Aarhus University

Research & Development Project

PERCEPTUAL IMAGE EVALUATION

Author:

Line Aggerbo Johansen
201302451

Supervisor:

Christian Fischer Pedersen

January 18, 2016

TABLE OF CONTENTS

1

INTRODUCTION

This report describes the process of a Research and Development project carried out at Aarhus University. The project is developed as a subproject of a Master's Thesis which will be carried out in in spring 2016. The purpose of the Master's Thesis is to conduct a statistical comparison of subjective image quality evaluation with objective image quality metrics, e.g. MSE, PSNR, and SSIM.

The purpose of this Research and Development project is to analyze multiple subjective image quality assessment techniques and from this choose a single assessment technique to be implemented. The outcome of this project is an application which will be used for evaluation in the Master's Thesis project.

All design and implementation choices made throughout this project is reasoned and described in this report together with screen shots from the final implementation of the application.

2

ASSESSMENT TECHNIQUES

When collecting subjective image quality evaluations, numerous techniques exist. Three of the existing techniques (Absolute Category Rating, Degradation Category Rating, and Subjective Assessment Methodology for Video Quality) are described in this chapter. The benefits and drawbacks of using each assessment technique according to this project are also discussed in this chapter.

In a subjective image quality assessment scenario, an observer is presented for one or more stimulus (video or image) at a time and must rate the presented stimulus.

2.1 Absolute Category Rating

Absolute category rating (ACR) is a discrete rating system most commonly used for video quality testing [7] but is also used for image quality testing [12]. The rating scale used in ACR has five categories; bad, poor, fair, good, and excellent. It can be a five-grade rating scale from 1 (bad) to 5 (excellent). This is referred to as the ACR-5. It can also consist of a higher number of grades like the ACR-9 or ACR-11 which are presented together with the ACR-5 in both [7] and [12]. The labels are similar for all the ACR rating scales.

The process of ACR is to present each stimulus to the observer one at a time and only once. For stimuli being images, each image is presented to the observer for approximately 10 seconds followed by a gray screen from where the observer must rate the presented stimulus. The duration of the rating process depends on the rating scale but should not exceed 10 seconds. The process is illustrated in Figure 2.1.

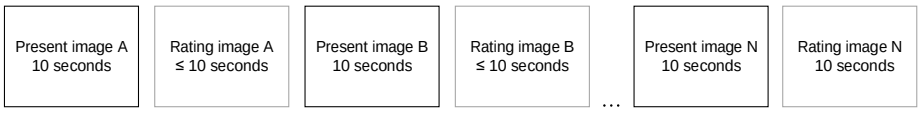

Figure 2.1: The process of ACR. A stimulus is presented for 10 seconds followed by a gray screen from where the observer can rate the stimulus. This process is repeated until all stimuli have been rated.

At the end of the process the observer should have assigned a rating score to all presented stimuli. When all observers have assigned a rating score for each stimulus a mean opinion score (MOS) is found for each stimulus by averaging all rating

scores.

2.2 Degradation Category Rating

Degradation category rating (DCR) is similar to ACR a discrete rating system. This system is most commonly used when comparing an original stimulus with a modified version of the same stimulus [7]. Due to the comparison, the labels for the DCR system is a little different; very annoying, annoying, slightly annoying, perceptible but not annoying, and imperceptible. Besides the labels, the rating scale is the same as in ACR where 1 is 'very annoying' and 5 is 'imperceptible'.

The process of DCR is much similar to the ACR system. The difference lies within the number of presented stimuli (one for ACR and two for DCR). First the observer is presented to the original stimuli followed by the modified version. Both stimulus must be presented before the rating can take place. Between the two stimulus there is a short break of two seconds where the observer is presented to a gray screen. Similar to ACR the the presentation time for the stimulus is approximately 10 seconds when stimuli are images. As with ACR a MOS can be found for each stimulus. The process is illustrated in Figure 2.2.

