Image processing algorithms as artistic tools in digital cinema

Itziar Zabaleta Razquin

DOCTORAL THESIS UPF / 2021

THESIS SUPERVISOR
Marcelo Bertalmío

Dept. of Information and Communication Technologies

Acknowledgement

First of all, I would like to thank my supervisor Marcelo Bertalmío for its encouragement, help and guidance, that made possible the research leading to this thesis.

Very special thank you for my colleagues, for always being patient, supportive, for having fun and sharing adventures during all these years. This work would have not been possible without you: Adrián, Alex, Trevor, Javier, Raquel, Praveen, Waqas, Gabriela, Antoine, and the Nora team.

A warmly thank you to my family and close friends, who have always helped me and supported me.

Abstract

The industry of cinema has experienced a radical change in the last decades: the transition from film cinematography to its digital format. As a consequence, several challenges have appeared, but, at the same time, many possibilities are open now for cinematographers to explore with this new medium.

In this thesis, we propose different tools that can be useful for cinematographers while doing their craft. First, we develop a tool for automatic color grading. It is a statistics-based method to automatically transfer the style from a graded image to unprocessed footage. Some advantages of the model are its simplicity and low computational cost, which make it amenable for real-time implementation, allowing cinematographers to experiment on-set with different styles and looks.

Then, a method for adding texture to footage is created. In cinema, the most commonly used texture is film grain, either directly shooting on film, or adding synthetic grain later-on at post-production stage. We propose a model of "retinal noise" which is inspired by processes in the visual system, and produces results that look natural and visually pleasing. It has parameters that allow to vary widely the resulting texture appearance, which make it an artistic tool for cinematographers. Moreover, due to the "masking" phenomenon of the visual system, the addition of this texture improves the perceived visual quality of images, resulting in bit rate and bandwidth savings. The method has been validated through psychophysical experiments in which observers, including cinema professionals, prefer it over film grain emulation alternatives from academia and the industry.

Finally, we introduce a physiology-based image quality metric, which can have several applications in the image processing field, and more specifically in the cinema and broadcasting context: video coding, image compression, etc. We study an optimization of the model parameters in order to be competitive with the state-of-the-art

quality metrics. An advantage of the method is its reduced number of parameters, compared with some state-of-the-art methods based in deep-learning, which have a number of parameters several orders of magnitude larger.

Resumen

La industria del cine ha experimentado un cambio radical en las últimas décadas: la transición de su soporte fílmico a la tecnología del cine digital. Como consecuencia, han aparecido algunos desafíos técnicos, pero, al mismo tiempo, infinitas nuevas posibilidades se han abierto con la utilización de este nuevo medio.

En esta tesis, se proponen diferentes herramientas que pueden ser útiles en el contexto del cine. Primero, se ha desarrollado una herramienta para aplicar *color grading* de manera automática. Es un método basado en estadísticas de imágenes, que transfiere el estilo de una imagen de referencia a metraje sin procesar. Las ventajas del método son su sencillez y bajo coste computacional, que lo hacen adecuado para ser implementado a tiempo real, permitiendo que se pueda experimentar con diferentes estilos y 'looks', directamente on-set.

En segundo lugar, se ha creado un método para mejorar imágenes mediante la adición de textura. En cine, el grano de película es la textura más utilizada, ya sea porque la grabación se hace directamente sobre película, o porque ha sido añadido a posteriori en contenido grabado en formato digital. En esta tesis se propone un método de 'ruido retiniano' inspirado en procesos del sistema visual, que produce resultados naturales y visualmente agradables. El modelo cuenta con parámetros que permiten variar ampliamente la apariencia de la textura, y por tanto puede ser utilizado como una herramienta artística para cinematografía. Además, debido al fenómeno de enmascaramiento del sistema visual, al añadir esta textura se produce una mejora en la calidad percibida de las imágenes, lo que supone ahorros en ancho de banda y tasa de bits. El método ha sido validado mediante experimentos psicofísicos en los cuales ha sido elegido por encima de otros métodos que emulan grano de película, métodos procedentes de academia como de industria.

Finalmente, se describe una métrica de calidad de imágenes, basada en fenómenos fisiológicos, con aplicaciones tanto en el campo del

procesamiento de imágenes, como más concretamente en el contexto del cine y la transmisión de imágenes: codificación de vídeo, compresión de imágenes, etc. Se propone la optimización de los parámetros del modelo, de manera que sea competitivo con otros métodos del estado del arte . Una ventaja de este método es su reducido número de parámetros comparado con algunos métodos basados en deep learning, que cuentan con un número varios órdenes de magnitud mayor.

Resum

La indústria de cinema ha experimentat un canvi radical en les últimes dècades: la transició del seu suport fílmic a la tecnologia del cinema digital. Com a conseqüència, han aparegut alguns desafiaments tecnològics, però, a el mateix temps, infinites noves possibilitats s'han obert amb la utilització d'aquest nou mitjà. En aquesta tesi, es proposen diferents eines útils en el context de cinema. Primer, s'ha desenvolupat una eina per aconseguir *color grading* automàticament. És un mètode basat en estadístiques d'imatges, que transfereix l'estil d'una imatge de referència a metratge sense processar (o a què encara no se li ha aplicat *color grading*). Els avantatges de l'mètode són la seva senzillesa i baix cost computacional, que el fan adequat per a ser implementat a temps real, permetent que es pugui experimentar amb diferents estils i 'looks', directament on-set.

Segon, s'ha creat un mètode per millorar imatges mitjançant l'addició de textura. En cinema, el gra de pel · lícula és la textura més utilitzada, ja sigui perquè la gravació es fa directament sobre pel · lícula, o perquè ha estat afegit a posteriori en contingut gravat en format digital. En aquesta tesi es proposa un mètode de 'soroll retinià', inspirat en processos de sistema visual, que produeix resultats naturals i visualment agradables. El model compta amb paràmetres que permeten variar àmpliament l'aparença de la textura, i per tant pot ser utilitzat com una eina artística per cinematografia. A més, a causa de el fenomen d'emmascarament de el sistema visual, a l'afegir aquesta textura es produeix una millora en la qualitat percebuda de les imatges, la qual cosa suposa estalvis en ample de banda i taxa de bits. El mètode ha estat validat mitjançant experiments psicofísics en els quals ha estat elegit per sobre d'altres mètodes que emulen gra de pel · lícula, tant d'acadèmia com d'indústria.

Finalment, es descriu una mètrica de qualitat d'imatges, basada en fenòmens fisiològics, amb aplicacions tant en el camp de l'processament d'imatges, com més concretament en el context de el cinema i la transmissió d'imatges: codificació de vídeo, compressió d'imat-

ges, etc . Es proposa l'optimització dels paràmetres de el model, de manera que sigui competitiu amb altres mètodes de l'estat de l'art. Un avantatge d'aquest mètode és el seu reduït nombre de paràmetres comparat amb alguns mètodes basats en deep learning, que compten amb un número diversos ordres de magnitud major.

Contents

List of Figures

List of Tables

1 Introduction

Many possibilities have emerged for cinematographers with the transition to digital format. This new medium offers many options to explore in terms of special effects, color manipulation, etc. In this thesis, we focus on different methods that can be useful for movie creators as artistic tools. The algorithms behind these methods explore active topics of research in the image processing field. First, we present an efficient solution for automatic color grading, then, we propose a method to add texture to footage, and finally, a physiology-based quality metric and its optimization process are explained.

Usually, footage is processed at post-production stage: colors are modified, effects are added, and the visual "look" of the movie is created. However, leaving all these decisions to post-production stage prevents cinematographers from experimenting on-set and making the necessary adjustments. The result is that more and more movies tend to have a similar look. The proposed tools are designed to be used on-set, allowing experimentation.

Color grading is a common process in cinema, video, and photography, which consists in altering the colors of images in order to develop an appropriate style for the image. Usually, this process is done manually and it is very costly in terms of budget and time. Therefore, automatic color grading methods are very useful in the

cinema and photography context. Moreover, real-time color grading allows cinematographers to experiment with these looks on-set. There exists a vast literature about methods to transfer the style between images. However, some of them produce results whose lack of photo-realism is not acceptable for the high-quality standards of cinema. Other methods, which produce photorealistic results, are computationally very costly, so they cannot be implemented for real-time applications. Moreover, few methods exist which transfer the style to video footage, as most of them are oriented to transform still images.

The second problem we address in this thesis is the addition of texture to images. Usually, the goal of adding texture is to generate visually pleasing images and to improve their perceived quality. The most commonly used texture in cinema and photography is film grain. However, shooting in film nowadays is unpractical and expensive, therefore, synthesized film grain is added to digital images at post-production stage. The main disadvantage of film grain models is the high computational cost, which makes them unpractical for real-time implementations, preventing cinematographers from using it on-set.

Our last application focus on image quality assessment. Some simple but effective quality metrics are based on the mean square error (e.g. PSNR), however, these metrics are not very well correlated with perceived visual quality. Therefore, these metrics have been improved using models that mimic the early stages of the visual system. Some state-of-the-art approaches, based on deep-learning, produce significant results but they have a large amount of parameters to be optimized. Moreover, most of these models are not based on the human visual system processes.

1.1 Contributions

Our first contribution in this thesis is the creation of a method for automatic, real-time color grading. With this purpose, we have de-

veloped a style transfer algorithm based on image statistics, which is amenable to produce real-time results due to its low computational cost. The code of the method can be found at: `https://github.com/izabalra8/VideoStyleTransfer`.

Our second contribution is the proposal of a model for generating "retinal noise", which is a texture inspired by processes in the visual system. The model is simple, its computational cost is low, and it has parameters that allow cinematographers to experiment onset with different textures. The method was patented under the patent name: "Computer-implemented method for adding texture to a digital image". The code of the method can be found at: `https://github.com/izabalra8/retinalNoise`.

Our last contribution is related to image quality assessment. With this purpose, a physiology-based quality metric is presented, and its optimization process is explained. The model, unlike most vision models in the vision science literature, assumes a non-linear response of neurons. It counts with a reduced number of parameters (4 parameters), compared to the state-of-the-art deep-learning based methods. The method achieves significant results in benchmark image quality databases such as TID2008, TID2013, LIVE, or CSIQ. The code of the method can be found at: `https://github.com/izabalra8/INRF-IQmetric`.

1.2 Publications

This thesis is based on our work described in the following papers.
Journals

- Itziar Zabaleta, Mateo Cámara, César Díaz, Trevor Canham, Narciso García, and Marcelo Bertalmío. Retinal noise emulation: A novel artistic tool for cinema that also improves compression efficiency. IEEE Access, 8:67263–67276, 2020.

- Itziar Zabaleta and Marcelo Bertalmío. Photorealistic style

transfer for video. Signal Processing: Image Communication, 95:116240, 2021.

Conferences

- Itziar Zabaleta and Marcelo Bertalmío. In-camera, photorealistic style transfer for on-set automatic grading. SMPTE Annual Technical Conference & Exhibition, 2018.
- Itziar Zabaleta and Marcelo Bertalmío. Photorealistic style transfer for cinema shoots. Colour and Visual Computing Symposium (CVCS), 2018 (best student paper award).

1.3 Thesis outline

Hereby, we introduce the structure of the thesis, following the Introduction. Chapter 2 is an introductory section describing the evolution of the cinema since its beginning to the present day. This description serves as a context for understanding the current necessities for digital cinema. Chapter 3 explains some important concepts concerning digital cinema, which are important for the following chapters. Chapter 4 is a description of processes present in the human visual system, and the consequent visual models derived from them. The chapter finishes with an explanation of the most commonly used color spaces, which will be mentioned in the different chapters of this thesis. In Chapter 5, we propose an efficient automatic color grading method, which produces photorealistic results with a low computational cost. Chapter 6 illustrates how the perceived quality of images can be improved by the addition of certain type of texture. We also propose a "retinal noise" model for adding texture, which produces visually pleasing images, and can be used with coding efficiency purposes. Chapter 7 describes a physiology-based image quality metric, and the optimization of the model parameters is explained in detail.

2 From film to digital cinema

2.1 Beginning

The creation of images is inherent to human history, since prehistoric times with cave paintings as one of the earliest examples, until the current century, with images being of crucial importance in the culture. Pictographic communication not only reflects the contributions of a specific culture, through paintings, sculptures, drawings, etc, but it also derives in intellectual and scientific progress.

For instance, during the Renaissance, Italian artists developed the laws of perspective as a consequence of their attempts of giving a faithful representation of the reality. During that time, the *camera obscura* was created. It was used as an aid for drawing and painting, to trace real-world scenes before transferring them to canvas. This invention was developed further into the photographic camera in the first half of the 19th century, and it was the technology of what sometime later would produce moving images. A *camera obscura* is a darkened room or a box with a small hole at one side through which light enters and reflects on the opposite interior wall. An image of the external objects is projected within the wall inside the box or room. In the 17th century, the camera obscura became small enough

to be portable, and around that time lenses were also introduced to focus light. Therefore, the basics for the photographic camera existed, but it was not until the 19th century, with the development of light-sensitive materials, that photography was invented.

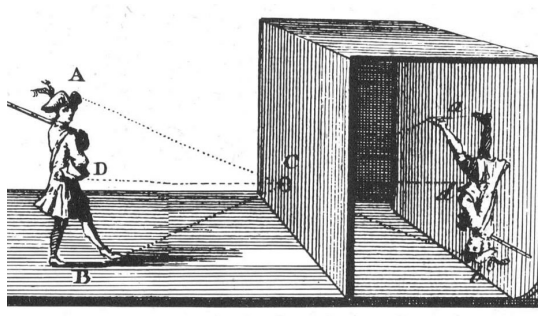

Figure 2.1. Illustration of how camera obscura works. Figure from `https://commons.wikimedia.org/wiki/File:001_a01_camera_obscura_abrazolas.jpg`.

Along with the invention of photography, other optical devices were developed at that time. The *magic lantern* dates from the 17th century. It is a device to project light through a transparent material onto a surface, and it could be considered a predecessor to the modern cinema projectors. At that time, this device was used as an aid for artists, in an attempt to achieve realism in depictions of the surrounding world.

During the late eighteenth and early nineteenth centuries, in the context of the first industrial revolution, amusement and education were bound together. At that time, there was no separation between art and science, or between popular amusement and the development of scientific instruments. In fact, the emergence of photography and film is related to optical devices such as magic lanterns, phantasmagoria, dioramas, etc created with the purpose of popular entertainment. Magic, visual illusions, the scientific study of visual phenomena and a development of technology occurred simultaneously during those years [91]. In 1824, British doctor Peter Mark Roget in his book

"The persistence of vision with regard to moving objects", explained the visual illusion by which the brain interprets rapid sequential, still images as continuous motion. The *thaumatrope*, a popular disc-shaped toy at that time, demonstrated Roget principle: when spun, separate images on either side of the disc appeared to be in the same picture. This optical toy is considered as the earliest antecedent of motion pictures and animation, the images of which are comprised of many individual still images. Some years later, between 1830 and 1860, a list of motion-image devices appeared: the *Phenakistiscope* or the *Praxinoscope* are some examples. The illusion of moving images was created by using rotating drums where sequential images were displayed for being viewed through some slots.

2.2 The industry of motion pictures

In 1870, Henry R. Heyl invented the *Phantasmatrope*, a magic lantern device that for the first time projected a motion-picture onto a screen for an audience. At the same time that the technology for screening was developed, advancements in photography were done. In 1872, Eadweard Muybridge recorded a sequence of galloping horses images that were projected using a device called the *Zoopraxiscope*. In 1890, inspired by Muybridge's work, Thomas Edison along with William Kennedy Laurie Dickson created the *Kinetograph*, a device that recorded motion pictures in 35mm film strips, and some years later they invented the *Kinetoscope*, a viewing apparatus that made use of the electric bulb for illumination, and the *Phonoscope* which incorporated a phonograph.

At the same time in France, the Lumière brothers invented the *Cinematographe*, the first mass-produced motion picture film camera that at the same time served as a projector. They also made their own films and they gave their first screenings in 1895, causing a sensation between audiences. In the following year, several movie theatres were opened in different cities of Europe. Since 1920 there

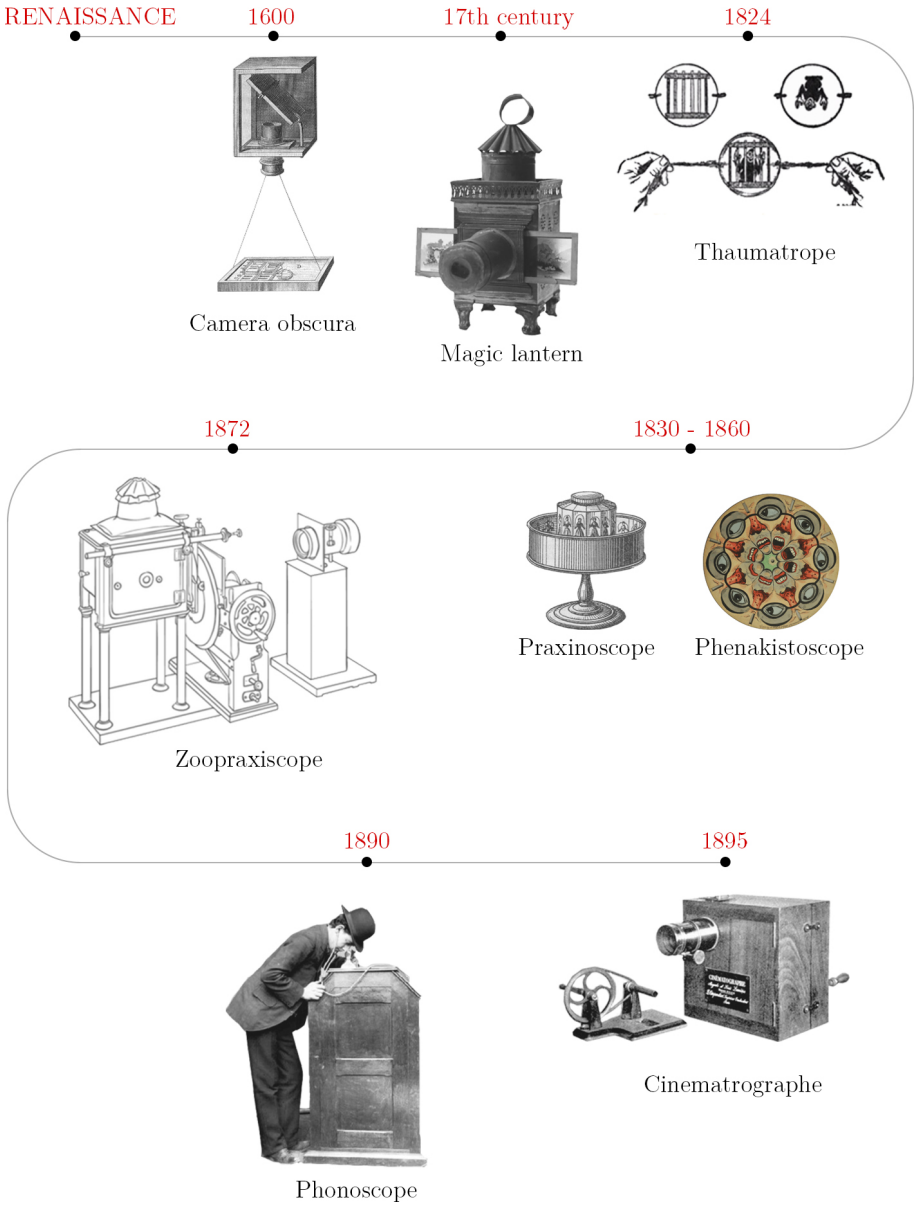

Figure 2.2. Timeline with the main discoveries in the cinema history.

were some steady advances in the film technology, and by the end of the decade, sound movies became the standard in cinemas. By 1920, films were a huge industry, especially after the emergence of Hollywood. From the 1910s until the 1960s, the world market would be dominated by American movies in the period known as "the golden age" of Hollywood.

2.3 Technological advances

The technology of film has remained essentially unchanged since its invention. Basically, film is a long strip of transparent plastic, coated on one side with a gelatin emulsion containing microscopically small light-sensitive silver halide crystals. This ribbon is mechanically transported through a camera, where a shuttering mechanism exposes it to light, activating the halide crystals depending on the amount of light received. A series of separate, sequential still images are recorded on the film strip. Afterwards, the film is passed through a chemical bath to develop the images recorded on it. The film is finished after being edited or cut into the final parts. This final version is duplicated in a laboratory and copies of it are distributed to theatres [47].

The first motion pictures were shot using a single photographic emulsion that produced black-and-white images. **Color film** was invented at the beginning of the 20th century, but it was not widely used for commercial motion-picture production until the early 1950s. Initially, the techniques to create the effect of color in film were tinting, that consisted in dying the emulsion of the film base giving the image a uniform monochromatic color, or the toning process, that replaced the silver particles in the film with metallic salts creating a color effect in the dark areas of the image.

The development of color film came after some advances in the study of colors and light. In 1861, the three-color method was firstly suggested by the physicist James Clerk Maxwell, and it was the basis

for the additive color system. In color theory, there exist two methods to produce color, the additive system, and the subtractive system. Additive colors are obtained by mixing different amounts of light colors. On the contrary, subtractive colors are created by absorbing some light wavelengths and reflecting others.

