour and Visual Computing Symposium 2022, September 08–09, 2022, Gjøvik, Norway.

Attended Colourlab strategy seminars to discuss long term plans for the lab. April 21- 22, 2022, Brumunddal, Norway.

IV – RESEARCH EXCHANGE PROGRAMME (REP)

Visited Sebastian Bosse, Head of Interactive & Cognitive Systems Group at Fraunhofer HHI – Heinrich Hertz Institute. December 06–13, 2022, Berlin, Germany.

When I visited Sebastian Bosse, I first gave a scientific talk about my ongoing work with subjective data to his lab. Following the talk, Dr. Bosse and I discussed whether the four categories I created had some lower level differences. The features of interest was colourfulness which I had previously heard about and the local standard deviation which was new to me.

The local standard deviation is a measure of how different the values of the pixels are in different parts of an image. By calculating the standard deviation for each pixel's neighborhood, we can see which parts of the image have a lot of variation in the pixel values and which parts have less variation. This can be useful for finding interesting features of the image, like areas with a lot of texture or contrast. For example, if we have a picture of a rocky mountain landscape, the local standard deviation can help us see which parts of the image have the most texture, like the rocky cliffs or the rough ground.

We aim to perform further analysis on these topics and submit them to a future conference.