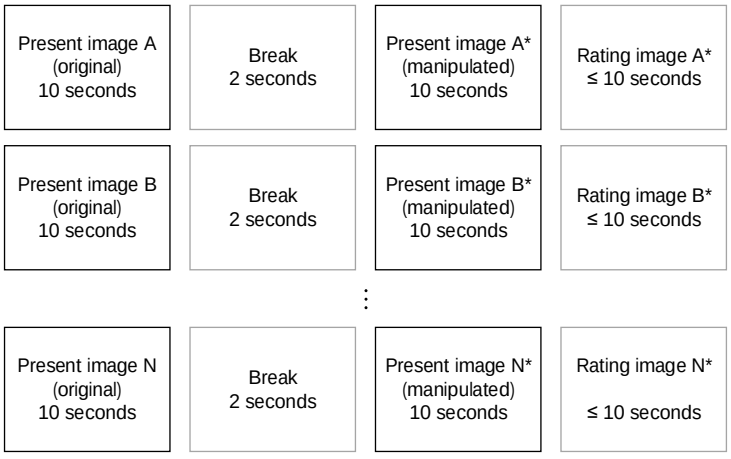

Figure 2.2: The process of DCR. Two stimuli are presented to the observer with a small break in between. The observer rates the second stimulus (modified) compared with the first stimulus (original).

2.3 Subjective Assessment Methodology for Video Quality

Subjective assessment methodology for video quality (SAMVIQ) is another assessment technique which differs from the other assessment techniques in multiple ways even though the labels; bad, poor, fair, good, and excellent are the same as the ones used in ACR. As indicated by the name this technique is developed for use in video quality evaluation but has also been used for image quality evaluation as in [12].

The rating scale in SAMVIQ differs from the two aforementioned techniques. The range of the scale goes from 0 to 100 where the label 'bad' is associated with the score 10 and the label 'excellent' is associated with the score 90. This scale has shown to be unnecessarily large and the observers ratings might be grouped in 5 or 9 groups over the rating scale range [12].

The process of SAMVIQ also differs from earlier described techniques. In SAMVIQ the process is split into a number of scenes where each scene contains a number of different variations of the same stimulus. The explicit reference is the original stimulus and is known to the observer. The hidden reference is also the original stimulus but is unknown to the observer. Within a scene, the observer can browse freely between the different variations of the stimulus. Also the observer is able to evaluate the stimulus while observing it. When all stimulus has been rated, another scene might be presented and the process repeats itself. Notice that the order of the different variations of a stimulus changes for every scene. The process is illustrated in Figure 2.3.

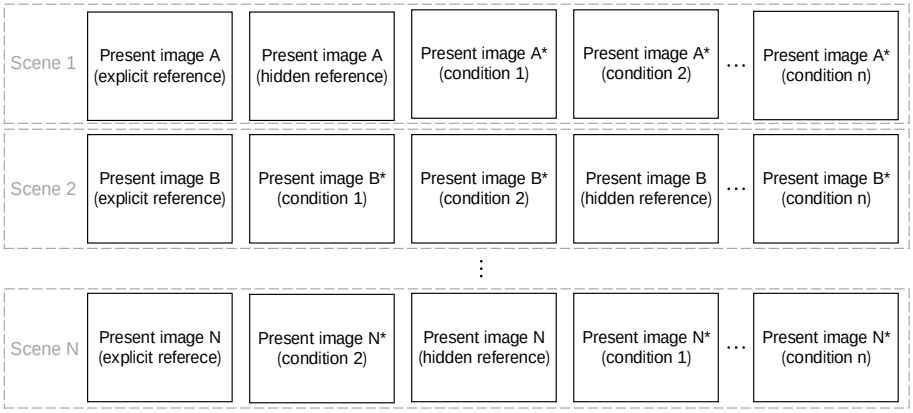

Figure 2.3: The process of SAMVIQ. Stimuli are split into scenes containing the same stimulus under different conditions (modified in different ways). All stimuli are rated before the observer gets to see the next scene.

The illustration in Figure 2.3 shows the use of a hidden reference. In SAMVIQ the use of a hidden reference is mandatory [8].

2.4 Discussion

The choice of assessment technique highly depends on the application which it is used for. In this project the focus is on evaluation of images which have been subject to a certain kind of lossy compression. The decrease in image quality, hence the error, between the compressed and uncompressed images, will vary according to the specific compression algorithm. The hardest thing to evaluate is the smaller errors which is difficult to see with the naked eye. A key aspect is therefore to provide the observer with the best opportunity to see and evaluate these types of errors.

The greatest benefit of ACR is that it is easy to understand and implement. The drawback, with respect to this project, is that the ACR does not have an explicit reference which is recommended in evaluations when the impairments of the images are small [7]. Both DCR and SAMVIQ has the ability of using an explicit reference.