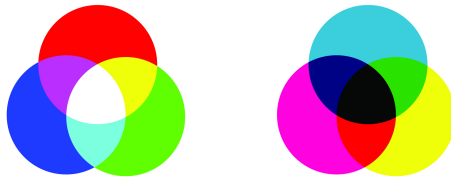

Figure 2.3. Additive color system on the left, subtractive system on the right.

Around 1900, the first systems for achieving color images appeared, and they were based on the additive color system. Red, green, and blue filters (sometimes two opposite colors instead of three: red and blue or red and green) were used to capture each color component information separately, into normal black-and-white film. These color components were projected using similar filters to reconstitute the final color image. An early example of this technology is the *Kinemacolor*, its mechanism is shown in Fig.2.4. These systems required an intense use of light and they were replaced by the subtractive color system some years later.

The subtractive color process uses either separate negatives to capture each color component by using filters (as it is done in the additive system) or a single film coated with three layers of color-sensitive emulsion, where each layer reacts to each different color stimuli. The final color image is created by superimposing three positive images, obtained by using the corresponding layer information from the negative, with an opposite color dye. For many years the three-layer film process, owned by Technicolor, monopolized the color

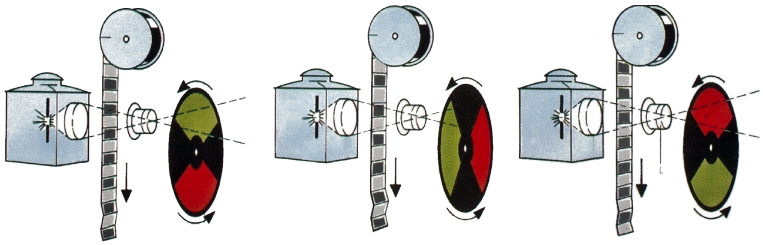

Figure 2.4. A schematic series showing how the two-color Kinemacolor additive motion picture process operated. Figure from [13].

cinematography, but after 1950 some alternatives appeared, such as Eastman Kodak *Kodachrome*, *Eastman color*, etc.

Since the 1920s, as the quantity of film and filmmakers was growing, film stock manufacturers began to diversify their products, for black and white and color cinema. The availability and diversification of film stocks offered some freedom to filmmakers to make a selection based on the desired aesthetic. Advancements in film emulsion and grain structure provided a wide range of available film stocks.

Another essential variable in the film stock is its **speed or sensitivity** to light [90]. A film with a lower speed index requires more exposure to light to produce the same image density as a faster speed index film. At the same time, the use of higher sensitivities generally leads to coarser film grain. Texture in analog images depends on the film stock chosen by the photographer. The film speed varies from ISO 50, which is slow and least sensitive to light, to 800, which is very fast and extremely sensitive to light.

The **post-production** techniques used in the laboratory to process the film stock can also offer a considerable variance in the resulting images [90]. Push-processing is a film developing technique consisting in increasing the recommended developing time of the film, compensating for under-exposure in the camera. The image produced shows an increased amount of grain, higher contrast, reduced quality, and saturated and distorted colors. On the contrary, pull-processing, consisting in overexposure and underdevelopment, results in images

Figure 2.5. Film grain used for artistic effect. Figure from `https://en.wikipedia.org/wiki/Film_grain`.

that display the opposite change in visual properties. The bleach by-pass is a chemical effect achieved by skipping the bleaching function while processing a color film, so the silver is retained in the emulsion along with the color dyes, resulting in a black-and-white image over a color image. Cross processing consists of processing film in a chemical solution intended for a different type of film, so the images show unnatural colors and high contrast.

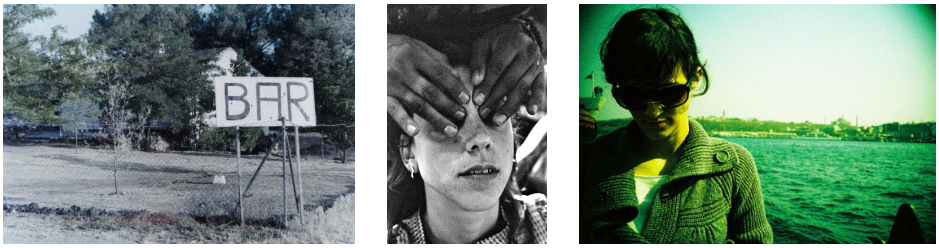

Figure 2.6. Bleach by-pass, push processing, and cross processing techniques in film photography. Figures from `https://handwiki.org/wiki/Push_processing`, and `https://crossprocessing.info/`.

Film gauge is the physical characteristic of film stock that defines its width [90]. The major movie film gauges are: 8 mm for amateur cinema, 16 mm for semi-professional, 35 mm for professional, and 65 mm for shooting epic photography, rarely used except in special event venues. A larger film gauge corresponds to a higher image resolution and technical quality.

All these factors plus the wide variety of cameras at the time contributed to the aesthetic of the resulting images in film cinematography. It allowed filmmakers to experiment, find and test new possibilities for creative expression. There exist numerous additional aspects that contribute to the art of film cinematography.

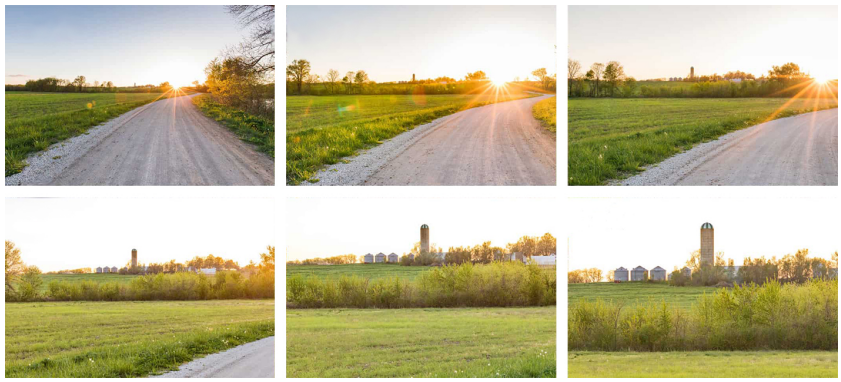

Figure 2.7. Increasing the focal length, the field of view decreases. From left to right, top to bottom: 16mm, 35mm, 50mm, 85mm, 135mm, and 200mm. Adapted figure from `https://www.colesclassroom.com/focal-length-basics-every-photographer/`.

Lenses can be attached to the camera to give a certain look or effect to the images, different lenses are used for different purposes. Focal length can be varied by the use of lenses. The **focal length** determines the angle of view and the field of view. Cinematographers can use wide-angle lenses (shorter focal length), that appear to expand the distance between objects and they produce perspective

distortions, normal lenses, that produce a field of view similar to the one we perceived with our naked eye, or long lenses or telephotos, resulting in compressed distances between objects and a magnification of the subject. Lenses can be divided into two categories, prime lenses or zoom lenses. Prime lenses or lenses with fixed focal length produce superior quality images, on the other hand, zoom lenses allow to change the focal length within a shot or quickly between setups for shots.

Figure 2.8. Decreasing the aperture of the lens, the depth of field increases. From left to right, top to bottom: f/3.5, f/5.6, f/10, and f/22. Shooting with wide aperture values, in the range of f/1.8 to f/3.5, will result in an image with a narrow depth of field. Conversely, narrow aperture values, in the range of f/18 to f/22 will result in a wider depth of field. Figure from `http://www.boostyourphotography.com/2014/10/depth-of-field.html`.

Focal length and diaphragm aperture affect the **depth of field** of a scene, that is the distance between the nearest and the farthest objects that are in sharp focus in an image. In cinematography, the use of tighter apertures to create every detail of the foreground and background in sharp focus is known as deep focus, on the contrary,

the use of a small depth of field is known as shallow focus. Depth of field is also affected by the format size, for instance, a 70mm film has less depth of field than 35mm for a given field of view.

2.4 The transition to digital cinema

It is hard to say precisely when digital cinematography started. In terms of digital projection, during the 20th century there were some early attempts at electronic projection, consisting in using video as a substitute for film in theatrical movie production and exhibition [47]. However, they were unsuccessful because the television image resolution was not enough for being projected at the big screen. Moreover, there existed some differences between the broadcasting and the motion-picture formats that made this projection impractical: the standard frame rate (frames per second) for motion-picture is 24 fps in a progressive manner, while for broadcasting video, some usual frame rates are 30, 60, 25, or 50 fps, employing interlace scanning. Although large-scale electronic image projection improved over the time, by 1960s was still not practical in economical terms, as cathode ray tubes (CRTs) projectors were used [47] (these projectors use a small, high-brightness cathode ray tube as the image generating element, and the image is then focused and enlarged onto a screen using a lens kept in front of the CRT face). It was not until the 1990s, when the D-ILA (Digital Image Light Amplifier) technology was designed and commercialized, that digital projectors started being used in theatres. In 1999, for the first time, two digital projectors of this type were used to project *Star Wars Episode I: The Phantom Menace*. Although being projected digitally, the movie was shot on film. The convenience of using digitally captured content was huge, as the movie already contained a large amount of computer-generated imagery, and the digital technology would imply saving money and speeding up the production work-flow.

The transition to digitally shot films was also progressive. In 1987,

the film *Julia and Julia* was shot using a Sony analog High Definition Video System (HDVS) camera, and then transferred to 35mm film for exhibition. However, this system met with little success due to image artifacts originated from the differences between broadcasting video and film formats. In the 1990s, the first digitally shot and post-produced films were released, but afterwards, they were converted to film for exhibition. George Lucas was one of the predecessors of fully digital cinematography, he included footage filmed with high-definition digital cameras into traditionally film shot movies. As it was mentioned above, he also used the digital projection technology for theatrical exhibition for the first time. Nowadays, the digital technology dominates the market: the major camera manufacturers offer a wide variety of Ultra High Definition video cameras specifically designed for digital cinema. In the last years, digital screening and digital shooting format has overtaken the film based technologies, some examples of this tendency can be observed in Fig. 2.9.

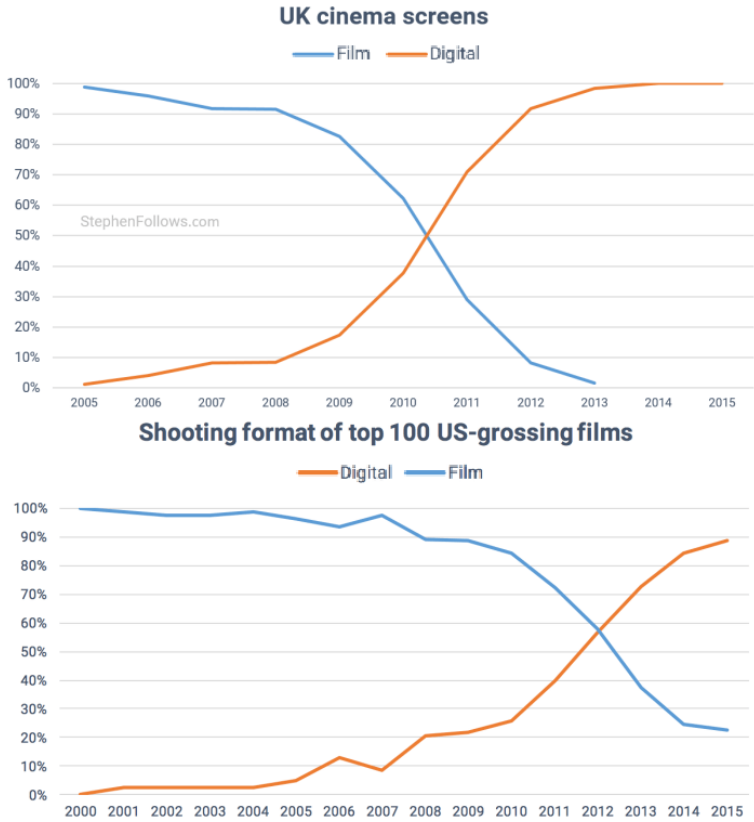

Figure 2.9. The graphic above shows the tendency in movie shooting among the top 100 grossing films in US, the graphic below how movies are displayed in the last years in UK. Figure from `https://stephenfollows.com/film-business-became-digital/`.

3

Digital cinema

The term digital refers to the way computers process information, that is as a series of zeros and ones, in contrast with the term analog, which refers to film and non-digital video (such as VHS tape) that are based on continuously variable signals [47].

Although there is no single definition of digital cinema, it generally involves four major categories which are digital production, post-production, distribution, and exhibition of theatrical movies.

Professional film-making consists of three main phases: **pre-production**, which encompasses all aspects that are prepared before the camera starts shooting, such as screenwriting, casting, costume design, location scouting, etc; **production**, the period of time when the film is shot, which includes direction, camera operation, acting, sound recording, lighting, etc; and **post-production**, the phase that comes after production and before releasing the film in its final form and it includes editing, color-grading, sound edition, mastering, etc. In the following sections, the aspects of cinematography that are concerned with images will be explained: image recording and image post-processing.

3.1 Production

The term production refers to the phase when the raw footage is recorded. The final resulting images are affected by decisions made at this stage of the workflow: the type of camera used, the selected recording format, etc.

3.1.1 Digital cameras

One of the first decisions made in film-making is the camera used for shooting. Since the 2010s, digital movie cameras have replaced film cameras in the motion picture industry. Digital cinema cameras capture footage in digital format rather than shooting on film stock as the traditional movie cameras do. They can be characterized by their resolution, the most common ones used by cinema camera manufacturers being: Standard Definition (SD), High Definition (HD), Full HD, Ultra High Definition (UHD), 4k, 6k, until 8k at the current moment. Digital cinema video formats are specified in terms of horizontal resolution, as multiples of 1024 pixels. For instance, a 2K image is 2048 pixels wide, a 4K image is 4096 pixels wide, and its corresponding vertical resolutions depend on their aspect ratios (relationship of the width of the picture to its height). The other resolution formats usually refer to 1280×720 pixels for HD, 1920×1080 pixels for Full HD, and 3840×2160 for UHD.

Although digital camera manufacturers do not make available the exact processes of their pipelines, there are three main phases common in the image formation of a digital camera: 1) the image acquisition, which explains how the light passes through the optics of the camera and reaches the sensor, 2) the transformation of light into electrical signals at sensor level, and 3) the in-camera color processing pipeline, which are the steps that occur inside the camera to obtain the final image [27].

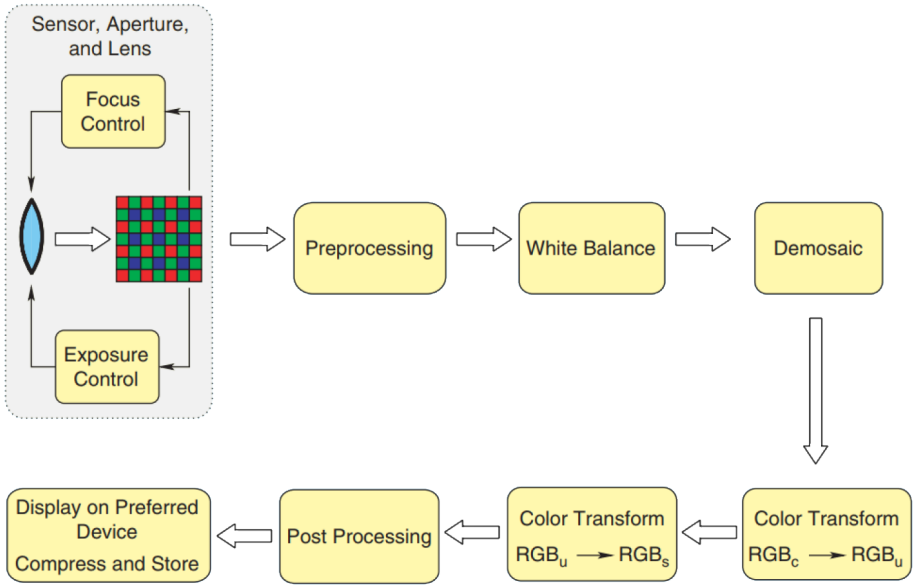

Figure 3.1. Digital image processing pipeline. Adapted figure from [64].

Exposure control

In photography, exposure is the amount of light per unit area reaching a frame of photographic film or the surface of an electronic image sensor. In the case of digital cameras, the exposure is controlled through two settings: the aperture and the shutter speed. Moreover, there is a third camera setting that affects the sensor output lightness values: the ISO number.

The aperture stop is an opaque part of an optical system that blocks certain rays. In a camera, a device called diaphragm serves as the aperture stop. The **lens aperture** is usually specified as an f-number, that is the ratio of the focal length f to the diameter D of the clear aperture of the diaphragm:

$$N = \frac{f}{D}$$

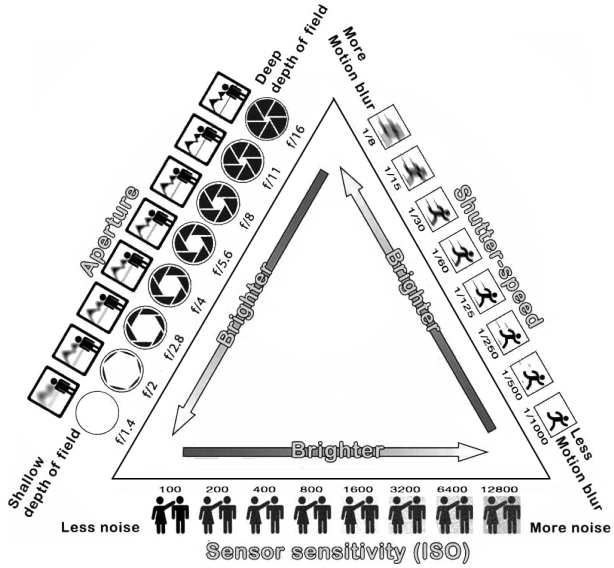

Figure 3.2. Adapted figure from `www.photoblog.com/learn/exposure-triangle-guide/`.

where N is the f-number. The **shutter speed** (or exposure time) is the length of time when the sensor (or the photographic film) is exposed to light, that is when the camera shutter is open when taking a photograph. The shutter speed is measured in fractions of a second. In the case of digital cameras, the **ISO value** controls the relationship between the exposure and the output image lightness. The lower the number, the less sensitive the camera is to light. In other words, with a smaller ISO number the output lightness value is lower compared to a higher ISO value.

Sensor

Once the light passes through the optics of the camera, it reaches the sensor. An image sensor is a device that converts light (photons) into electrical signals (electrons)[3]. It is formed by an array of cells. The electrical charge is accumulated in each cell while the sensor is

being exposed to light, and then, it must be converted into voltage through the scanning of the image array.

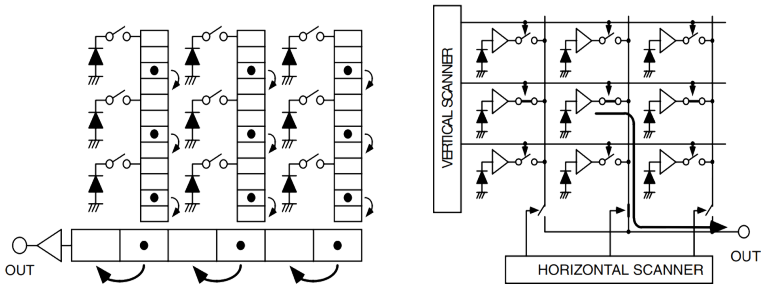

Figure 3.3. Two camera sensor types: on the left a CCD, and on the right a CMOS sensor. Figure from [53].

There exist two types of image sensors depending on how this scanning is performed:

1. The **CCD sensor**: It transfers the signal vertically from each column of cells, then it is transferred horizontally and converted into voltage at one single output amplifier.

2. The **CMOS sensor**: It has an amplifier at each cell location and it performs the conversion into voltage at each amplifier at the same time.

Once the voltage has been measured for each cell location, it has to be converted into digital values. The bit depth specifies how much information is kept for each pixel in the image, that is the range of integer values used for keeping the pixel information. For instance, an 8-bit image is represented by values in the range 0 to 2^8 (0 to 256), a 12-bit image is represented by values in the range 0 to 2^{12} (0 to 4095), etc. Therefore, a larger bit depth implies a wider palette of colors to represent the image.

At this point, the sensor has transformed the photons into digital values as a measure of the incident light intensity, but there is not

color information yet [3]. Color is a property of light relative to its wavelength and not to the light intensity. There exist two types of configurations to capture colors:

1. **Three-sensor systems**: The incoming light is separated into short, medium, and long wavelengths using a beam splitter, and three different sensors are used to capture each wavelength information. See Fig. 3.4.

2. **Color filter arrays** (CFA): The array of cells of the sensor is covered by a mosaic of individual color filters, making each cell in the array capture one color channel (red, blue, or green). The values of the other two channels are interpolated later on in a process called demosaicking. This system is the most popular since it needs only a single sensor. The most common CFA is the Bayer pattern, which is a 2×2 *'RGGB'* pattern that is repeated over the entire sensor, as it can be observed in Fig. 3.4.