One difference of DCR and SAMVIQ is the rating scale. As mentioned before, experiments show that the rather large rating scale in SAMVIQ is superfluous [12] and the 5, 9 or 11 grade rating scale in DCR would be enough.

Despite the large rating scale the SAMVIQ assessment technique is chosen for this project due to the process of the technique. In DCR the reference is first shown followed by the image for evaluation and then these two steps are repeated. This is useful when evaluating many different images. In SAMVIQ the reference image together with different versions of this image are shown simultaneously. This is ideal for this project where the same image will be compressed in different ways.

A comparison of ACR and SAMVIQ showed that SAMVIQ needed less observers to obtain the same normalized confidence interval as ACR. The ACR-11 also needed fewer observers to obtain the same normalized confidence interval as the ACR-5 according to the same comparison [12]. This indicates that the 5 grade rating scale from ACR and DCR requires a higher number of observers than the 100 grade rating scale from SAMVIQ.

3

DESIGN

The principals of SAMVIQ should be incorporated in a software application which the observer can use to evaluate images. The first choice to make is whether the application should be PC or web based. PC applications are confined to a physical location and hence have usability constraints. Web applications on the other hand make it convenient for the observer to access the application from any location using the Internet. Due to the exact evaluation scenario is unknown it is chosen to make the application web based in order to enhance flexibility.

The programming language used for the application is specified by the supervisor of this project to be Python. To ease the development process it is chosen to base the development on a web application framework. The chosen framework is Django [4] which is the largest Python-based web framework. Online remarks of the Django framework includes; "Django aims to include all the batteries a web application will need so developers need only open the box and start working, pulling in Django's many modules as they go." [2] and "Django is great for developers who want to build something quickly with powerful built-in tools"[9]. Alternatives to Django would be Flask[11] and Pyramid[10].

3.1 The web application

The framework architecture is inspired by the Model View Controller (MVC) architecture and consists of four instances; a model, a view, a template, and an URL dispatcher. The model describes the data in the database. The template defines how observers see things in the browser. The view controls what observers see and the communication between the template and the model. The URL dispatcher maps the requested URL to a view function and calls it. The architecture is illustrated in Figure 3.1.

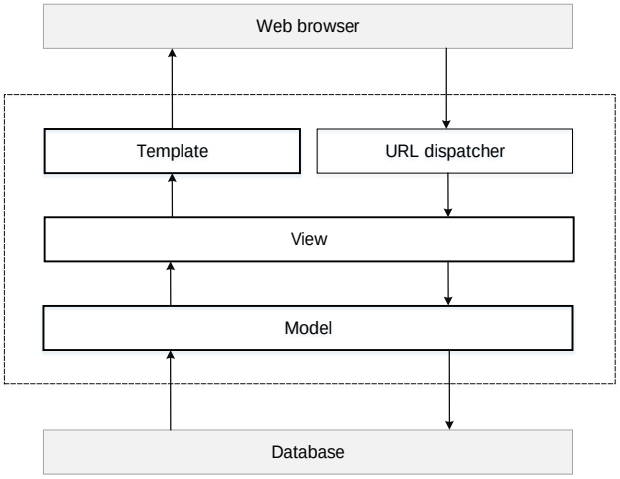

Figure 3.1: The architecture of the Django framework.

3.1.1 Model

The model defines the data and the interaction with the database. The database is a SQLite database which consists of three tables; Observer, Image, and Rating. This is illustrated in Figure 3.2.

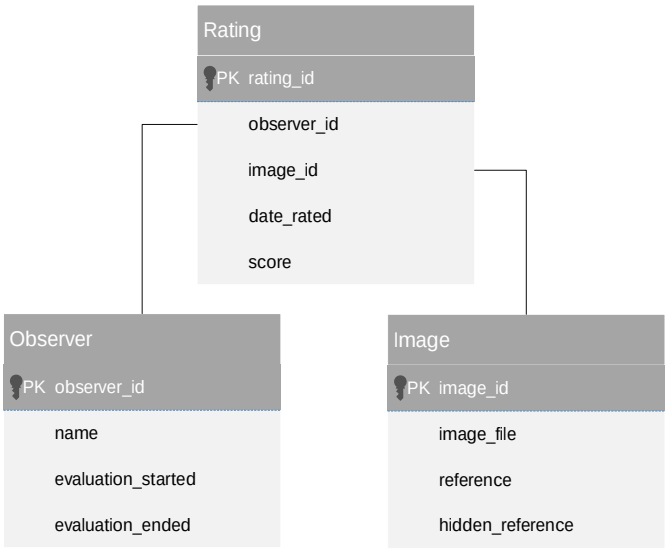

Figure 3.2: The architecture of the database.

The Observer table will contain information about all observers who have completed the quality evaluation. Each observer is created with a unique id, a name and the

start and end time of the evaluation. The time stamps will be used to investigate the amount of time the observer has spent on the evaluation. This is relevant only if the evaluation is suspicious.

The Image table will contain information about all images in the evaluation. Each image is created with a unique id, a path to the image(image_file), and the information of whether or not the image is being used as a reference or a hidden reference.

The Rating table will contain information about all ratings which have been made throughout all evaluations. Each rating will be created with a unique id, two foreign keys (observer_id and image_id) that refer to a primary key in the Observer and Image table respectively, the time of the rating and the score of the particular image.

3.1.2 Template

The template is the front-end of the web application which typically consists of HTML/CSS. Django offers a powerful template engine that supports the developer in doing presentation logic. The design of the template is done with respect to descriptions of the SAMVIQ interface found in [6] and [5]. A mockup of the template is illustrated in Figure 3.3.

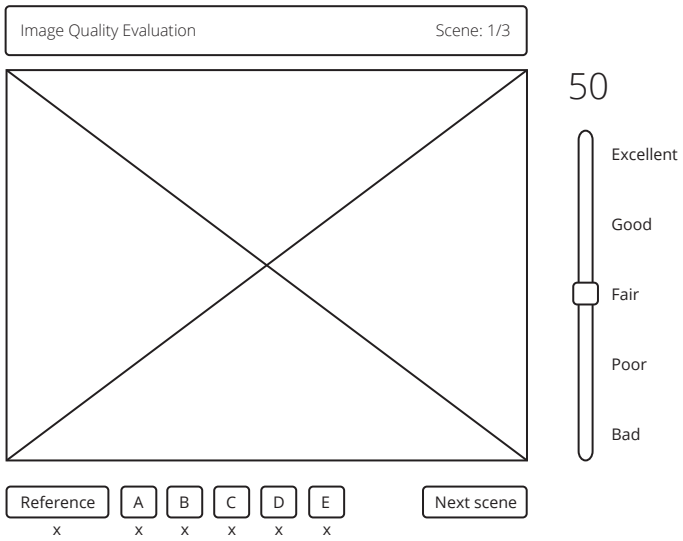

Figure 3.3: Mockup of the template.

According to both articles the template should consist of a large image in the middle. Below the image navigation buttons should be placed, one labeled as 'Reference' and the rest labeled alphabetically. A button for changing to the next scene should also be present.

Both articles state that the rating mechanism as a slider located on the right side of the image. The slider has the five rating labels spread out over the total range. The score is shown both at the top of the slider and under the button which links to the image.

According to both articles the background color is chosen to be mid-range gray although this is not shown in the mockup. Above the image, information about the process is shown in order to let the observer be aware about the number of remaining scenes to evaluate.

3.1.3 View

The view performs the requested action from the observer which typically involves reading from and writing to the database. This view instance is responsible for creating a new observer when a new evaluation is started. It also has to save all ratings done in a scene before changing to the next scene.

4

IMPLEMENTATION

The three parts of the web application; model, template and view is implemented with respect to the design choices made in chapter 3. The final implementation of the web application will be presented in this chapter.

The template is implemented in the languages HTML, CSS, and JavaScript. HyperText Markup Language (HTML) describes the elements which form the building blocks of the template. The Cascading Style Sheet (CSS) sets the visual style of the building blocks including aspects such as colors and fonts. JavaScript is used to manipulate the HTML elements. In this application JavaScript allows for each scene to contain multiple images while only presenting one at a time. A screen shot of the final template is shown in Figure 4.1.