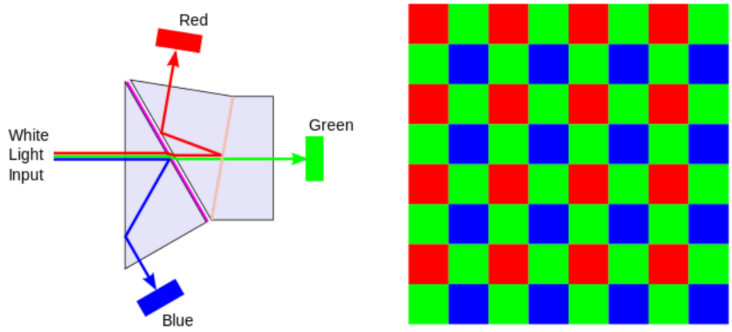

Figure 3.4. Capturing color information. A three-sensor system on the left, a Bayer color filter array on the right. Figure from [3].

Most of the professional cinema cameras, such as ARRI, RED, or Blackmagic cameras, use CMOS sensors with a single sensor system.

Camera processing pipeline

The values captured by the sensor can be directly stored before applying any transformation to them, these values form what is known as RAW image. In professional cinema productions, the RAW footage is edited at post-production stage: color grading, tone-mapping and gamut mapping are applied in the post-production laboratory. However, a chain of transformations can be applied to the RAW data inside the camera, in order to get an image ready for display. This is the usual practice in not-professional productions or in TV broadcasting. These steps and the order in which they are applied may vary from one camera manufacturer to another, but in general, they are listed as:

1. **White balance**: This step is based on the property of the human visual system of color constancy (or chromatic adaptation), that is the ability to perceive as constant the color of an object under different illumination conditions. Our perception of the color of the object is independent of the illuminant and matches the reflectance values of the object.

   The triplet RGB value captured by the camera sensor is:

$$
\begin{aligned}
R &= \int_{\omega} r(\lambda) I(\lambda) S(\lambda)\, d\lambda \\
G &= \int_{\omega} g(\lambda) I(\lambda) S(\lambda)\, d\lambda \\
B &= \int_{\omega} b(\lambda) I(\lambda) S(\lambda)\, d\lambda
\end{aligned}
\tag{3.1}
$$

   where ω is the spectral range over which the camera is sensitive, $r(\lambda), g(\lambda)$ and $b(\lambda)$ are the spectral sensitivities of the red, green and blue filters used by the camera (see Fig. 3.5), $I(\lambda)$ is the power distribution of the illuminant, and $S(\lambda)$ is the spectral reflectance of the object.

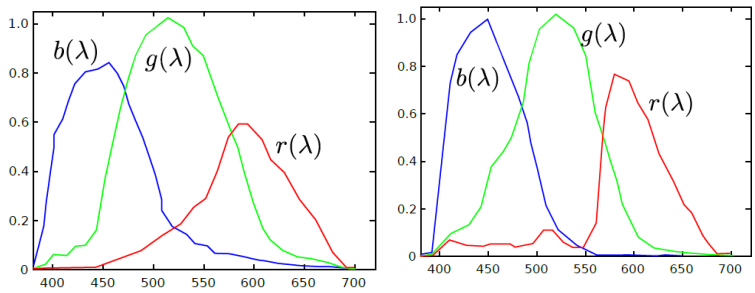

Figure 3.5. On the left: the spectral sensitivity of Nikon D5000 camera sensor, and on the right: the spectral sensitivity of Canon EOS 500D camera sensor. Image from [27].

A simplified model of these equations are:

$$\begin{aligned} R &= I(\lambda_R)S(\lambda_R) \\ G &= I(\lambda_G)S(\lambda_G) \\ B &= I(\lambda_B)S(\lambda_B) \end{aligned} \tag{3.2}$$

where λ_R, λ_G, and λ_B are the corresponding peak sensitivities for each filter.

If the illuminant values are known, the reflectance values can be recovered by doing:

$$\begin{aligned} S(\lambda_R) &= R/I(\lambda_R) \\ S(\lambda_G) &= G/I(\lambda_G) \\ S(\lambda_B) &= B/I(\lambda_B) \end{aligned} \tag{3.3}$$

Therefore, the matrix form of the white balance transformation is:

$$\begin{bmatrix} R' \\ G' \\ B' \end{bmatrix} = \begin{bmatrix} \frac{1}{I(\lambda_R)} & 0 & 0 \\ 0 & \frac{1}{I(\lambda_B)} & 0 \\ 0 & 0 & \frac{1}{I(\lambda_B)} \end{bmatrix} \begin{bmatrix} R \\ G \\ B \end{bmatrix} \tag{3.4}$$

where R, G, B are the values detected by the sensor and R', G', B' are the resulting white-balanced values.

In a real scenario, the illuminant values are unknown, so in order to apply the previous formulas, these values have to be estimated. This estimation can be done manually or automatically. The illuminant estimation is an ill-posed problem and there exist numerous methods that aim to estimate it, a review on them can be found in [33].

2. **Demosaicking**: As it has been shown in Fig. 3.4, in the color filter array image each pixel contains information about one color channel: red, green, or blue. The other two channel values must be interpolated in a process known as demosaicking, so a 3-channel image is created from the original 1-channel CFA image. One of the simplest demosaicking methods is based on bilinear interpolation (see Fig. 3.6), which is an eight neighborhood filter that obtains the missing value of a certain color channel by taking the average of the adjacent pixels of the same color channel.

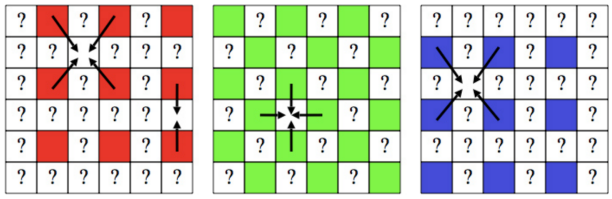

Figure 3.6. Demosaicking by bilinear interpolation. Image from `https://slazebni.cs.illinois.edu/spring19/assignment0.html`.

This approach produces modest results, that may show artifacts: false colors or zipper-effect. More complex models incorporate directional information into the reconstruction, so edges are estimated correctly and fine details are preserved. There exists an extensive literature on demosaicking techniques. A survey on several approaches for this task has been done in [48].

3. **Color transformation**: At this point, the obtained color image values are defined in the RGB color space of the camera sensor. During color correction, the image is firstly converted from the camera RGB color space to XYZ tristimulus values, and then from the CIE XYZ color space to a standard RGB color space (e.g. RGBs, ITU-R BT.709), for display purposes. This conversion is needed because usually, the spectral sensitivity functions of the camera sensor color channels are not identical to those of the displaying output color space.

   Typically, this transformation is a chain of two multiplications by 3×3 color conversion matrices.

$$\begin{bmatrix} R_s \\ G_s \\ B_s \end{bmatrix} = B \cdot A \cdot \begin{bmatrix} R_c \\ G_c \\ B_c \end{bmatrix} \tag{3.5}$$

   where R_s, G_s, B_s are the values in a standard RGB color space, A is the color transformation matrix from the camera RGB to CIE XYZ color space, and B is the transformation matrix from CIE XYZ to standard RGB color space, and R_c, G_c, B_c are the values of the sensor in the camera RGB color space.

4. **Encoding techniques**: Image encoding aims to compress the demosaicked, white-balanced image in the most efficient manner, so the usage of bits is optimized. Perceptual encoding is essential to maximize the perceived image quality.

The most common encoding techniques are based on brightness perception models (explained more in detail in Section 4.3).

The *Weber-Fechner's law* describes the relation between the sensation S, and the base stimulus I:

$$S = k' log\left(\frac{I}{I_T}\right) \tag{3.6}$$

where k' is a constant, and I_T is the threshold value that makes the perceived stimulus become zero. This law implies a logarithmic relationship between physical stimulus and perceived magnitude.

Steven's law, on the other hand, states that perceptual sensation and the physical stimulus are related through a power law:

$$S = kI^a \tag{3.7}$$

where S is the sensation, k is the proportionality constant, and a is an exponent that depends on the type of stimulus. In the case of *lightness sensation*, as explained in [48], a has a value of 0.42.

These two perception models imply that the brightness function has a compressive nature, meaning that for darker values its slope is higher, and as the luminance increases it gets progressively lower. The consequence is that a small change in luminance in a dark region will be more noticeable than the same change in a bright region [4].

Camera sensors transform light into numerical values that are proportional to the light intensity. The camera signal has to be quantized into a certain number of bits to provide a digital output. In practice, the signal is transformed by a non-linearity that follows a brightness function, and quantization is applied afterwards: this process is called *perceptual linearization*. In

this way, more bits are used at the darkest regions, where we are more sensitive to differences, while fewer bits are used at the brightest regions, where we are less sensitive to differences. See Fig. 3.7.

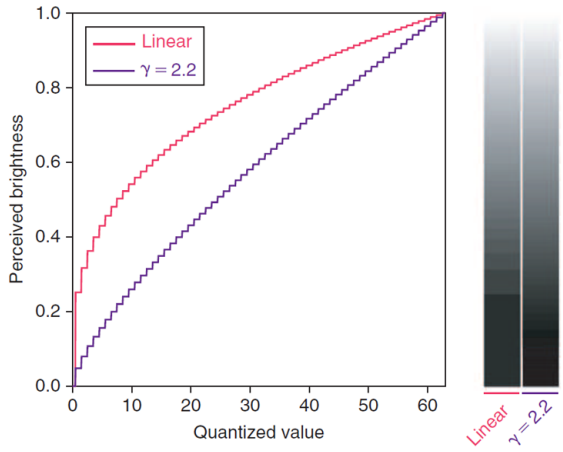

Figure 3.7. Comparison of perceived brightness of quantized steps using linear (red) and gamma encoding (blue). Image from [70].

The function that the camera applies before quantization is known as *opto-electro transfer function* (OETF), which takes into account the perception models explained before. In this subsection, the most commonly used OETFs will be explained: gamma correction, for SDR content; and logarithmic encoding, PQ, and HLG for HDR image sequences.

Gamma correction is a power-law transform, except for low luminances where it is linear, so as to avoid having an infinite slope at luminance zero (which can cause numerical problems).

Gamma correction was originally used due to the cathode ray tube (CRT) displays. The relation between the device input voltage and the luminance of the screen for these displays is

$$L = kV^{\gamma} \tag{3.8}$$

where L is the screen luminance, V is the voltage, and γ is the exponent of the power function, which has a value of around 2.4. At the same time, the luminance values captured by the camera are linearly proportional to light intensity. Therefore, for correct luminance reproduction on the display, the camera luminance signal V_s must be transformed by applying the inverse of the γ value:

$$V_c = V_s^{1/\gamma} \tag{3.9}$$

where V_c is the corrected voltage, V_s is the source voltage from the camera sensor, and $1/\gamma$ is the gamma correction exponent, that has a value of around 0.42, and it is known as encoding gamma. The γ is known as decoding gamma, and $1/\gamma$ is called encoding gamma. The process of compensating the CRT display response to luminance is known as *gamma correction*.

Although gamma correction is attributed to the CRT non-linear response, our perception of lightness (accordingly to Steven's law, see Eq. 4.7), follows a non-linearity with respect to the luminance in the scene that is very similar to the gamma correction exponential function: both functions are a power law of exponent approximately 0.42. While CRT displays are obsolete, gamma correction is still used in the camera output to emulate the perception of luminance in HVS.

In order to implement gamma correction, different transfer functions can be used. For instance, the BT.709, used for high definition television (HDTV) is

$$V' = \begin{cases} 4.5V & \text{if } 0 \leq V \leq 0.018 \\ 1.099V^{0.45} - 0.099 & \text{if } 0.018 \leq V \leq 1 \end{cases} \tag{3.10}$$

where V denotes pixel value in any of the color channels.

The standard used for screen monitors and internet is called

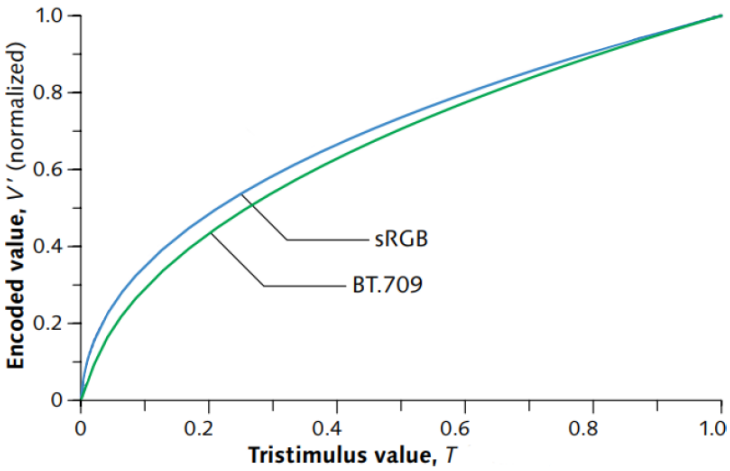

Figure 3.8. Graph of transfer functions BT.709 and sRGB. Adapted figure from [61].

sRGB,

$$V' = \begin{cases} 12.92V & \text{if } 0 \leq V \leq 0.0031308 \\ 1.055V^{1/2.4} - 0.055 & \text{if } 0.0031308 \leq V \leq 1 \end{cases} \tag{3.11}$$

where V denotes pixel value in any of the color channels. See Fig. 3.8.

Gamma correction has been the most used encoding technique for a long time, however, due to quantization issues, it is not suitable to work with high dynamic range (HDR) imaging. [4].

Logarithmic curves are commonly used for encoding high dynamic range content. The logarithmic color space has its origin in the first digital film systems, which consisted in scanning film negatives for subsequent computer-based post-processing. The data was stored in logarithmic format, directly corresponding to density of the original negative.

The reason to still use the logarithmic curve is that, as it has been seen in Eq. 4.6, it follows the Weber-Fechner's brightness

perception law and, at the same time, it retains the most dynamic range of information from the camera sensor, in the case of high luminance values. Each camera manufacturer may have its own definition for this logarithmic function, but its general form, common to the most popular log-encoding approaches, is expressed as:

$$V' = c \cdot log_{10}(a \cdot A \cdot V + b) + d \tag{3.12}$$

where V denotes pixel value in any of the color channels, and the parameters a,b,c, and d are constant real values varying for different camera manufacturers and camera settings. If the resulting images from this transformation are directly displayed, they appear under-saturated and with low contrast, so a look-up table (LUT) needs to be applied for preview in on-set monitors.

The **PQ (Perceptual Quantization)** is a family of curves. Each curve depends on the peak luminance of the display where the image is going to be presented. The PQ general formula can be expressed in the form of a Naka-Rushton equation [4]:

$$V(L) = K_1 \left(\frac{c_1 + c_2(\frac{L}{k_0})^{m_1}}{1 + c_3(\frac{L}{k_0})^{m_1}} \right)^{m_2} \tag{3.13}$$

where L is the luminance value, c_i is the minimum contrast that can be detected on an image of luminance L_i, $i \in \left[1, 2^N\right]$, and N is the number of bits for encoding the transform.

By construction, this linearization transform ensures that quantization errors are not visible, as they are at the JND threshold. Therefore, the PQ curve does not model brightness perception. The experiments show that PQ requires fewer bits than gamma correction to avoid banding artifacts.

The **HLG function** is a curve based on classic brightness perception models. It can encode a wide dynamic range, while still

making the signal compatible with regular SDR displays. The HLG function is expressed as [4]:

$$HLG(I) = \begin{cases} \sqrt{3I}, & 0 \leq I \leq 1/12 \\ a \cdot log(12I - b) + c, & 1/12 < I \leq 1 \end{cases} \tag{3.14}$$

where I is proportional to relative light intensity in a camera color channel (R, G, or B). See graphic in Fig. 3.9.

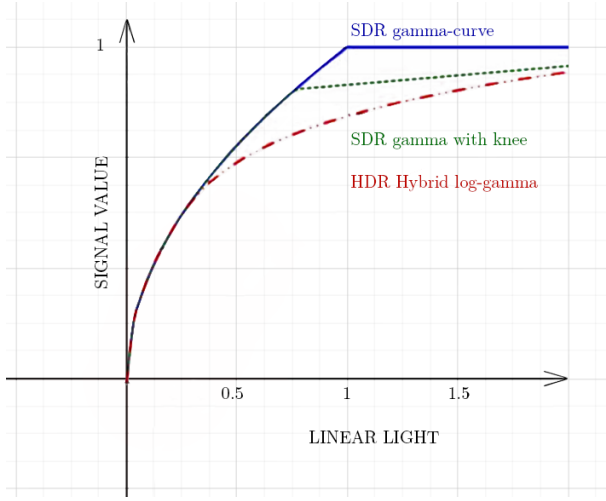

Figure 3.9. HLG and gamma correction. Figure from [4].

5. **Displaying:** The last step in the camera pipeline is the adaptation of the footage to the device where it will be displayed. Some basic concepts will be defined before introducing the most common final transformations in the camera pipeline: *dynamic range* is the ratio between the brightest and the darkest values that can be captured or reproduced, and *color gamut* is the range of colors achievable on a given display.

   Therefore, **tone mapping** is the process of reducing the dynamic range of the input for displaying purposes, while pre-

serving the perceived detail of the image. (A more detailed description of the tone mapping process can be found in Section 5.3). For accurate on-set monitoring of High Dynamic Range (HDR) footage, it is very common that 3D LUTs are created beforehand by cinematographers and colorists [4]. And **gamut mapping** consists of altering the range of colors of the original content to adapt it to the display color gamut. In TV broadcasting or low-budget movie productions, this process is carried out within the camera. For professional cinema productions, both tone mapping, and gamut mapping are performed off-line by expert technicians at post-production stage.

3.2 Post-production

This phase includes all the transformations made to footage after production and before distribution. During post-production the image content is altered by the colorist, so the footage is adapted to the device where is going to be displayed, and also its "visual look" is defined. These transformations are done in a process called *color grading*. However, during pre-production and production phases, before the actual color grading takes place, grades can be created by the director of photography (DoP), and they can be stored in a variety of ways:

- **Lookup tables (LUTs)** are saved image-processing operations to set looks for on-set display. Moreover, they can be passed to colorists during post-processing stage, for reference, or as a starting point for color grading. As it has been explained before, LUTs can be used on-set for displaying HDR content. See Fig. 3.10.

- **Color decision lists (CDLs)** are an industry-standard file format for the exchange of basic primary color grading information. A CDL defines the parameters slope, offset, and power,

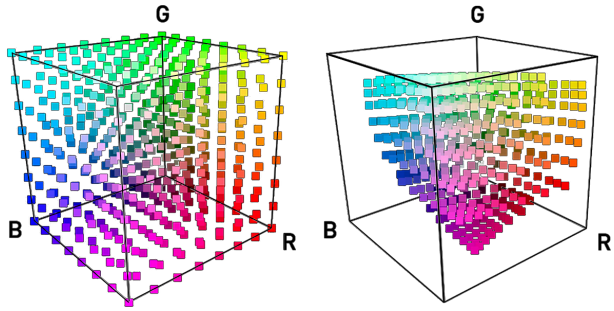

Figure 3.10. Visualization of a LUT color transformation. Adapted figure from `https://www.inventome.com/read/the-truth-about-luts`.

following the color correction formula

$$i' = (i \times s + o)^p \tag{3.15}$$

where i' is the color graded pixel code value, i is the input pixel code value, s is slope, o is offset, and p is power. These three parameters are defined for each R, G, and B channels. The CDL consists of these 9 parameters, plus an extra parameter, saturation, that is applied in R, G, and B channels in combination. A CDL is used to organize primary grade adjustments for a collection of shots, and it is also a starting point for adjusting colors in post-production.

Ideally, the intended color look is conceived beforehand, in pre-production stage, and it is stored as a LUT or a CDL. Then, during post-production and usually in a color suite, this look is implemented and carefully rendered by the colorist in a process called color grading. Color grading consists of adjusting the image in specific ways to improve its appearance, or to create stylistic effects. Colorists use color grading for artistic purposes to ensure that the footage conveys a specific atmosphere, mood, or look. It also includes the processes of tone-mapping and gamut mapping, that adapt the footage to the displaying

device. This whole process is very costly in terms of budget and time, and it may require a significant amount of work from very skilled artists and technicians.

3.3 Distribution

Once the footage has been recorded and post-processed, it needs to be compressed. Uncompressed video represents quite large quantities of data. Different properties of the image contribute to the size of this data:

- The **image quality** or video resolution indicates the number of pixels that form the image, and it is expressed as the number of horizontal pixels. The number of rows of pixels depends on the aspect ratio (relationship of the width of the image to its height). For instance, a 2K image with 1.78:1 aspect ratio has 2048×1152 pixels, or 23.6 Megapixels.

- The **bit depth** specifies much color information, or how many bits are available for each pixel, in an image. For example, an image with a bit depth of 8 has 2^8, or 256, possible values for each channel. See Fig. 3.11.

- The **bit rate** is the number of bits per second. It is a measure of video file size, and it is affected by the image resolution , the bit depth, the video frame rate (number of frames per second), and the method used for compressing the video.

The amount of data needed to represent images or video is reduced in a process called **compression**, so this image content can be stored in a disc or can be watched with good quality by streaming. The goal of compression is not only to meet a bitrate requirement, but it must also do it while at the same time keeping the image quality above a certain level and using methods of affordable computational complexity [3]. Compression is possible due to the redundancy that exists in any video content, and to some properties of "natural" images:

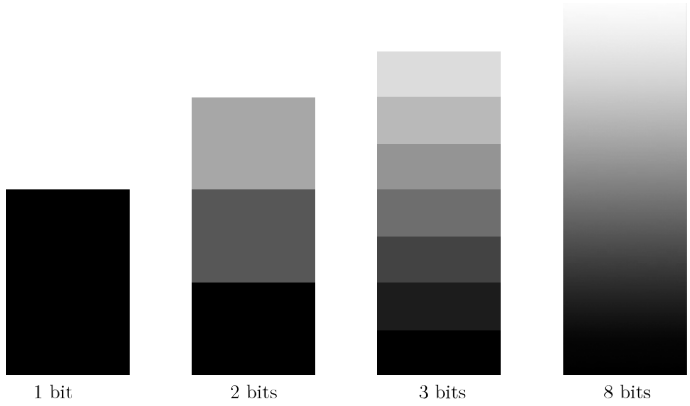

Figure 3.11. Comparison of different bit depths.

- Temporal redundancy: In a video, from one frame to the next, the percentage of pixels that change their value is small compared to the total number of pixels. Therefore, each frame can be very well approximated by a version of the previous frame.

- Spatial redundancy: In homogeneous regions, each pixel is similar to its neighbours, also at image contours, or within textured regions. Compression can be achieved by expressing pixels in terms of their neighbours.

Compression also exploits some perceptual properties of the HVS, so less data is used to represent regions that are perceived less. Visual masking is a phenomenon of perception that occurs when the visibility of one image, called a target, is reduced by the presence of another image, called a mask. Several types of masking are exploited for compression purposes:

- Luminance masking: According to Weber's law, Eq. 4.4, an increment in a stimulus is perceived only if it is larger than a certain threshold value. This property is used when quantizing a signal, assigning more bits to darker values, and less for brighter regions.

- Texture masking: Artifacts are more visible in uniform regions than in textured ones. Therefore, quantization is adapted to the intensity variation of image regions, using fewer bits in more variable regions.

- Temporal masking: The HVS takes some time to adapt to abrupt scene changes. During this period the sensitivity to details is lower, so images can be represented in a coarser way.

- Color masking: The sensitivity to luminance is higher than to chrominance. This property can be used for more efficient quantization.

3.4 Analog vs. digital cinema

As it has been exposed in [43], until the end of the 20th century, most film professionals and critics preferred celluloid film. Digitally recorded images were considered technically and aesthetically inferior. Until recently, production companies preferred digital technology for budgetary reasons, while directors of photography and filmmakers often chose analog film. There have been some technical reasons for this preference: until the launch of the ARRI Alexa in 2011, the dynamic range capacity of digital cameras was inferior compared to analog film stock. However, modern digital cameras are comparable or sometimes superior to film stock in terms of dynamic range. Therefore, the difference between analog and digital film is not only a question of dynamic range but rather lies in the characteristics of film grain and pixel: the random spatial distribution of film grain is described as pleasing to the eye, compared to the stability of the pixels fixed grid, which is perceived as cold and sterile. Some other differences between these two formats are color reproduction, or the mechanical movement that is present in analog cameras. Despite these differences, in the last years, digital has become the standard for shooting and projecting, mainly due to lower

costs. In their work [43], the authors state that an empiric study comparing the two technologies has not been conducted. Therefore, they compare some narrative films recorded with analog and digital cinematography, in terms of cognitive and emotional reactions, enjoyment, and immersive experiences. They also test whether the type of projection influences audience reactions. Their main observations are the following:

- The two capturing technologies produced similar emotional and immersive experiences during digital projection.
- There exist significant differences in the memory of visual details, with higher recall scores for the digitally captured versions.
- The mechanical projection of celluloid film produced higher levels of emotional reactions.

The conclusion of the study is that the gap between analog and digital aesthetics has been closed with the current advances in digital technology. On the other hand, the transition to this new technology creates numerous challenges for cinematographers. Digital cinematography and the aid of computers is a great advantage in many aspects, however, cinematographers are becoming increasingly frustrated by some artistic limitations that the digital medium imposes, and that current movie production trends promote. Since the beginning of cinema and for many decades, there was a wide variety of cameras, film stocks and film developing options that allowed cinematographers to experiment, find and test new possibilities for creative expression, often carefully thought out in advance, while the limitations of film in terms of dynamic range required a mastering of the craft of lighting the scenes which also fostered artistic creativity; cinematographers performed the bulk of their work at pre-production and during the film shoot, with relevant but usually minor adjustments during post-production. Currently, virtually all professional productions resort to the same digital cinema camera model, causing the default look to

be quite homogeneous to begin with. And these cameras have ever increasing dynamic range capabilities, so there is less and less need to light the scenes. Consequently, producers are pressing directors of photography to complete more and more shots per day, just ensuring that the image quality is good in the barest possible sense (detail visibility, focus, and so on), but as much as possible leaving artistic decisions regarding contrast and color for the color-grading stage in post-production.

As a result, cinematographers have increasingly less opportunities to properly exercise their craft: on the set there is pressure not to devote too much time for lighting and just make sure everything is properly visible, while in post-production the cinematographer must communicate the artistic intent to the colorist, who must be able to translate it into operations performed on the color grading suite (and the time devoted for this is also being progressively reduced). The net result is that more and more movies tend to have a similar look, with directors of photography growing dissatisfied with the diminishing role their craft seems to be taking.

4 The human visual system: Perception and visual models

This chapter starts with the definition of visible light and a summarized description of the human visual system. Then, some visual processes in the retina are developed more in depth, and it concludes with the explanation of some perception models of brightness and color.

4.1 Light

Electromagnetic radiation is characterized by its wavelength (represented by λ) and its intensity. Light, or *visible light*, is defined as the electromagnetic radiation with wavelengths within the spectrum that can be perceived by the human eye, that is between 380 nm and 740 nm.

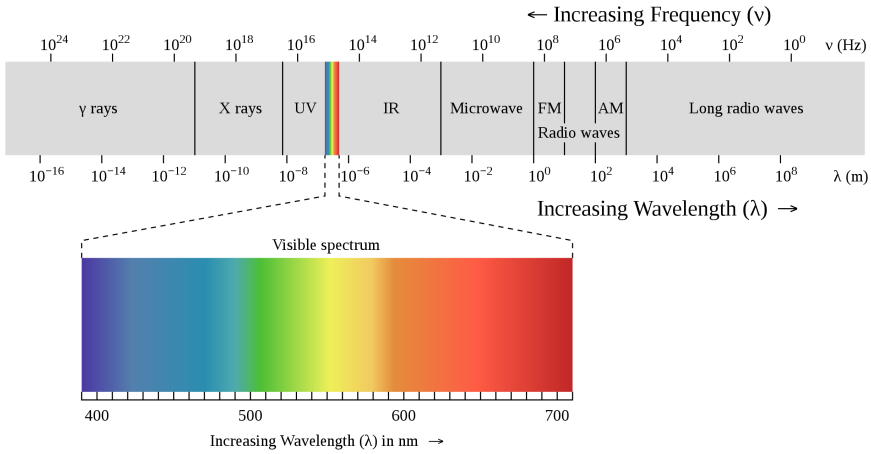

Figure 4.1. Electromagnetic spectrum and visible light. Figure from [3].

Some definitions related to light will be given before going any further:

- The *irradiance* function, $I(\lambda)$, describes light by its power spectrum, that is, for each wavelength λ, it defines the amount of power I the light has. See Fig. 4.2.

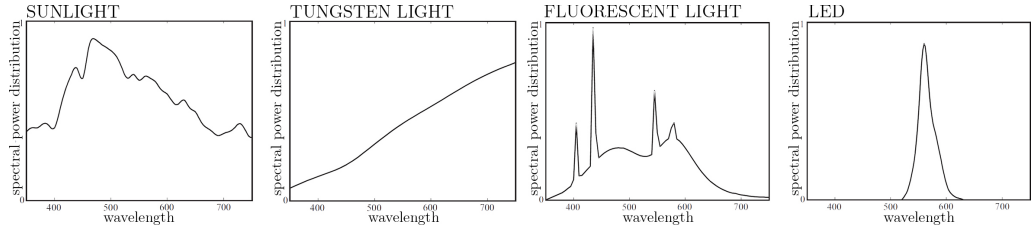

Figure 4.2. Irradiance function $I(\lambda)$ of various common types of illuminations. Adapted figure from [45].

- The *reflectance* function, $R(\lambda)$, describes the light absorption properties of a surface, that is, for each wavelength λ, it states

the percentage of photons that are reflected by the surface. See Fig. 4.3.

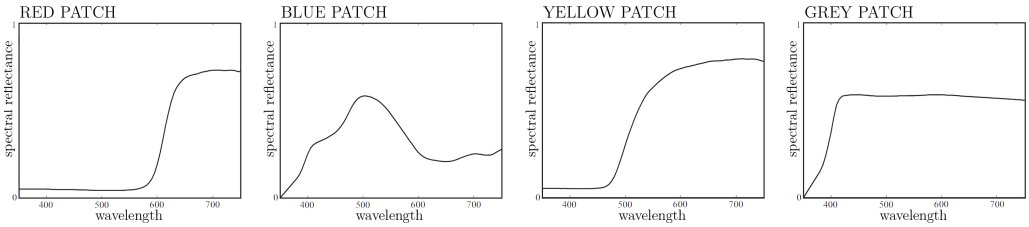

Figure 4.3. Spectral reflectance of various colored patches. Adapted figure from [45].

- The *radiance* function, $E(\lambda)$, defines the light reflected by a surface that reaches our eyes. It is defined as:

$$E(\lambda) = I(\lambda)R(\lambda) \tag{4.1}$$

- The *luminosity function* describes the spectral sensitivity of the human eye. As Fig. 4.4 shows, $V(\lambda)$ has a higher response on wavelengths in the middle of visible light, and it decreases in the extremes of the visible spectrum.

- *Luminance* is a measure of the luminous intensity of light per unit area. It is calculated as the integral of a weighted radiance over the visible spectrum:

$$\int_{380}^{740} V(\lambda)E(\lambda)d\lambda \tag{4.2}$$

  where $V(\lambda)$ is a luminosity function.

- *Brightness* is a subjective measure. It is the visual sensation according to which an area appears to exhibit more or less light [61].

- *Lightness* is defined as the brightness of an area judged relative to the brightness of a similarly illuminated area that appears to be white or highly transmitting [61].

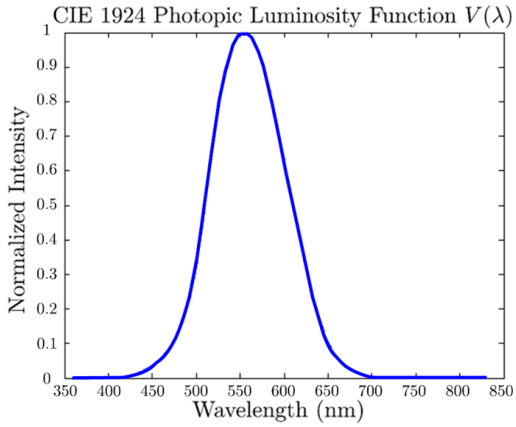

Figure 4.4. The standard CIE photopic luminosity function $V(\lambda)$. Image from [27].

4.2 Biological vision

Visual perception is the ability to perceive the surrounding environment using light within the visible spectrum. The *visual system* refers to all the physiological components involved in vision. It comprises the eye and parts of the central nervous system: the retina and the photoreceptor cells, the optic nerve, the optic tract, and the visual cortex.

4.2.1 Human visual system

The optical system of the human eye acts similarly to the optics of a camera. Light enters the eye through the outermost coat of the eye: the cornea, or the clear, curved layer located in front of the iris and pupil. The front of the eye is protected by the cornea. After passing through it, light travels through the pupil. The pupil forms a variable diaphragm, that regulates the amount of light going through. The iris controls and determines the maximum aperture of the pupil size. Behind the iris sits the lens. By changing its shape, the lens

focuses on near or far objects. The structure of the eye and its main components are shown in Fig. 4.5.

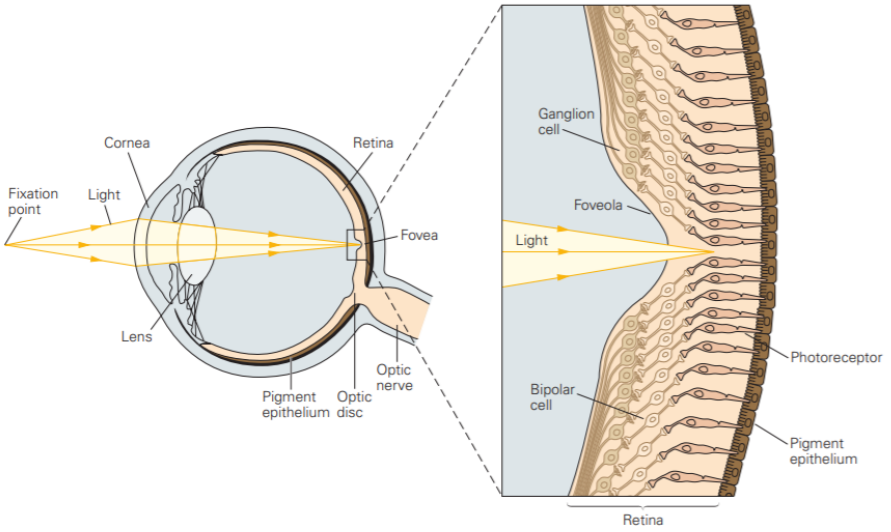

Figure 4.5. Cross-sectional diagram of the human eye. Image from [35].

The retina has photoreceptor neurons that convert light into electrical signals, which are carried to the brain by the optic nerve. These signals reach the brain through the lateral geniculate nucleus (LGN) in the thalamus, which transmits the information to the visual cortex, more specifically to the primary visual cortex (V1), where the color representation is formed. Some cells in the LGN respond to color, and for some time the chromatic properties of these neurons were thought to correspond to the perception of color, but more recently has been shown that this is not the case [4]. Recent studies demonstrate that some color perception phenomena, such as color contrast and color constancy, can be explained by the presence of double-opponent cells, which are found in the primary visual cortex (V1).

4.2.2 The retina

There are two main types of photoreceptors in the retina, depending on their sensitivity to light luminance:

- *Rods* work with low and mid-low luminances. At high luminances are active but saturated.

- *Cones* work with high luminances, at low luminances are not active because their pigments are less sensitive compared to rods pigments. There exist three types of cones: S-cones for short wavelengths, M-cones for medium wavelengths, and L-cones for long wavelengths. The spectral absorbance functions: $l(\lambda), m(\lambda)$, and $s(\lambda)$ describe the sensitiviy to light of each sort of cone photoreceptor, as a function of wavelenght. See Fig. 4.6.

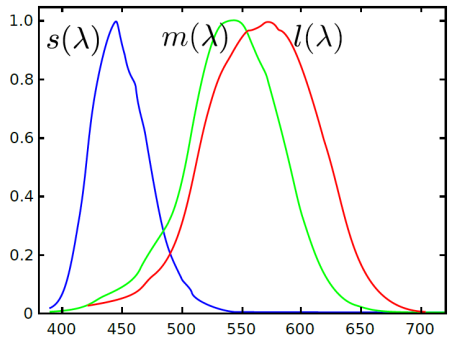

Figure 4.6. Spectral sensitivities (normalized) of S, M, and L-cones as functions of wavelengths. Adapted figure from `https://en.wikipedia.org/wiki/Spectral_sensitivity`.

The sensation produced by a light of power spectrum $E(\lambda)$ is determined by a triplet of values, the *tristimulus values*, given by the following formula:

$$
\begin{aligned}
L &= \int_{380}^{740} l(\lambda)E(\lambda)d\lambda \\
M &= \int_{380}^{740} m(\lambda)E(\lambda)d\lambda \\
S &= \int_{380}^{740} s(\lambda)E(\lambda)d\lambda
\end{aligned}
\tag{4.3}
$$

where $l(\lambda), m(\lambda)$, and $s(\lambda)$ are the spectral absorbance functions of each photoreceptor type.

Low luminance vision, or *scotopic vision*, is mediated only by rods and it is therefore color-less. In a low-medium range of luminances, or *mesopic vision*, both rods and cones are active, and in high-luminance, or *photopic vision*, cones are active and rods are saturated [3].

4.2.3 Low-level visual processing in the retina

The retina is composed of five cell types that are arranged in three cellular layers separated by two in-between layers, called plexiform layers (See Fig.4.7).

The photoreceptor cells, rod and cones, in the outermost layer, absorb light and convert it into electrical signals. These signals are passed to bipolar cells (BCs), which in turn connect to retinal ganglion cells (RGCs) in the innermost layer. In addition to this vertical pathway, the retinal circuit includes many lateral connections provided by horizontal cells (HCs) in the outer plexiform layer and amacrine cells (ACs) in the inner synaptic layer. Retinal ganglion cells are the output neurons of the retina and their axons form the optic nerve that transmits the visual signal from the retina to the brain.

Photoreceptors transform the light reaching the retina into electrical signals. The response of photoreceptors is nonlinear and, for a single cell without feedback, it can be well approximated by the Naka-Rushton equation [75], which is a particular instance of a *di-*

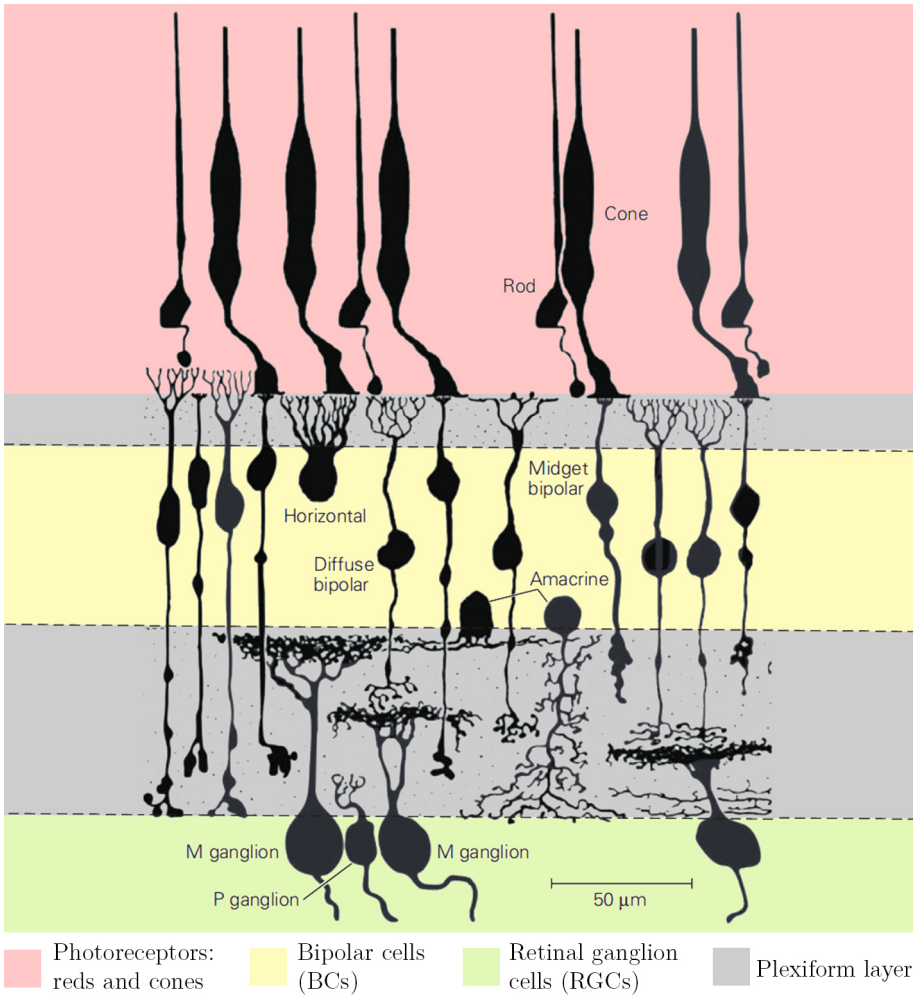

Figure 4.7. Neurons in the retina of the macaque monkey. Adapted figure from [35].

visive normalization operation [11], i.e. a process that computes the ratio between the response of an individual neuron and some weighted average of the activity of its neighbors, and this in turns allows the photoreceptor response to adapt to the average light level, therefore, optimizing its operative range.

The *lateral inhibition* or center-surround processing, in which a cell response is modeled as the difference between the activity of the cell closest neighbors and the activity of the cells in the near ring-shaped surround, allows to encode and enhance contrast, therefore being key for efficient representation, and is present at every stage of visual processing from the retina to the cortex. Lateral inhibition is often modeled as a linear operation, a convolution with a kernel shaped as a difference of Gaussians (DoG). See diagram in Fig. 4.8.

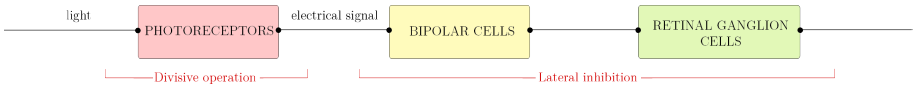

Figure 4.8. Diagram of the divisive normalization operation, and lateral inhibition in the retina.

4.3 Brightness perception models

This section adapts and summarizes some concepts explained in "Vision models for high-dynamic-range and wide colour gamut imaging" [4], Chapter 5. As it has been defined at the beginning of this chapter, the term *brightness* refers to the visual sensation according to which an area appears to exhibit more or less light, therefore, it is a subjective measure. *Brightness perception* is the relation between the intensity of light , which is a physical magnitude, and how bright it appears to us, which is a subjective magnitude. This relationship is not linear, and different models have been proposed so far. However, it is not a solved problem since the models depend on the type of

experiments conducted, the viewing conditions, image background, etc.