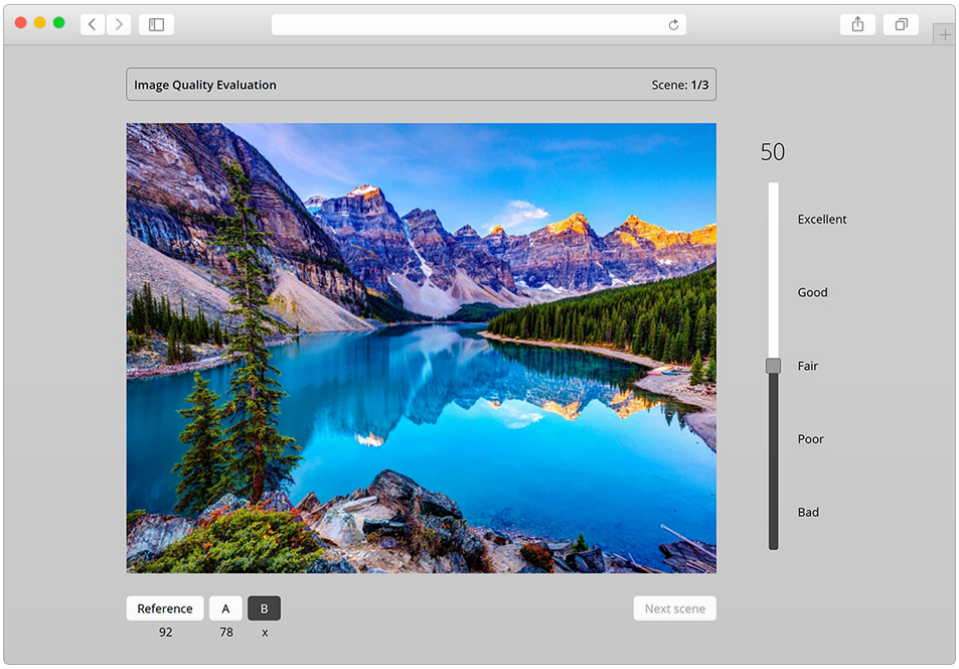

Figure 4.1: Screenshot of web application. The image for evaluation has not been modified and is a random nature image [1].

The web application is implemented such that the number of images in every scene is dynamic. Each scene might contain up to 10 images which is the maximum number of images for a scene in SAMVIQ [8].

Noteworthy features of the template is among others that the button which links to next scene is disabled until all images are rated as the SAMVIQ process states [5]. Another feature is that the rating score is shown dynamically above the slider and below the corresponding button.

The result of the implemented model is shown in Figure 4.2, Figure 4.3, and Figure 4.4. The database can be investigated using the free database browser 'SQLite Browser'[3]. The information in the shown figures is dummy data created during the application test phase.

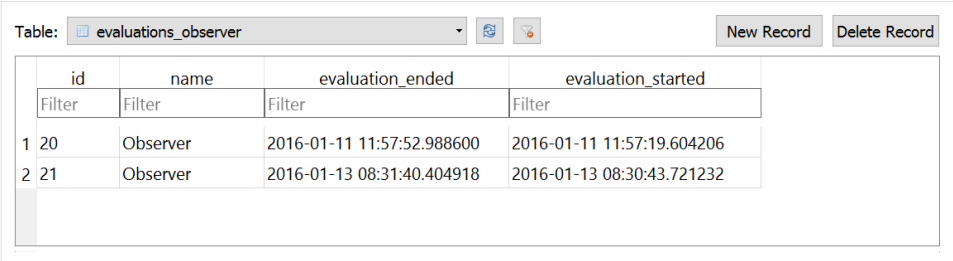

Figure 4.2: List of information about all observers in the Observer table.

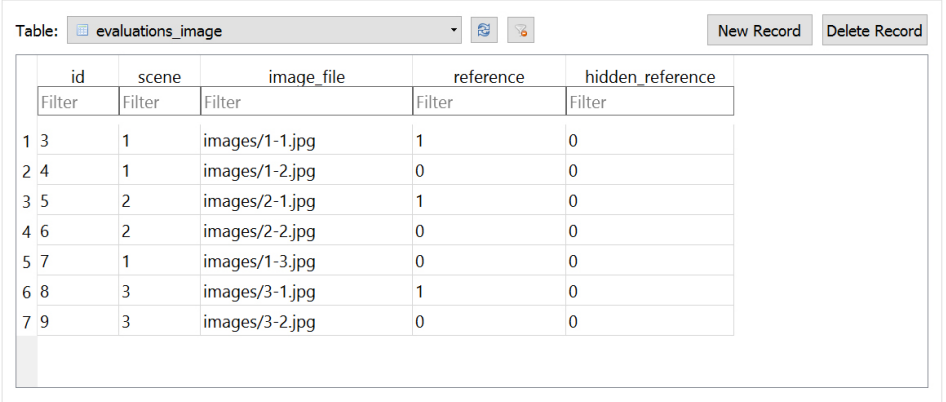

Figure 4.3: List of information about all images in the Image table.