The **Weber-Fechner law** is a hypothesis in the field of psychophysics (or quantitative study of perception), that explains the relation between physical stimulus and perceived magnitude. Weber's law states that the perceived change in stimuli is proportional to the initial stimuli:

$$\frac{\Delta I}{I} = k \tag{4.4}$$

where I is the base stimulus, and k is the constant of proportionality. Assuming that all the ΔI, or *just noticeable difference* (JND) are equal, Fechner derived that each JND produces the same increase ΔS in the amount of sensation S, then:

$$\frac{\Delta I}{kI} = \Delta S \tag{4.5}$$

Fechner also assumed that the total magnitude S of the sensation can be computed by adding up the contributions of all the ΔS increments. Therefore, by integrating the previous equation, and assuming that the perceived stimulus becomes zero at some threshold stimulus I_T, the *Weber-Fechner's law* is obtained:

$$S = k' log(\frac{I}{I_T}) \tag{4.6}$$

where k' is a constant. See Fig. 4.9.

This law implies a logarithmic relationship between physical stimulus and perceived magnitude.

Steven's law, on the other hand, states that perceptual sensation and the physical stimulus are related through a power law:

$$S = kI^a \tag{4.7}$$

where S is the sensation, k is the proportionality constant, and a is an exponent that depends on the type of stimulus. In the case of lightness sensation, as explained in [48], a has a value of 0.42. See Fig. 4.10.

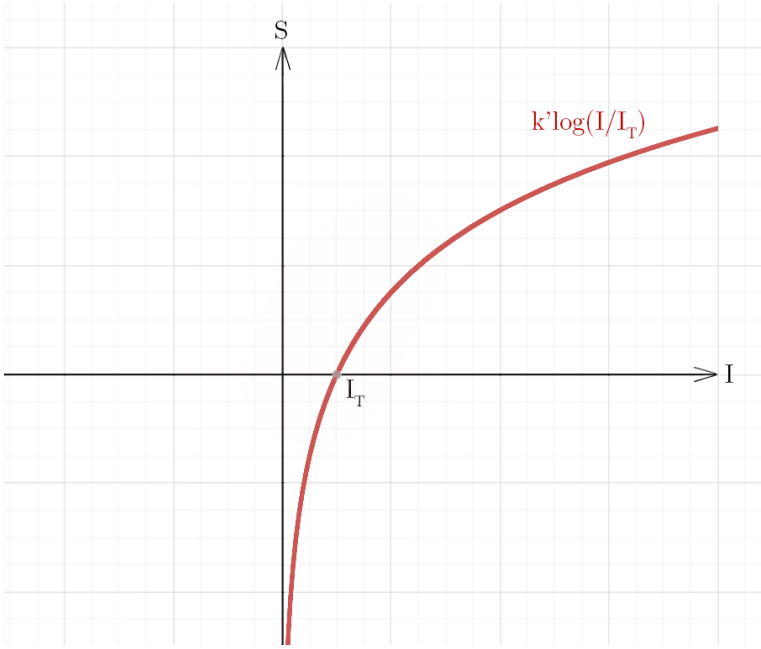

Figure 4.9. Weber-Fechner's law. Graph of the sensation S as a logarithmic function that depends on the intensity I.

4.4 Color perception and color spaces

This section adapts and summarizes some concepts explained in "Vision models for high-dynamic-range and wide colour gamut imaging" [4], Chapter 6. Color is an attribute of the visual perception, therefore is a subjective quality, not a physical property of light. Color sensations are associated with light wavelengths, reflectance of objects, and the sensitivity of the cone cells in the human eye. There exist models that predict color appearance in controlled environments, but for the case of natural images in arbitrary viewing conditions, there is not a vision model that can explain all the perceptual phenomena that come into play. Therefore, the color appearance problem remains very much open, see Fig. 4.11.

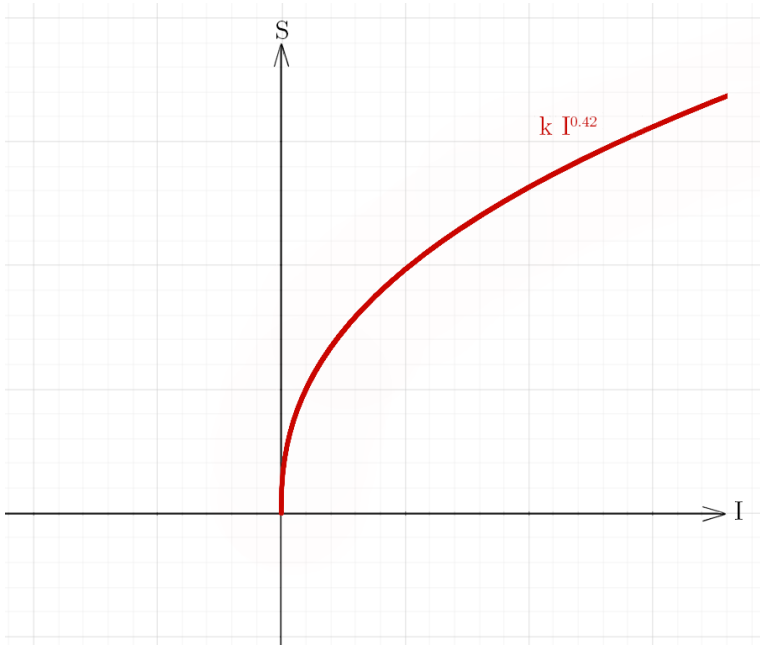

Figure 4.10. Steven's law. Graph of the sensation S as an exponential function that depends on the intensity I.

The trichromatic theory states that each color in the visible spectrum can be defined using a combination of three primaries. This is a property of HVS and not a property of light. It is a consequence of the fact that there exist three types of cones, and each of them detects specific wavelengths, see Eq. 4.3. Around 1930, experiments were performed by Wright and Guild to study trichromacy: subjects were asked to adjust the intensity of a set of red, green, and blue monochromatic lights (650, 530, and 460 nm), in order to color-match a given monochromatic light. The color matching functions $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$, were obtained by averaging the resulting functions of each subject (see Fig. 4.12).

Therefore, each perceived color can be defined by triplets of numerical values, corresponding to the contribution of the different

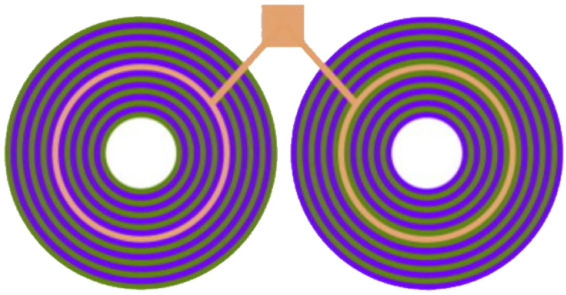

Figure 4.11. Wavelength does not determine color: the inner rings are identical, yet they appear to us as having different colors. Figure from [4].

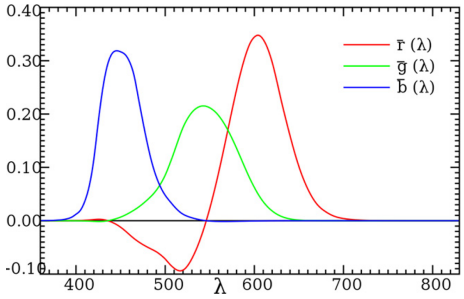

Figure 4.12. Color matching functions $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$. Figure from [27].

types of wavelengths. This mathematical model, which associates colors with tristimulus values (primaries in some additive color model), is known as *color space*.

4.4.1 The first standard color spaces

In 1931, the CIE, using the data collected from Wright and Guild's experiments, proposed two color matching functions: CIE RGB and CIE XYZ.

The **CIE RGB** color system is defined by the previously mentioned color matching functions $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$. The *RGB* tris-

timulus for a light source with spectral distribution $E(\lambda)$ are expressed as follows:

$$\begin{aligned} R &= \int_{380}^{740} \bar{r}(\lambda)E(\lambda)d\lambda \\ G &= \int_{380}^{740} \bar{g}(\lambda)E(\lambda)d\lambda \\ B &= \int_{380}^{740} \bar{b}(\lambda)E(\lambda)d\lambda \end{aligned} \tag{4.8}$$

To represent any visible color, the red component of the CIE RGB color space sometimes becomes negative. To address this limitation, a new color space called CIE XYZ is defined, whose values are always positive.

The **CIE XYZ** tristimulus values XYZ, for a light source with spectral distribution $E(\lambda)$ are:

$$\begin{aligned} X &= \int_{380}^{740} \bar{x}(\lambda)E(\lambda)d\lambda \\ Y &= \int_{380}^{740} \bar{y}(\lambda)E(\lambda)d\lambda \\ Z &= \int_{380}^{740} \bar{z}(\lambda)E(\lambda)d\lambda \end{aligned} \tag{4.9}$$

where the color matching functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ are defined as a linear combination of $\bar{r}(\lambda)$, $\bar{g}(\lambda)$, and $\bar{b}(\lambda)$, by imposing certain constraints:

- $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ must be always positive.
- $\bar{y}(\lambda)$ is equal to the standard luminosity function $V(\lambda)$, see Fig. 4.4.
- $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$ are normalized so a white light has equal tristimulus values $X = Y = Z$.

Perceived colors can be defined in terms of *luminance*, and *chromaticity*, which is the quality of color regardless of its luminance.

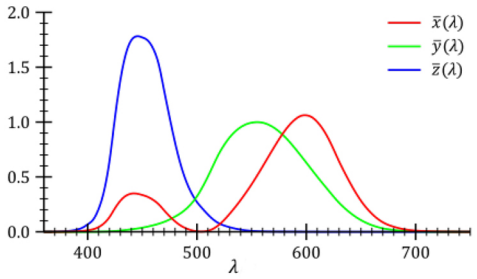

Figure 4.13. Color matching functions $\bar{x}(\lambda)$, $\bar{y}(\lambda)$, and $\bar{z}(\lambda)$. Image from [27].

We now define the values x, y, z as:

$$\begin{aligned} x &= \frac{X}{X+Y+Z} \\ y &= \frac{Y}{X+Y+Z} \\ z &= \frac{Z}{X+Y+Z} \end{aligned} \tag{4.10}$$

It can be easily observed that for lights E_1 and $E_2 = \alpha E_1$, their corresponding x, y, z values are identical. For this reason, the tristimulus values x, y, z are called the *chromaticity coordinates*, as they do not change if the light stimulus only changes its luminance. By construction, $x + y + z = 1$, then, all the chromaticity information is contained in the pair (x, y). Therefore all the chromaticity information can be represented in a plane, called the CIE xy chromaticity diagram (see Fig. 4.14).

It is worth remarking that any visible color can be defined by its chromaticity coordinates (x, y), and its luminance Y.

4.4.2 Perceptually uniform color spaces

The CIE XYZ is not a perceptually uniform color space; that is, the distance between two points in XYZ space is not proportional to the

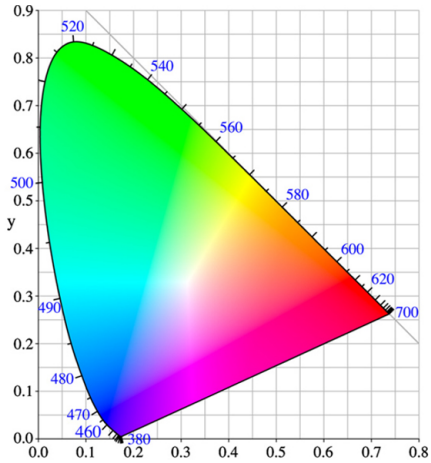

Figure 4.14. CIE xy chromaticity diagram. Figure from [4].

perceived difference between the colors corresponding to the points. Perceptual uniformity is a very useful property for color reproduction systems as it allows to define error tolerances.

In 1976, the CIE introduced the CIE $L^*a^*b^*$ color space, or **CIELAB**. It was intended as a perceptually uniform color space, with the channel L^* representing lightness, a^* representing red-green response, and b^* representing yellow-blue response.

The formula to transform the CIE XYZ values to CIELAB is:

$$\begin{aligned} L^* &= 116 f\left(\frac{Y}{Y_n}\right) - 16 \\ a^* &= 500\left(f\left(\frac{X}{X_n}\right) - f\left(\frac{Y}{Y_n}\right)\right) \\ b^* &= 200\left(f\left(\frac{Y}{Y_n}\right) - f\left(\frac{Z}{Z_n}\right)\right) \end{aligned} \tag{4.11}$$

where X_n, Y_n, Z_n are the tristimulus values of a reference white, and f is the function:

$$f(x) = \begin{cases} x^{\frac{1}{3}} & \text{if } x > \left(\frac{6}{29}\right)^3 \\ \frac{1}{3}\left(\frac{29}{6}\right)^2 x + \frac{4}{29} & \text{otherwise} \end{cases} \tag{4.12}$$

First of all, CIELAB performs a normalization with respect to the tristimulus values X_n, Y_n, Z_n of a reference white. It emulates the adaptation to the ambient illuminant, therefore, the normalization is an approximation to the color constancy property of the HVS.

After that, CIELAB applies a power-law of value $\frac{1}{3}$ to estimate the lightness. It can be recalled, from Steven's law (Eq. 4.7), that lightness perception is related to luminance through a power-law.

Finally, the signals a^*, and b^* are calculated, based on the theory of color opponency. This theory, developed by Hering in the late 19th century, states that there exist two opponent axes for color perception: the red-green axis, and the yellow-blue axis. Each color can be represented by two values, corresponding to the proportions of the color on each opponent channel. The a^*, b^* coordinates can be positive or negative, therefore CIELAB colors are often expressed in cylindrical coordinates $L^*C^*h^*$, where:

- $C^* = \sqrt{a^{*2} + b^{*2}}$
- $h^* = arctan\left(\frac{b^*}{a^*}\right)$

where C^* represents the *chroma*, which is the color intensity, or degree of colorfulness with respect to a white color of the same brightness, and h^* is the *hue*, which represents the basic color. An angle $h^* = 0°$ corresponds to red, $h^* = 60°$ corresponds to yellow, $h^* = 120°$ corresponds to green, etc (see Fig. 4.15).

The CIELAB color space was introduced to overcome the non-uniformity of the previous ones, still, it is not fully uniform. Experiments show that in some parts of the color space (mainly around blue), CIELAB suffers from *cross-contamination*, that is changing a color attribute produces perceptible effects on the other attributes.

In 1998, Ebner and Fairchild proposed the **IPT** color space [19], to improve hue uniformity. This model, based on some experimental results, first transforms the (X, Y, Z) values to (L, M, S) cone tristimulus values, then the cone responses are transformed by a power law of value 0.43, and finally, a linear transformation is applied to

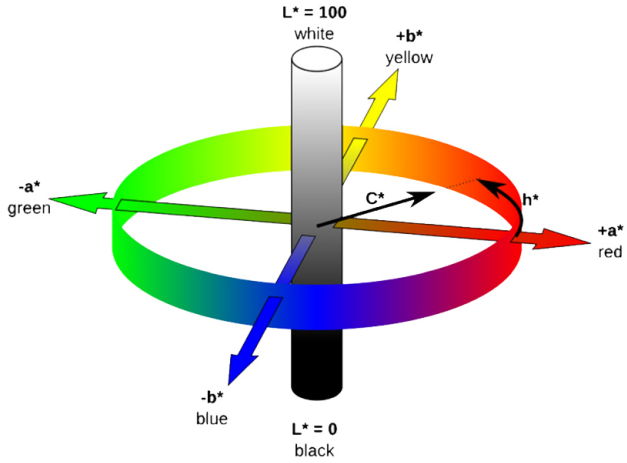

Figure 4.15. CIE $L^*a^*b^*$ color space in both cartesian and cylindrical coordinates. Figure from [4].

obtain the (I, P, S) tristimulus values. The formula to transform the CIE XYZ values to IPT is:

$$\begin{bmatrix} L \\ M \\ S \end{bmatrix} = \begin{bmatrix} 0.4002 & 0.7075 & -0.0807 \\ -0.2280 & 1.1500 & 0.0612 \\ 0 & 0 & 0.9184 \end{bmatrix} \begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \tag{4.13}$$

$$\begin{aligned} L' &= \begin{cases} L^{0.43} & \text{if } L \geq 0 \\ -(-L)^{0.43} & \text{if } L < 0 \end{cases} \\ M' &= \begin{cases} M^{0.43} & \text{if } M \geq 0 \\ -(-M)^{0.43} & \text{if } M < 0 \end{cases} \\ S' &= \begin{cases} S^{0.43} & \text{if } S \geq 0 \\ -(-S)^{0.43} & \text{if } S < 0 \end{cases} \end{aligned} \tag{4.14}$$

$$\begin{bmatrix} I \\ P \\ T \end{bmatrix} = \begin{bmatrix} 0.4000 & 0.4000 & 0.2000 \\ 4.4550 & -4.8510 & 0.3960 \\ 0.8056 & 0.3572 & -1.1628 \end{bmatrix} \begin{bmatrix} L' \\ M' \\ S' \end{bmatrix} \tag{4.15}$$

In the same year, Ruderman et al. [74] proposed the $\mathbf{l\alpha\beta}$ color space, based on the efficient coding theory, which states that the human visual system is optimally designed to process natural images. They gathered a set of natural images and found a logarithmic color space, wherein decorrelation produced three principal orthogonal axes. The first axis corresponds to radiance information, and the two other axes are reminiscent of the color opponency theory: one corresponds to the yellow-blue chromatic-opponent mechanism and the other to the red-green one.

The formula to transform the LMS information to $l\alpha\beta$ values consists of two stages: first, the LMS values are transformed to logarithmic values (the reason for this is that the natural images data showed a great deal of skew in the LMS cone space, being concentrated near the origin):

$$\begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{M} \\ \boldsymbol{S} \end{bmatrix} = \begin{bmatrix} \log L \\ \log M \\ \log S \end{bmatrix} \tag{4.16}$$

Then, the logarithmic values are transformed to $l\alpha\beta$ by a matrix multiplication:

$$\begin{bmatrix} l \\ \alpha \\ \beta \end{bmatrix} = \begin{bmatrix} \frac{1}{\sqrt{3}} & 0 & 0 \\ 0 & \frac{1}{\sqrt{6}} & 0 \\ 0 & 0 & \frac{1}{\sqrt{2}} \end{bmatrix} \begin{bmatrix} 1 & 1 & 1 \\ 1 & 1 & -2 \\ 1 & -1 & 0 \end{bmatrix} \begin{bmatrix} \boldsymbol{L} \\ \boldsymbol{M} \\ \boldsymbol{S} \end{bmatrix} \tag{4.17}$$

The advantages of the $l\alpha\beta$ color space are:

- Decorrelation between axes.
- Logartihmic color space, which means that uniform changes in channel intensity are equally detectable, following the Weber's law of perception 4.4.
- Description of the data according to its perceptual information, following the chromatic-opponency theory.

4.5 Conclusions

In this chapter, a basic knowledge of light and how the visual system works has been developed, in order to understand some commonly used models of brightness perception and color perception. The development of these models is important due to their applications. As it has been seen in Chapter 3, brightness perception models are crucial for camera manufacturers, to develop efficient quantization functions in the camera pipeline. On the other hand, accurate perceptual color spaces are of great importance for color reproduction systems, as they allow to define error tolerance. Extensive research is being carried out to develop uniform color spaces, and it follows two independent lines: finding a uniform lightness scale, and devising a uniform chromaticity diagram for colors of constant lightness [4]. The perceptual color spaces previously exposed are based on experiments that used data whose dynamic range was limited, compared to High Dynamic Range (HDR) imaging. Therefore, there currently exists an active line of research to develop color spaces specifically designed for HDR systems.

5 Photorealistic style transfer

In this section, we describe the importance of a process called *color grading* in cinema, and the current solutions in the academic literature and in the cinema industry. Our contribution in this chapter is to propose a method to automatically transfer the style from a reference image to the original unprocessed footage, easing the workload of cinematographers. The computational cost of the method is low, making it amenable for real-time implementation. It can be used on-set, allowing cinematographers to experiment with different looks and styles.

This chapter is an adaptation of our article "Photorealistic style transfer for video" published in *Signal Processing: Image Communication* [95], and the conference paper "In-camera, photorealistic style transfer for on-set automatic grading" published in *SMPTE 2018* [94]. The code of the method can be found at: `https://github.com/izabalra8/VideoStyleTransfer`.

5.1 Motivation

Color plays a critical role in cinema, as it greatly affects how we perceive the film and the characters. It can build harmony or tension

within a scene, or bring attention to a key theme. If it is chosen carefully, a well-placed movie color palette evokes mood and sets the tone for the film. The first movie to be fully digitally color graded is the Coen Brothers' 2000 film, *O' Brother, Where Art Thou?*, that uses a sepia-tinted color palette to evoke its setting of rural Mississippi during the time of the Great Depression. Before deciding to digitally manipulate the footage, the director of photography (DoP) of the film tried to obtain the look of the movie by using different types of photo-chemical processes at a film lab, but his attempts were unsuccessful due to the huge difference between the recorded images and the desired look: the colors of the footage needed to be completely altered, as it can be observed in Fig. 5.1. They finally decided to use a digital intermediate process. This process consists in digitizing the film negative after it has been edited, so the colors and other image characteristics can be modified digitally. Once the whole film has been color graded and assembled on the digital station, it is recorded back out to film again, using a laser film recorder. Then, the print is approved by the cinematographer, and multiple prints are made that go to the movie theatres.

In the ideal scenario for color grading, the intended color look would be conceived beforehand, then, carefully enhanced and rendered by the colorist during post-production in a stage called color grading. In the past, color grading was a photochemical process performed at a photographic laboratory, but nowadays it is generally performed digitally in a color suite. Usually, in professional cinema, 3D lookup tables (LUTs) are created under the supervision of the DoP to set a look as well as to endow images with specific looks. Later on, these 3D LUTs along with the ungraded footage are passed to the post-production stage, where the LUTs are used as a starting point for the final color grading of the movie. This whole process is very costly in terms of budget and time and it may require a significant amount of work from very skilled artists and technicians.

Figure 5.1. Top image: footage before color grading, bottom image: color graded image of the final movie. Images from the documentary *Painting With Pixels: (O' Brother, Where Art Thou).*

5.2 Current methods

In the last years, style and color transfer have received significant attention from the image processing and computer vision community. The methods proposed so far have numerous applications: generate more realistic renderings, some methods can be used for panorama stitching, tone mapping may employ color transfer techniques, they can be used for color stabilization (match images from the same scene taken with different camera configurations), etc. In the field of cinema, style transfer techniques applied to the video footage can be very helpful, in terms of time and work savings. However, there are still open research problems on how to extend the current color transfer approaches to video content, or how to achieve realistic results, free of artifacts, acceptable for cinema quality standards, etc.

We will summarize some of the most significant approaches in this field, that are also important for explaining our proposed style transfer method.

5.2.1 Color and style transfer

Since Reinhard et al. [69] presented their pioneering work about *color transfer*, this has been an active research topic. They describe their method as a form of color correction for borrowing the color properties of one image from another image (see Fig. 5.2). Their technique takes advantage of the decorrelation property of the $l\alpha\beta$ color space (explained in Chapter 4), and transfers simple statistical moments (mean and standard deviation) between each channel of the two images, i.e. from reference to source image. This method works well in many scenarios, but it is restricted in $l\alpha\beta$ color space, which is constructed with the aim of decorrelating natural images on average, but not to decorrelate the specific images that are being modified.

Some methods try to overcome this restriction by applying Principal Component Analysis (PCA) to individual images [36], [92], finding a dedicated color space for each image. The PCA-based method

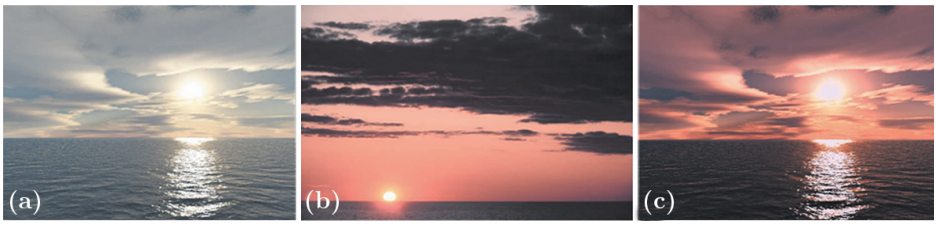

Figure 5.2. Color transfer by Reinhard et al. [69]. a) Source image, b) reference image, and c) color transferred image. Adapted figure from [69].

proposed by Kotera [36] will be explained more in detail, as it will be important to understand our proposed method. In statistics, *Principal Component Analysis* is a technique that uses an orthogonal transformation to convert a set of observations of possibly correlated variables into a set of values of linearly uncorrelated variables, called principal components. This transformation is calculated through the covariance matrix of the data (which is a symmetrical matrix), finding its spectral decomposition. In the eigenvector basis (formed by orthogonal vectors), the covariance matrix becomes diagonal, therefore the variables are uncorrelated in this basis. In this context, the principal components are the eigenvectors of the covariance matrix, that also satisfy:

- The first principal component is the direction that maximizes the variance of the projected data (or equivalently, minimizes the average squared distance from the data points to the vector line). Therefore, it corresponds to the eigenvector associated with the greatest eigenvalue.
- The i-th component has the highest variance possible, under the constraint that it is orthogonal to the first $i-1$ components.

In the work of Kotera [36], the principal components of the source color cluster are matched to that of the reference by multiplying by a matrix M, obtained as follows:

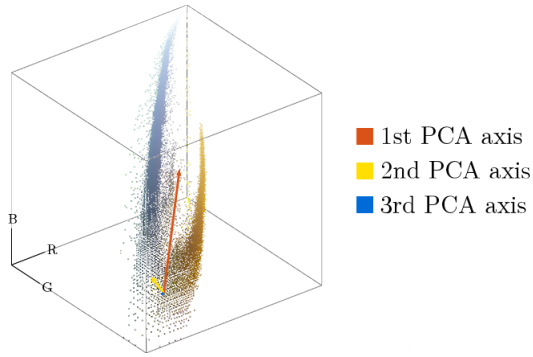

Figure 5.3. PCA of an image. The first PCA axis (in red) has the highest variance (9%), the second PCA axis, shown in yellow, has the second highest variance (2%), and the third PCA axis, shown in blue, has a variance of 0.1%. The PCA axes have been scaled by their variance. The origin of the PCA vectors indicates the mean of the observations.

$$M = A_{ref}^{-1} \cdot S \cdot A_{src} \tag{5.1}$$

where A_{src} and A_{ref} are the eigenvector matrices for the source and reference color clusters respectively, and S is a diagonal matrix with the eigenvalues' ratio, given by:

$$S = \begin{pmatrix} \sqrt{\frac{\lambda_{1_{ref}}}{\lambda_{1_{src}}}} & 0 & 0 \\ 0 & \sqrt{\frac{\lambda_{2_{ref}}}{\lambda_{2_{src}}}} & 0 \\ 0 & 0 & \sqrt{\frac{\lambda_{3_{ref}}}{\lambda_{3_{src}}}} \end{pmatrix} \tag{5.2}$$

As it is shown in Fig. 5.4, the matrix M has two functions: It matches the PC axes by rotating the cluster along the eigenvectors (matrices A_{ref}^{-1} and A_{src}), and it matches the variances by scaling the color distribution according to the eigenvalues' ratio (matrix S). As it is explained [36, 92], a PCA-based method (and more generally, any statistics-based method) is a global transformation, therefore,

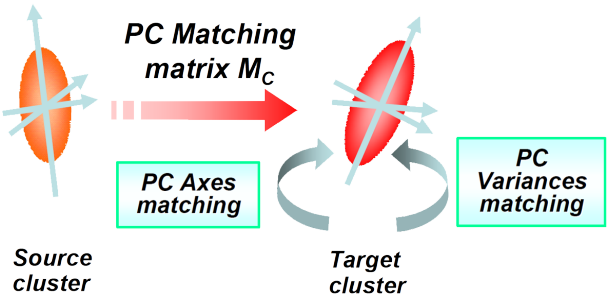

Figure 5.4. Color matching using PCA. The principal components of the source image are matched to the principal components of the reference, and the standard deviation along each axis is transferred from the reference to the source. Figure from [36].

it has some limitations. In some cases, the color characteristics of the image cannot be represented by a single cluster. Results can be improved by having user interaction such as segmenting the images [7, 36, 44], or by recovering dense pixel correspondences between the images [31, 84]. Vazquez-Corral and Bertalmío [84] propose a method for color stabilization (match the colors of images of the same scene taken with different cameras). They assume that, in a standard camera pipeline, the output values can be expressed as:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix}_{out} = \left(A \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{in} \right)^{\gamma} \tag{5.3}$$

where RGB_{in} are the camera RAW values at a given pixel location. Then, using SIFT, they find pixel correspondences between the images that have to be color matched. The corresponding pixels p_1 and p_2 satisfy:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix}_{p_1} = \left(A_1 \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{p} \right)^{\gamma_1} \quad ; \quad \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{p_2} = \left(A_2 \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{p} \right)^{\gamma_2} \tag{5.4}$$

Therefore, to obtain the values of p_1 from the pixel values of p_2 the following formula has to be applied:

$$\begin{bmatrix} R \\ G \\ B \end{bmatrix}_{p_1} = \left(A_1 \cdot A_2^{-1} \begin{bmatrix} R \\ G \\ B \end{bmatrix}_{p_2}^{\frac{1}{\gamma_2}} \right)^{\gamma_1} \tag{5.5}$$

Assuming that the values γ_1 and γ_2 can be estimated from the images, the only unknown variables are A_1 and A_2. Let us call H the matrix $H = A_1 \cdot A_2^{-1}$. The specific values of A_1 and A_2 are not needed to obtain the transformation that matches image 2 to image 1, an estimation of the value of H is enough. Then, the transformation is obtained by an optimization procedure, finding the matrix H that minimizes the error in equation 5.5 for all the pairs of corresponding pixels p_1, p_2 found in the previous step.

Optimal transport is another popular framework for performing color transfer between images. The work by Frigo et al. [24] uses the Monge-Kantorovich formulation to map a pair of meaningful color palettes. Optimal transport theory is also used in the work of Ferradans et al. [23] and the work of Rabin et al. [63], where they incorporate a spatial regularization of colors and a relaxation of the bijectivity constraint in order to avoid artifacts. An alternative framework to this theory has been recently proposed by Grogan et al. [30], where they use a cost function defined as the $\mathcal{L}_2$-divergence between two color distributions (modeled as compact Gaussian mixtures). It must be noted that in these types of approaches a minimization of the distance between color distributions can lead to the creation of color artifacts.

In the last years, there has been a line of research that has increased in popularity: style transfer using convolutional neural networks, which was first introduced by Gatys et al. [26]. These methods were initially designed to transfer the style of an artwork to a photograph, and they are able to produce good non-realistic results from images with very different content and style; however, they fail in the case of producing photorealistic results free of painting-like

distortions. A semantic segmentation of the source and reference images has been proposed [44], in order to adapt these methods to transfer the style between photographs. Although the results look visually more satisfying compared to the previously mentioned approach, user interaction is required and they are not completely free of painting-like artifacts, see Fig. 5.5.

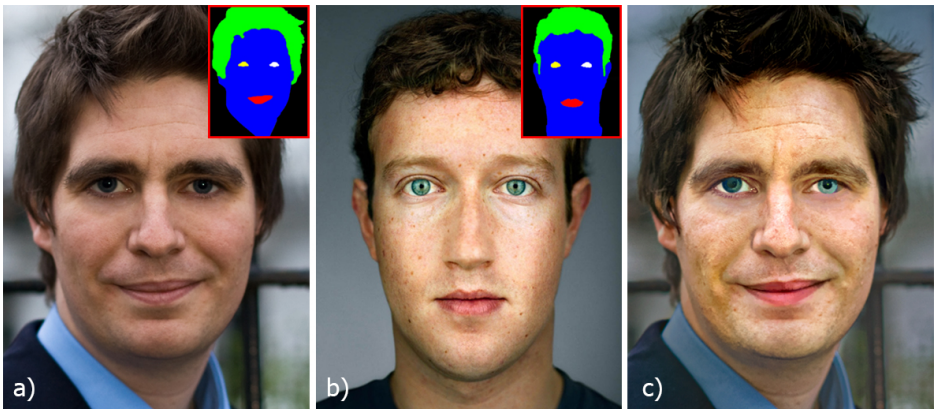

Figure 5.5. Photorealistic style transfer result, using a deep-learning approach combined with a segmentation mask, by Luan et al. [44]. a) Source image, b) segmentation mask, c) reference image, d) resulting image. Convolutional neural network approaches may produce artifacts, creating non-photorealistic results. Adapted figure from [44].

Recently, Li et al. [41] proposed a state-of-the-art method that aims to overcome these limitations by adding a smoothing step to the stylization step (see Fig. 5.6); moreover, it is faster than the previous photorealistic deep-learning approaches. However, when segmentation masks are not used, this method often generates inconsistent results with noticeable color and spatial artifacts.

Figure 5.6. Deep-learning approach of style transfer, using a segmentation mask, by Li et al. [41]. a) Source image, b) reference image, c) resulting style transferred image. Adapted image from [41].

5.2.2 Video style transfer methods

While many image color transfer methods exist, only a few video color transfer algorithms have been proposed so far. One state-of-the-art color transfer method for video is the one proposed by Bonneel et al. [7]. It consists of a per-frame chromaticity color transfer based on image statistics, followed by a curvature-flow smoothing to avoid color bleeding and flickering that may appear when applied to videos. As they explain in their work, some color styles have spatially varying characteristics that cannot be replicated with global color adjustments, therefore they use a user-specified segmentation to produce results that are more faithful to the style (see Fig. 5.7).

Figure 5.7. Video style transfer, using segmentation masks, by Bonneel et al. [7]. a) Source image, b) reference image, c) resulting style transferred image.

5.3 Proposed approach

We propose a method that can ease the workload of cinematographers and colorists during shooting and post-production. In many cases, the director wants to emulate the style and look present in a reference image, e.g. a still from an existing movie, a photograph, or even a previously shot sequence in the current movie. Given a color graded reference image, our approach automatically transfers the style, in terms of tone, color palette and contrast to the source ungraded footage: the algorithm is applied directly on the unprocessed images and it generates a display-ready result that matches the style of the reference image (see Fig 5.8). It is a low computational cost process that can be implemented in-camera, so the lighting and scene elements can be adjusted on-set while seeing the resulting image on the screen. While the method is proposed as a substitute for some of the post-production tasks, it is compatible with further refinements, both on-set and in post-production.

Some preliminary concepts will be reviewed before going any further: tone-mapping and contrast normalization.

Tone-mapping is the process of reducing the dynamic range of an image while reproducing the visual appearance and preserving the perceived detail (as it was defined in Chapter 3). There exist two types of tone-mapping operators:

- Global operators, which apply the same transformation to all the pixels in the image. These methods are computationally efficient, as they can be implemented in lookup tables, at the cost of losing contrast and image detail.

- Local operators, whose parameters change for each pixel in the image accordingly to the local features of the image at that specific pixel. These methods are inspired by the local adaptation process of the human visual system, which is the ability to adapt to different luminance levels when viewing natural images. These local operators produce images with more contrast

Figure 5.8. After applying our style transfer method to the original image (on top), it can be observed that the colors in the background of the reference image (in the middle) are correctly transferred to the background of the source image, as it is shown in the resulting image (in the bottom).

and higher detail level, but they are computationally more expensive, and they may show artifacts, such as halos around high-contrast edges.

A naive solution for dynamic range reduction is to linearly scale the image values, but it would result in loss of detail, as it can be observed in the example of Fig. 5.9. Therefore, tone-mapping methods are usually based on visual appearance models, making use of some knowledge of psychophysics and perception.

Within the global operators, the method proposed by Cyriac et al [14] will be explained more in detail, as it will be mentioned in the following sections. This method performs a constrained histogram equalization, and it is based on some properties of natural image statistics. It has been found [34, 73] that natural images have a cumulative histogram that can be modeled as:

$$H(x) = x^{\gamma(x)} \tag{5.6}$$

where x is the luminance value, and γ is a piece-wise function. In [14], the function γ increases linearly with some slope γ_L until the intensity value M, and then, with a different slope γ_H, with a smooth transition between the two slopes around M, as follows:

$$\gamma(x) = \gamma_H + (\gamma_L - \gamma_H)\left(1 - \frac{x^n}{x^n + M_{lin}^n}\right) \tag{5.7}$$

On the other hand, *histogram equalization* is a well-established method for increasing the contrast (and therefore, the perceived detail), using the image histogram. Let H be the cumulative histogram of an image, histogram equalization consists in applying the transformation

$$I_{eq}(i,j) = H\left(I(i,j)\right) \tag{5.8}$$

to every pixel $I(i,j)$ of the image I, where (i,j) stands for each specific pixel location. This transformation spreads the histogram of the original image, so the intensity levels of the equalized image

Figure 5.9. HDR image depicting outdoor and indoor information. Comparison of linear scaling (top) with a tone-mapped result (bottom). It can be observed that linear scaling leads to a loss of detail information in the dark areas. Images from [70].

span a wider range of the intensity scale, resulting in contrast enhancement. However, complete histogram equalization can lead to unnatural looking images. Therefore, the method proposed by Cyriac el al. [14], applies a constrained histogram equalization defined as follows: first, the transformation $\gamma(x)$ (defined in Eq. 5.7) is calculated; the values γ_L, γ_H, and M are obtained from the cumulative histogram of the considered image. Then, a constrained histogram equalization is done by applying the transformation:

$$H\Big(I(i,j)\Big) = \Big(I(i,j)\Big)^{\gamma(I(i,j))} \tag{5.9}$$

as this is the function that models the average cumulative histogram of natural images.

Contrast normalization is a process present in the human visual system which consists in scaling the contrast by a factor that depends on the standard deviation of light intensity, where contrast is the difference between light intensity and its mean value (as explained in [8]).

In this context, we can now describe the proposed algorithm, which can be summarized as follows:

1. Linearize the encoded source video sequence.
2. Calculate a style transfer transformation for a selected frame in the sequence.
3. Apply the transformation calculated in the previous step to all the frames in the sequence.

5.3.1 Linearization of encoded source video

As a first step, our method applies to the input footage the inverse of the encoding non-linearity (the main encoding techniques have been explained in Section 3.1) so as to ensure that the source content is linear. This assumes that the non-linearity of the source material is

known, which is the case in all practical shooting scenarios. In the case that the source footage is RAW data from the camera (demosaicked and white balanced), this step is not required because the information is already in linear form.

5.3.2 Style transfer for a still image

Once the source footage has been converted to linear information, a style transformation is calculated for a selected frame (typically the first one) in the sequence. This transformation consists of luminance, color, and contrast transfer steps.

Luminance transfer

Let S_0 be a frame of the linearized footage (that we will call source image) and let R be the color graded image used as reference. Luminance transfer consists in applying a transformation to the source image S_0, so as to match the luminance of the reference image R.

The transformation calculated in this step is based on the tone-mapping approach proposed by Cyriac et al. [14], explained above. This tone-mapping algorithm performs a constrained histogram equalization to the original image, creating an image I_{eq}:

$$I_{eq}(i,j) = TM\Big(I(i,j)\Big) = \Big(I(i,j)\Big)^{\gamma(I(i,j))} \tag{5.10}$$

where $I(i,j)$ is the pixel luminance value at the pixel location (i,j), and the function γ has been defined in Eq. 5.7.

On the other hand, we will assume that the reference image R has been encoded with the standard gamma correction formula gc (as it has been defined in Chapter 3, Eq. 3.11):

$$gc(V) = \begin{cases} 12.92V & \text{if } 0 \leq V \leq 0.0031308 \\ 1.055V^{1/2.4} - 0.055 & \text{if } 0.0031308 \leq V \leq 1 \end{cases} \tag{5.11}$$

where V denotes pixel value in any of the color channels. Therefore, $gc^{-1}(R)$ will be the linearized version of the reference R.

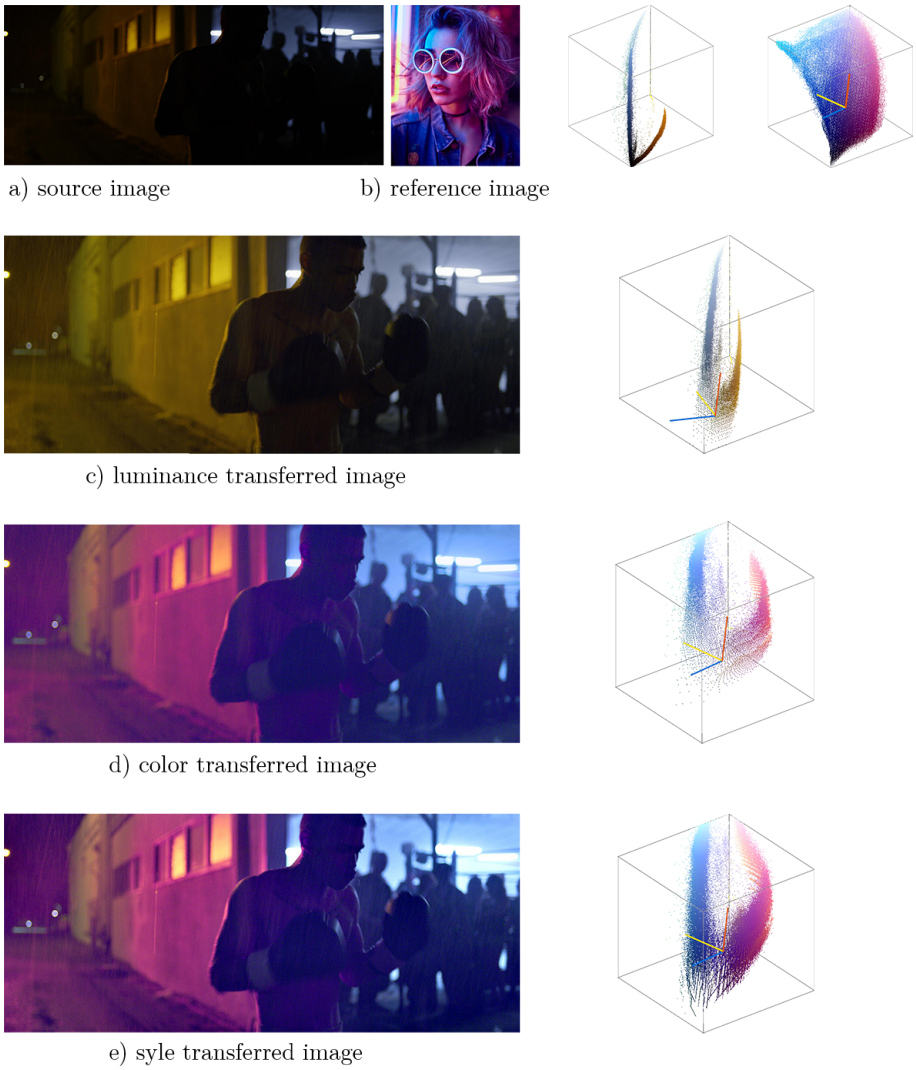

Figure 5.10. Our style transfer method consists of three steps: First, the luminance is transferred, then the colors are matched, and finally, the contrast is transferred from the reference image to the source image. See image credits on Section 5.7.