Table: evaluations_rating

New Record Delete Record

	id	date_rated	image_id	bserver_i	score
	Filter	Filter	Filter	Filter	Filter
1	34	2016-01-11 11:57:33.019586	3	20	71
2	35	2016-01-11 11:57:33.019586	4	20	58
3	36	2016-01-11 11:57:33.035250	7	20	88
4	37	2016-01-11 11:57:45.254142	5	20	61
5	38	2016-01-11 11:57:45.269746	6	20	66
6	39	2016-01-11 11:57:52.941744	8	20	63
7	40	2016-01-11 11:57:52.957348	9	20	70
8	41	2016-01-13 08:31:21.880135	3	21	100
9	42	2016-01-13 08:31:21.908155	4	21	81
10	43	2016-01-13 08:31:21.930792	7	21	24
11	44	2016-01-13 08:31:31.895668	5	21	78
12	45	2016-01-13 08:31:31.921579	6	21	28
13	46	2016-01-13 08:31:40.316666	8	21	87
14	47	2016-01-13 08:31:40.343683	9	21	100

Figure 4.4: List of information about all rating in the Rating table.

5

CONCLUDING REMARKS

The purpose of this project was twofold. First, three image quality evaluations assessment techniques were analyzed. The technique chosen for implementation was SAMVIQ which is best suited for rating images under various conditions where the impairments are small.

Secondly, an application was developed based on the chosen assessment technique. The application is web-based and is able to demonstrate the core concepts of the SAMVIQ methodology.

5.1 Future work

The web application is not quiet ready to be used for image quality evaluation assessment. The application will need a start page together with an initial training scene from which an observer will be carefully trained in the process of the evaluation [8].

It would be beneficial to have an end page for the web application from where the observer would be thanked for participating. This page could contain information about the evaluation and the use of the assessed data.

Bibliography

[1] Wallpapers Craft. https://wallpaperscraft.com/download/nature_mountains_sky_lake_clouds_81150/800x600, 2016.

[2] Ryan Brown. Django vs Flask vs Pyramid: Choosing a Python Web Framework. https://www.airpair.com/python/posts/django-flask-pyramid, 2015.

[3] Mauricio Piacentini et. al. DB Browser for SQLite. http://sqlitebrowser.org/, 2014.

[4] Django Software Foundation. Django. https://www.djangoproject.com/, 2016.

[5] Quan Huynh-Thu, Matthew Brotherton, David Hands, Kjell Brunnström, and Mohammed Ghanbari. Examination of the samviq methodology for the subjective assessment of multimedia quality. *Proc. International Workshop on Video Processing and Quality Metrics for Consumer Electronics (VPQM07), Scottsdale, USA*, page 17, 2007.

[6] International Telecommunication Union. Methodology for the subjective assessment of video quality in multimedi applications. *Recommendation ITU-R BT.1788*, pages 1–13, 2007.

[7] ITU-T Recommendation P.910. Subjective video quality assessment methods for multimedia applications. Technical report, 1999.

[8] Franc Kozamernik, Paola Sunna, E Wyckens, and DI Pettersen. Subjective quality of internet video codecs - Phase 2 evaluations using SAMVIQ. *EBU Technical Review*, (January), 2005.

[9] Clayton Parker. 4 Python Web Frameworks Compared. http://www.sixfeetup.com/blog/4-python-web-frameworks-compared, 2013.

[10] Pylons Project. Pyramid. http://www.pylonsproject.org/, 2011.

[11] Armin Ronacher. Flask. http://flask.pocoo.org/, 2014.

[12] David M Rouse, Romuald Pépion, Patrick Le Callet, and Sheila S Hemami. Tradeoffs in subjective testing methods for image and video quality assessment. In Bernice E. Rogowitz and Thrasyvoulos N. Pappas, editors, *Proc. SPIE*, volume 7527, pages 75270F–75270F–11, feb 2010.

A

APPENDIX: DETTE ER EN TEST