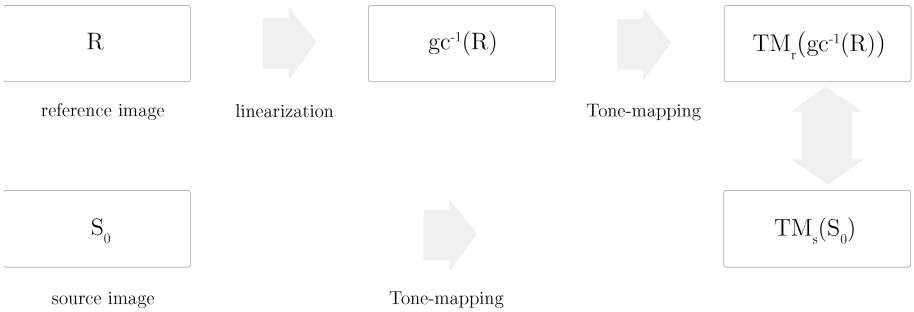

Figure 5.11. Estimation of the transformation to be applied to the source image S_0 so its luminance matches the luminance of the reference image R.

According to the diagram in Fig. 5.11, the histogram of $TM_s(S_0)$ can be considered approximately equal to the histogram of $TM_r\Big(gc^{-1}(R)\Big)$, as the tone-mapping method of Eq. 5.10 performs an approximate histogram equalization, therefore both histograms are approximately flat. Then:

$$\text{histogram}\Big(TM_s(S_0)\Big) \approx \text{histogram}\Big(TM_r(gc^{-1}(R))\Big) \tag{5.12}$$

where TM_s and TM_r are the tone-mapping functions calculated for the source image S_0, and for the reference image R, respectively.

Therefore, and following the diagram in Fig. 5.11, the transformed source image S_1 that matches the luminance of the reference R will be:

$$S_1 = gc(TM_r^{-1}(TM_s(S_0)) \tag{5.13}$$

Color transfer

We seek to transfer the colors of the reference image R to the image S_1 (resulting from the previous step). Following the same line as in

[36, 69, 92], we transfer the statistics (mean and standard deviation) along each channel separately in a decorrelated color space. Principal Component Analysis (PCA) is applied to the RGB source and reference images to find decorrelated spaces for each of them, but in our method there exist some modifications with respect to the traditional color transfer methods based on PCA analysis, as we will explain below. Additionally, our method can incorporate a region of interest previously selected by the user.

As it has been explained in Section 5.2.1, the transformation proposed by Kotera [36] for color transfer consists in a pixel-wise multiplication by a matrix M, that associates the following pair of points of the source and reference images:

- The points p_m and p'_m, that represent the mean value of the source and reference images respectively, (therefore $M \cdot p_m = p'_m$).

- The vectors $\sqrt{\lambda_{i_{src}}} v_i$ and $\sqrt{\lambda_{i_{ref}}} v'_i$, where v_i and v'_i represent the i-th PCA axes of the source and reference images respectively, λ_i the i-th eigenvalues, and $i = \{1, 2, 3\}$, (therefore $M \cdot \sqrt{\lambda_{i_{src}}} v_i = \sqrt{\lambda_{i_{ref}}} v'_i$).

The matrix M is associated with the linear transformation that converts these source image points to those of the reference. Therefore, M is fully determined by these 4 pair of points.

In the same line as the work of Kotera, our transformation consists in multiplying by a matrix M_{CT}. This matrix also associates pairs of points of the source S_1 and reference R images, but there are some differences between our method and the method by Kotera: in our case, the standard deviation along the first axis of the source image PCA is not matched to that of reference, as this axis contains the luminance information of the image that has been already matched in the luminance transfer step; our method can incorporate additional statistical information from regions of colors to be matched,

previously manually selected. Then, our color transfer function, represented by the matrix M_{CT}, associates the following pairs of points:

- The points p_m and p'_m, that represent the mean value of the source and reference images respectively.
- The vectors v_1 and v'_1, that represent the first PCA axis of the source and reference images respectively. As it has been explained in [9, 51], the first axis of PCA contains the luminance information of the image, therefore the standard deviation along this axis is not matched because the luminance information has been already transferred in the previous luminance transfer step.
- The vectors $\sqrt{\lambda_{i_{src}}}v_i$ and $\sqrt{\lambda_{i_{ref}}}v'_i$, where v_i and v'_i represent the i-th PCA axes of the source and reference images respectively, λ_i the i-th eigenvalues, and $i = \{2, 3\}$. As it has been explained in [9, 51], the second and third axis of PCA contains the chromaticity information of the image.
- If there is user input, there is an extra pair formed by the points p_s and p'_s, where p_s is the mean of the selected region of interest in the source image in the chromaticity channels (that is the projection of the point on the plane formed by the second and third axes of the PCA), and p'_s is the mean of the selected region of interest in the reference image in the chromaticity channels.

We extend M_{CT} as a projective transformation with size 4×4 (inspired by the color stabilization approach of Gil et al. [28] and Vazquez-Corral and Bertalmío [84]). Therefore, the matrix M_{CT} associated with the color transfer function is calculated by solving the system of equations formed by all the conditions listed above:

$$M_{CT} \cdot \begin{bmatrix} R \\ G \\ B \\ 1 \end{bmatrix}_{src} = \begin{bmatrix} R \\ G \\ B \\ 1 \end{bmatrix}_{ref} \tag{5.14}$$

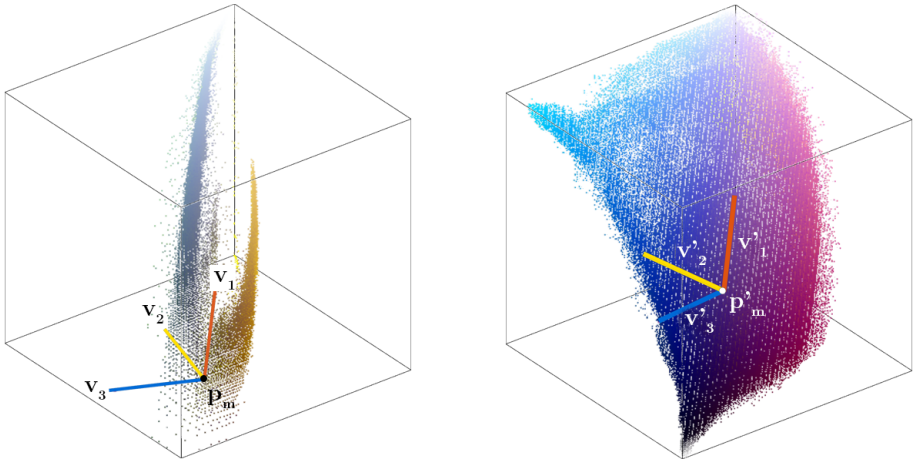

Figure 5.12. Source image on the left, reference image on the right. Statistical properties of the reference image (mean p'_m, PCA axes v'_i and variances along the axes) are transferred to the source image through the matrix M_{CT}.

for the points listed above (in homogeneous coordinates), and:

$$M_{CT} \cdot \begin{bmatrix} R \\ G \\ B \\ 0 \end{bmatrix}_{src} = \begin{bmatrix} R \\ G \\ B \\ 0 \end{bmatrix}_{ref} \tag{5.15}$$

for the vectors listed above (in homogeneous coordinates).

We have 16 unknowns for the 4×4 matrix M_{CT}. Each pair of point correspondences gives us 4 equations, and if there is no selection of a region of interest by the user, we have 4 point correspondences and therefore 16 equations: the matrix M_{CT} is calculated by solving this linear system of equations. In the case that the user selects a region of interest, there are 5 point correspondences, so 20 equations. Then, the system is overdetermined and the solution can be found by an optimization procedure that minimizes the distance between corresponding points. We use a least squares error approach to ap-

proximate the matrix M_{CT} that performs the color transfer, that is the matrix M_{CT} that minimizes:

$$\sum_j w_j \cdot (p'_j - M_{CT} \cdot p_j)^2 \tag{5.16}$$

where p_j and p'_j is each of the 5 pairs of points (or vectors) mentioned above.

Additionally, a weighted least squares approach can be used to give priority to some of the conditions that the matrix has to satisfy. By default, the weights are equally distributed with a value of $w_j = 0.2$, so each condition has the same importance as the others; however, these parameters can be adjusted by the user to give more priority to one condition over the others, see Fig. 5.13.

Then, if M_{CT} is the matrix calculated as it has been explained above, the resulting image S_2 from our color transfer transformation will be:

$$S_2(i,j) = M_{CT} \cdot S_1(i,j) \tag{5.17}$$

for each pixel location (i,j) in the image S_1.

Local contrast transfer

This step is based on the contrast normalization formula proposed by Cyriac et al. [14]. We assume that contrast information is kept in the luminance channel of the image, so the local contrast transformation is applied on the first axis of the PCA of the image S_2.

If we define *local contrast* of an image as the difference between light intensity and its local mean value, $I(x) - \mu(x)$, our approach transfers the standard deviation of the local contrast from the reference to the source image through the following formula:

$$S_3(x) = \mu(x) + (S_2(x) - \mu(x)) \cdot \frac{\sigma_{ref}}{\sigma_{src}}, \tag{5.18}$$

where S_2 is the image obtained after the color transfer step, $\mu(x)$ is the local mean of S_2 (μ is obtained by convolving the image with

Figure 5.13. From top to bottom: (a) Source and reference image, (b) style transfer result with no region of interest, (c) result with region of interest for the skin areas with weight $w = 0.2$, (d) result with weight $w = 0.4$. See image credits on Section 5.7.

a kernel W that is a linear combination of two Gaussian kernels, of standard deviation $\frac{1}{4}$ and $\frac{1}{16}$ of the height or width of the image, whichever dimension is smaller), σ_{src} is the standard deviation of the local contrast of S_2, and σ_{ref} is the standard deviation of the local contrast of the reference image R.

5.3.3 Video style transfer

Extending an image-based method to video sequences is not trivial: applying an independent style transfer to each frame of a video sequence might result in strong texture flickering and temporal incoherence, even if neighboring frames are similar in content and color. It is also computationally expensive to calculate the parameters for each frame without taking into account the common content between frames. To avoid that, we propose a simple but effective method (shown in Fig. 5.14) designed to handle these limitations. The style transfer transformations previously explained are calculated for a representative frame of the source sequence (e.g. simply the first frame of the video), then the same transforms are applied to each frame of the video, instead of calculating per-frame transformations. Temporal coherence is guaranteed by applying the same transformation to all the frames in the video. This approach has a very low computational cost and produces temporal-coherent and flickering-free results.

5.4 Results and experimental validation

We demonstrate style transfer results on a wide range of source and reference sequences, where the luminosity and color range vary from one to another. Fig. 5.8 and 5.10 are examples of our style transfer method. The algorithm has been tested with videos with resolution of 1920×1080 pixels, some of the resulting frames are shown in Fig. 5.14.

The computational cost of the algorithm is low so it is amenable for real-time implementation and in-camera processing. The lumi-

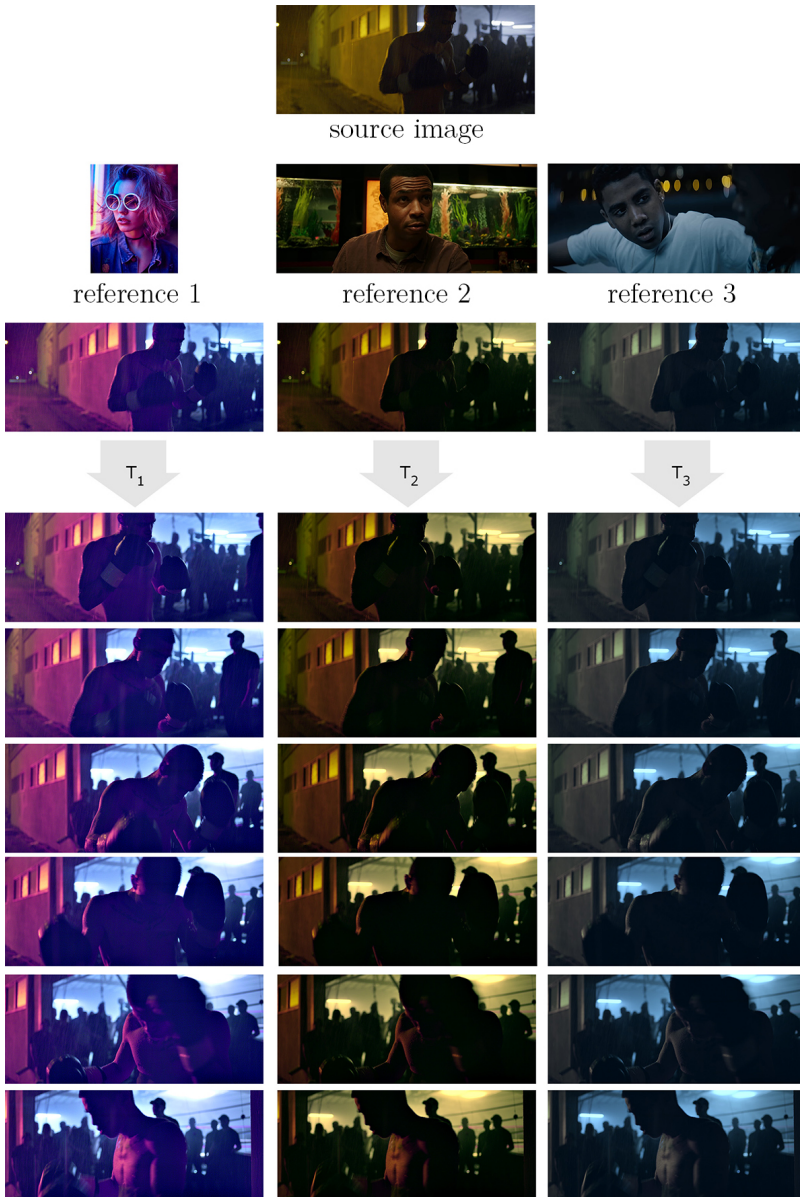

Figure 5.14. A transformation T is calculated for a given source image and reference image, the same transformation is applied to all the frames in the video. See image credits on Section 5.7.

nance transformation can be stored in a LUT, the color transformation matrix M_{CT} is obtained by an optimization procedure using a least squares error approach, and the contrast transformation uses a convolution, so the total transformation consists of a sequence of low-computational cost operations.

5.4.1 Psychophysical evaluation

We must remark that quantitative evaluation of stylization is very difficult since there is no ground truth. Therefore, we conduct psychophysical experiments in order to validate our results and to compare them with style transfer methods for still images from the state of the art in academia, like the optimal transport based algorithm by Grogan et al. [30] and by Rabin and Papadakis [63], or the deep-learning approach by Li et al.[41]; and in the industry as well, like the Color Match tool provided by the professional software Adobe Photoshop. For a fair comparison between methods, we use the automatic implementation of our method, so there is no region of interest selected by the user.

To test our approach we use a selection of 11 frames from Blackmagic Pocket Cinema footage and a selection of reference images. Those images reflect a variety of image scenarios, including outdoors and indoor scenes, portraits, daylight and night scenes, and different color palettes. For the evaluation, fifteen observers (3 women and 12 men) have taken part in the experiments, all of them having normal or corrected vision. We use a dark room where observers are instructed to sit approximately two picture heights away from a Sony PVM-250 25" Full HD OLED display (Rec709, gamma 2.4 calibrated). The experiment consists of two parts, in the first one we evaluate which is the preferred method, in the second one we measure the amount of artifacts that each method produces. In the first task, a two-alternative forced-choice comparison (2AFC) technique is used. For each comparison, the observer can navigate between four images: the source image, the reference image, and two style trans-

ferred images, each of them obtained from one of the methods. Then, observers are asked to choose between the style transferred images selecting the most similar to the reference image. In the second part of the experiment, the task consists in rating the amount of artifacts in the images, on a scale from 1 to 5: 1 for highest image quality, 5 for lowest image quality.

To compute accuracy scores from the raw psychophysical data we use the same approach as in [50], which is based on Thurstone's law of comparative judgment, and which we will now describe.

In order to compare n methods with experiments involving N observers, we create a $n \times n$ matrix for each observer where the value of the element at position (i, j) is 1 if method i is chosen over method j. From the matrices for all observers, we create a $n \times n$ frequency matrix where each of its elements shows how often in a pair one method is preferred over the other. From the frequency matrix we create a $n \times n$ z-score matrix: given a percentage of times that method i was chosen over method j, the corresponding z-score is the distance from the mean (on a scale whose units are the distribution's standard deviation) that corresponds to an area under the normal distribution's curve equaling the given percentage. The accuracy score A for each method is given by the average of the corresponding column in the z-score matrix. The 95% confidence interval is given by $A \pm 1.96\frac{\sigma}{\sqrt{N}}$, as A is based on a random sample of size N from a normal distribution with standard deviation σ. In practice $\sigma = \frac{1}{\sqrt{2}}$, because the z–score represents the difference between two stimuli on a scale where the unit is $\sigma * \sqrt{2}$ (in Thurstone's paper this set of assumptions is referred to as "Case V"); as the scale of A has units which equal $\sigma * \sqrt{2}$, then we get that $\sigma = \frac{1}{\sqrt{2}}$. The higher the accuracy score is for a given method, the more it is preferred by observers over the competing methods in the experiment.

Fig. 5.15 depicts the results of comparing the five methods mentioned previously. The experiments show that on still images, our method outperforms the methods from academia, and it is comparable to the Match Color tool from Photoshop, both in terms of fidelity

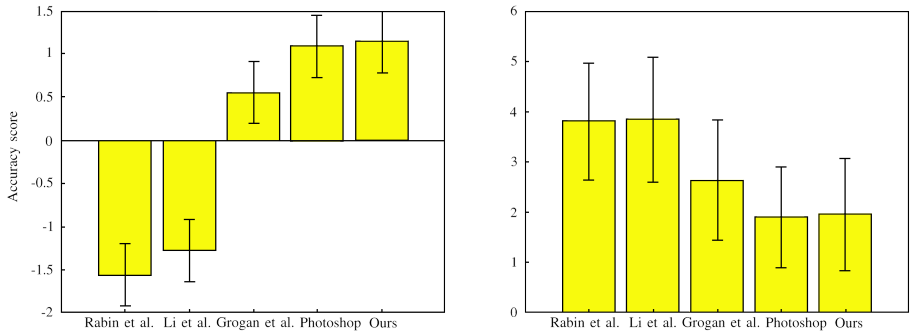

Figure 5.15. LEFT: Accuracy scores of competing methods, with 95% confidence intervals. The higher the accuracy score is for a given method, the more it is preferred by observers over the competing methods in the psychophysical experiment. RIGHT: Visual artifacts produced by competing methods, in a range from 1 to 5: 1 denotes the lowest amount of artifacts (highest image quality), 5 corresponds to highest amount of artifacts (lowest image quality.) Over each bar there is a 95% confidence interval.

to the reference and amount of artifacts.

Some visual comparisons of the images used for the experiments are shown in Figs. 5.16 and 5.17.

5.4.2 Video comparisons

In this section, a visual comparison has been made between our proposed method and the successful example-based video color grading approach of Bonneel et al. [7].

When the source video comes without foreground-background segmentation, both methods are on par in terms of visual quality and absence of artifacts, as the examples in Fig. 5.18 show.

On the other hand, when the source material has a user-specified segmentation in order to assist the matching, the style transfer procedure of Bonneel et al. [7] may produce visible color artifacts, while

a) Source image 1

b) Source image 2

c) Source image 3

Figure 5.16. Example results for the five different methods on some source images used in the psychophysical experiment. For each source image, the 6-panel block shows, from top to bottom and left to right: original image with reference shown as inset, results by: Rabin et al.[63], Li et al.[41], Grogan et al.[30], Photoshop, and proposed method. See image credits on Section 5.7.

c) Source image 4

d) Source image 5

b) Source image 6

Figure 5.17. Example results for the five different methods on some source images used in the psychophysical experiment. For each source image, the 6-panel block shows, from top to bottom and left to right: original image with reference shown as inset, results by: Rabin et al.[63], Li et al.[41], Grogan et al.[30], Photoshop, and proposed method. See image credits on Section 5.7.

our proposed method does not suffer from this shortcoming, as it can be observed in Fig.5.19.

Figure 5.18. Left: source frame, with reference shown as inset. Middle: result by Bonneel et al. [7], without using segmentation. Right: result by proposed method. See image credits on Section 5.7.

5.5 Limitations

Despite its overall good performance in a wide range of scenarios, the proposed method has some limitations that we want to illustrate in this section.

For scenes where foreground and background have very different color palettes, our method may not be able to transfer and separate colors properly, as it is based on a global color transformation that does not consider spatial locality, see Fig. 5.20. In these cases, a foreground/background segmentation procedure would probably solve the problem.

In our proposed method a single transformation is calculated for one frame and applied to all the frames in the video sequence, therefore the results might be sub-optimal if the raw content shows significant changes along the frame sequence. This is illustrated in Fig. 5.21, where between the first and the last frames in the video there

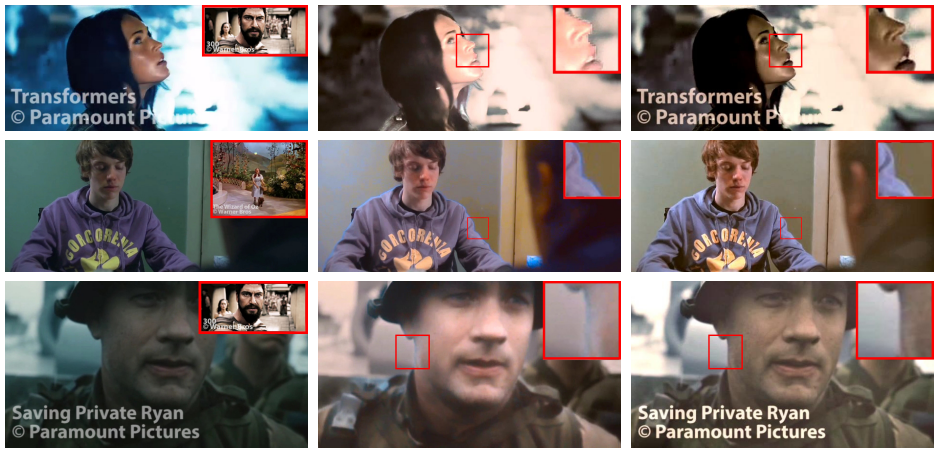

Figure 5.19. Left: source frame, with reference shown as inset. Middle: result by Bonneel et al. [7], using segmentation; zoomed-in detail shows artifacts produced by the method. Right: result by our proposed method, notice absence of artifacts. See image credits on Section 5.7.

is an apparent color difference in the skin tone. A possible solution for this problem would be to define a number of keyframes for the sequence, compute one style transfer transformation for each, and compute a temporal average of these transforms as the final process that is applied to the source video.

Finally, if the content of the source and the reference image are quite different, the results obtained by our algorithm might not be satisfactory, e.g. see Fig. 5.22.

5.6 Conclusions and future work

We have presented a method for transferring the style from a color graded reference image to unprocessed video footage that produces results that look natural and are free of artifacts. The computational complexity of the method is very low, appearing suitable for a real-

Figure 5.20. Style transfer failure due to a very different color palette between foreground and background. See image credits on Section 5.7.

Figure 5.21. Top: Original footage, initial and last frame of a video sequence in a video sequence. Bottom: Resulting images showing skin color variation along time. See image credits on Section 5.7.

Figure 5.22. Underwhelming result due to large difference between source and reference images. See image credits on Section 5.7.

time implementation in order to be used on the set. A psychophysical validation was conducted, showing that the proposed method outperforms algorithms from the state of the art in the academic literature and is comparable to the methods in the industry. Also, for video content, a visual comparison has been made between our method and one of the most successful methods of academia for automatic color grading. Both are on par in terms of visual quality, and when using segmentation masks, our method outperforms the other method as it is free of artifacts.

As future work we intend to incorporate to our method a number of extensions, like allowing for foreground/background segmentation and the use of keyframes, in order to overcome some of the limitations that have been observed.

5.7 Video credits

Image sequences are property of RED, Blackmagic, A24, and Paramount Pictures.

6

Adding texture to digital footage

We will start by outlining some reasons to add texture to digital images: on the one hand, there aesthetical reasons; on the other hand, we will explain how the addition of texture can be used with coding efficiency purposes, and it can also improve the perceived quality of images. Then, we will describe the most common approaches for adding texture to images. Finally, we will propose a method to add "retinal noise" to images that serves a double purpose: one is aesthetic, as it has parameters that allow to vary widely the resulting texture appearance, which make it an artistic tool for cinematographers; the other purpose is to improve the quality of compressed video by masking compression artifacts, which allows to lower the encoding bit rate while preserving image quality, and to improve image quality while keeping the bit rate fixed. This chapter is based on our work "Retinal noise emulation: a novel artistic tool for cinema that also improves compression efficiency" [96]. The method has been patented under the patent name: "Computer-implemented method for adding texture to a digital image". The code of the method can be found at: `https://github.com/izabalra8/retinalNoise`.

6.1 Motivation

Nowadays, some directors still use analog film for shooting their movies. At the same time, in cinema, it is standard practice to improve the appearance of digital images by adding noise that simulates film grain. The film grain texture is highly valued in the cinema industry and between photographers, and synthesized film grain is often added to digital images to reduce the "digital look".

On the other hand, in the media industry, there is a constant push in cinema, broadcast, and streaming services towards ever higher resolution, frame rate, and dynamic range. The accompanying increase in data volume that these new formats bring imposes considerable demands on transmission bandwidth and memory, and as a result the problem of compressing video is as relevant as ever. Specifically, Ultra High-Definition (UHD)/4K video requires bit rates nearly 10 times higher than Standard Definition (SD) video and Full HD (FHD)/2K around four or five times higher [1, 12]. In addition, video traffic is expected to account for 82% of all IP traffic by 2022. Moreover, by that year, UHD/4K video will be around 22% of IP video traffic, and FHD/2K video around 57% [12]. Faced with these remarkable data, content and service providers are constantly searching for ways to provide increasingly higher quality video and quality of experience (QoE) at restrained bit rates.

In this respect, the fact that most users are not able to perceive some objective quality drops under some conditions [72, 81] can be exploited. In fact, the visibility of image distortions is reduced by the presence of another stimulus, a masking pattern. This phenomenon is called "visual masking" and it is a well known property of visual perception. Visual masking takes several forms, as it depends on different properties of the image stimuli: luminance, color, temporal variations, spatial patterns. Visual masking is a key perceptual phenomenon for the design of image and video compression algorithms [80] (see Section 3.3), and in particular texture masking or pattern masking has been successfully applied for video coding [52].

Moreover, an established way to improve the appearance of digital images is to add to them a certain amount of fine-detail texture, and user studies have shown that observers indeed prefer images with some noise [37, 85]. In cinema and TV fiction the standard practice is to always add texture to digitally-shot content, and this texture invariably takes the form of film grain. This particular choice of texture aims to mimic the look of film, which is still considered as the gold standard by many cinematographers.

6.2 Related work

Film grain is the resulting texture from the silver halide based analog photographic process: the film emulsion contains photosensitive silver halide crystals, and when a photon hits a crystal, small particles of metallic silver are created. The non-homogeneous density of these particles produces what is known as film grain. As we will describe now, most of the methods to add texture to images consist in either synthesizing film grain or adding scannings of it. The texture is added at post-production stage.

In the movie industry, the most popular methods for film grain emulation are based on assembling a database of scannings of different types of film stock with varying forms of grain, that are superimposed on the digital image that is processed. For these methods there is a compromise between speed and the realism of the result: the fastest algorithms overlay grain that is somewhat independent from the content of the digital image, which may produce noticeable artifacts especially when there is motion, and the methods producing the more visually pleasing results are computationally very intensive, requiring special hardware.

The state-of-the-art and most relevant work in the academic literature is done by Newson et al. [54]. They propose a physical model of film grain which explains the distribution of grain in the film emulsion, in order to produce realistic-looking synthesized grain.

First, using stochastic geometry, the grain is modelled as a sequence of randomly distributed disks, whose centers are distributed in the plane following a Poisson law, and their radii follow a log-normal distribution. Secondly, a filtering step is done to produce the output grey-levels from the previous binary model. One advantage of this approach is that is able to produce grainy images at any resolution. The main limitation of the method is its computational cost, therefore the authors propose two alternative implementations whose computational complexities depend on the grain radius, and they also propose parallelization for a more efficient execution.

(a) Zoom 1x (b) Zoom 5x (c) Zoom 30x

Figure 6.1. The method by Newson et. al [54] is able to render images at any desired resolution. Figure from [54].

On the other hand, image content with high-frequency detail, like film grain, requires a higher bit rate in order to be compressed properly, i.e. for a given visual quality level, "clean" video requires less bits than video with film grain. For this reason, there are several works in the field of video compression [15, 55, 56] that improve coding efficiency by synthesizing film grain. The process consists, first, of denoising the input video (removing the film grain), then modeling the noise with a set of parameter values which are transmitted alongside the denoised video, and finally at the decoder re-synthesizing the film grain noise and adding it back to the decoded denoised video. We are not aware of any of these methods actually being used in practice in video streaming, and they require practical solutions to very challenging problems like video denoising.

Due to their computational complexity, existing methods impose a restriction on the creative work of filmmakers, preventing them from having on the set an accurate representation of how the movie will look in post-production after the synthetic texture is added. The emulation of film grain is also limiting because the artists are not able to really experiment with a wide diversity of texture options, nor to introduce novel looks.

6.3 Proposed framework: retinal noise

The proposed algorithm emulates retinal noise, and the motivation for using a retinal noise model is that the resulting images will have a more natural appearance, since noise is always present in retinal signals, in photopic (daytime) vision as well. Nonetheless, in our case the magnitude of emulated retinal grain that is added to day-like scenes is an artistic choice: it must be noticed that the method is proposed as an artistic tool and an aesthetic alternative to film grain emulation, rather than a physiologically accurate simulation of perceived noise. It has a very low computational complexity, so it is amenable for a real-time implementation that can be used on set. It has parameters that can be varied to achieve a wide range of texture appearances, allowing movie creators to try out new looks. Results are validated through psychophysical experiments in which observers, including cinema professionals, prefer our method over film grain emulation alternatives from academia and the industry.

Another contribution of our work is to show that the retinal noise emulation can also be used to improve the quality of compressed video by masking compression artifacts. Once the movie creator has taken the artistic decision to add a certain amount and type of this "retinal grain" to improve the look of the digital film, the movie distributor can use this fact to its advantage by encoding the "clean" content at a lower bit rate and adding the retinal grain after decoding, the exact same grain that the content creator decided was right for the

movie for aesthetic reasons, thus masking the visual artifacts produced by the reduced bit rate and yielding the same QoE of a higher bit rate. The extra data that is required at reception to introduce the retinal noise is negligible, as it only consists of the values for the user parameters (up to 5 floating point numbers per frame). This is completely novel because in the literature, as mentioned above, the grain is roughly estimated via a denoising process (which is an open problem), parametric models of film grain provide coarse approximations, and those works have a limited application because they are intended just for films with grain, whereas our approach can be used with any kind of content. We performed psychophysical experiments using color-graded professional cinema content shot in 4K, where the amount of retinal noise was selected by a motion picture specialist based solely on aesthetic preference. This content was encoded at different bit rates, and the retinal grain was added after decoding. The participants rated the quality of the resulting videos, and the results show that when reducing the bit rate, the loss of perceived quality is consistently smaller when the video has had retinal grain added to it than when it has not. Our method is shown to yield remarkable savings in bit rate, of over 22.5% on average.

Additionally, we can mention some other application scenarios for our method. As the proposed scheme is able to provide better subjective quality to compressed video, it can also be applied to scalable video, thus opening its use in the adaptive bitrate scenarios considered in streaming applications. Even more, it can be applied in novel multicamera scenarios, like high quality free viewpoint video, where the synthesis artifacts due to occlusions and missing data could be hidden by the addition of retinal noise. Finally, our method can be applied in-camera to enhance photographs and video, especially in the case of acquisition devices with limited capabilities.

As a closing point, we want to briefly discuss the possibility of our proposed work being replaced by a deep neural network (DNN) procedure. In our opinion, given that the applications discussed in this paper are all based on the perceived (aesthetic) appearance of

images and videos, the use of a DNN for these tasks would first require the ability of said DNN to represent aesthetic preference, and while there are some recent works in this regard, e.g. [62, 101], they have been shown to be unsuitable in the professional media production scenarios [4, 98] for which the method introduced in the current paper is intended.

6.3.1 The algorithm

The proposed method takes as input an image I, and creates an output image O with added texture emulating retinal noise. The transformations applied to the image are based on neurophysiological models of the visual system (a detailed description of these processes can be found at Section 4.2). The method can be summarized in the following stages:

1. Transform the input image I with a model of retinal processes, producing an intermediate image R that emulates the retinal output.

2. Add "retinal" noise to R to obtain a noisy image R_n.

3. Create the final output image O by applying to R_n the inverse of the previous transformations that emulate retinal processes.

The following transformations are applied separately in each RGB channel of the input image. In this section, the image I will represent each R, G, B channel of the input image in the range [0,1].

Given that our method is based on the emulation of retinal noise, we must start by ensuring that the input image I has values that are proportional to the intensity of light arriving at the retina. This is already the case if I is a RAW image, otherwise we assume that a nonlinear transform like gamma-correction has been applied to I with a standard exponent such as 1/2.2 [3] and we undo it, obtaining the linear image I_L:

$$I_L = I^{2.2}. \tag{6.1}$$

After this, the photoreceptor response $P(I_L)$ to the light stimulus I_L is emulated via the Naka-Rushton equation [75], yielding I_P:

$$I_P = P(I_L) = \frac{{I_L}^n}{{I_L}^n + {I_s}^n}, \tag{6.2}$$

where I_s is the semi-saturation constant and n controls the slope of the Naka-Rushton curve.

The lateral inhibition or center-surround organization of both bipolar cells and retinal ganglion cells is modelled as a convolution between the photoreceptor response and a kernel K similar to a DoG. The resulting image R, which will be our proxy for the clean retinal image, is obtained as follows :

$$R = K * I_P. \tag{6.3}$$

Motivated by the work of [97], we choose for K the following form:

$$K = \mathcal{F}^{-1}\left(\frac{1}{0.81 + 0.2\mathcal{F}(G_K)}\right) \tag{6.4}$$

where $\mathcal{F}$ is the Fourier transform and G_K is a 2D Gaussian kernel with standard deviation equal to 1/3 of the maximum of the image dimensions (height or width). A key advantage of this choice of kernel is that it is invertible, which will be very useful for us as we will see:

$$K^{-1} = \mathcal{F}^{-1}\left(0.81 + 0.2\mathcal{F}(G_K)\right). \tag{6.5}$$

We add to R a certain amount a of retinal noise n_r, that emulates the noise measured in the RGCs,

$$R_r = R + an_r, \tag{6.6}$$

and therefore the image R_r corresponds to the noisy image created in the retina.

For the noise signal n_r we use the same distribution as the noise observed in RGCs [58, 79], which has a constant standard deviation

(that does not depend on the input contrast), and we impose as well a bandpass frequency spectrum as approximately given by the contrast sensitivity function of the visual system [89]:

$$n_r = (G_c - G_s) * I_{\mathcal{N}}, \tag{6.7}$$

where $I_{\mathcal{N}}$ is a Gaussian noise image with standard deviation $\sigma = 1$, and G_c and G_s are 2D Gaussian kernels. As it is mentioned in [89], contrast sensitivity depends on the orientation, and this effect could be modeled with the 2×2 covariance matrices Σ_c and Σ_s of the kernels G_c and G_s. In practice, in our experiments we will use symmetric kernels, and in this case, the covariance matrices are not needed as the kernels can be described simply with the standard deviation of the Gaussians, σ_c for G_c and σ_s for G_s.

Recapping, the *noisy* retinal image R_r results from adding noise to

$$R = K * P(I^{2.2}), \tag{6.8}$$

so we can find the *noisy* light stimuli O that would directly produce R_r by undoing the previous chain of operations:

$$O = (P^{-1}(K^{-1} * R_r))^{\frac{1}{2.2}}. \tag{6.9}$$

The image O is the final output produced by our method (see Fig. 6.2), which as mentioned above is applied independently to each of the three color channels.

We would like to stress that all the operations performed by our algorithm are of low or very low computational complexity. The linearization, the Naka-Rushton transform, and their inverses can be encoded as 1D look-up tables (LUTs), so the method is essentially as fast as the time it takes to compute two convolutions, one with kernel K and the other with its inverse K^{-1}.

6.3.2 User parameters

The full list of parameters for our method, by order of appearance, is: $I_s, n, a, \sigma_c, \sigma_s$. Default values are proposed for these parameters so the

Figure 6.2. Image with added retinal grain on the left. Close-up of the image with retinal noise in the center, and close-up of the original image on the right.

method can be used as fully automatic. Moreover, these parameters can be modified by the user to control the visual aspect of the noise in the resulting image.

For the Naka-Rushton equation, we have fixed both its parameters: $n = 0.74$ following [22], and $I_s = 0.18$ given that 18% is the reflectance value for mid-gray, taking 100% as diffuse white. The intensity of the noise in the final output image O is controlled by the parameter a, which can vary in the range $[0, 1]$, and whose default value is set as $a = 0.015$. As a increases, the noise becomes more visible, as Fig. 6.3 shows. For the sizes of the Gaussians G_c and G_s used to generate n_r, we have chosen as default values $\sigma_c = 0.7$ and $\sigma_s = 1.5$. These parameters can be adjusted by the user allowing certain control in the size and distribution of the noise. Higher values of σ result in bigger size noise. The effect of using non-symmetrical Gaussian filters, defined by their 2×2 covariance matrices Σ_c and Σ_s, is a non-symmetrically distributed noise as it can be observed in Fig. 6.3 (d) and (f). Modifying these values alters the power spectrum of the noise, as it can be observed in Fig. 6.4.

6.3.3 Psychophysical evaluation

The goal of adding noise to images with a creative intent is to produce results that are visually appealing for observers. Therefore, we conduct psychophysical experiments in order to validate our results and

Figure 6.3. Resulting images with different parameter choices:
a) $\sigma_c = 1$, $\sigma_s = 2$, $a = 0.05$
b) $\sigma_c = 0.5$, $\sigma_s = 1$, $a = 0.05$
c) $\sigma_c = 0.5$, $\sigma_s = 1$, $a = 0.1$
d) Non-symmetrical kernels G_c and G_s, with covariance matrices $\Sigma_c = \begin{pmatrix} 0.2 & 0 \\ 0 & 0.05 \end{pmatrix}$ and $\Sigma_s = \begin{pmatrix} 1 & 0 \\ 0 & 0.25 \end{pmatrix}$, $a = 0.05$
e) $\sigma_c = 0.05$, $\sigma_s = 1$, $a = 0.05$
f) Non-symmetrical kernels G_c and G_s, with covariance matrices $\Sigma_c = \begin{pmatrix} 0.05 & 0 \\ 0 & 0.4 \end{pmatrix}$ and $\Sigma_s = \begin{pmatrix} 0.25 & 0 \\ 0 & 4 \end{pmatrix}$, $a = 0.05$

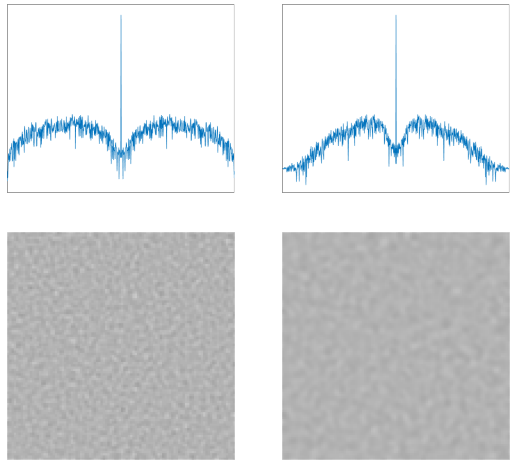

Figure 6.4. Proposed method applied to a flat grey image, with parameter values $\sigma_C = 0.7$, $\sigma_s = 1.5$ (left) and $\sigma_c = 1.2$, $\sigma_s = 2.6$ (right). Top: power spectrum. Bottom: resulting grain from our proposed method applied to a flat grey image.

to compare them with methods from the state of the art in academia, like the algorithm of Newson et al. in its implementation [54], and in the movie industry as well, like the film-grain emulation provided by the professional post-production software DaVinci Resolve 14.

For this study, the parameters of each method have been selected by a cinema expert to get the most appealing visual appearance of images according to his liking. For the method of Newson *et al.*, grain radius is set to $r = 1/200$, type of algorithm is pixel-wise and number of MonteCarlo iterations is set to $N = 1000$. For DaVinci ResolveFX texture film grain effect, the 35mm film settings have been used with an intensity of $I = 0.75$. For our method, the parameters used are $a = 0.015$, $\sigma_c = 0.7$ and $\sigma_s = 1.5$. Fig. 6.5 shows samples of four video frames, showing the four different versions used in the psychophysical experiments: a clean version, a version with retinal noise, a version with film-grain noise from the work by Newson *et al.*, and a version with film grain using the DaVinci ResolveFX texture effect. All the

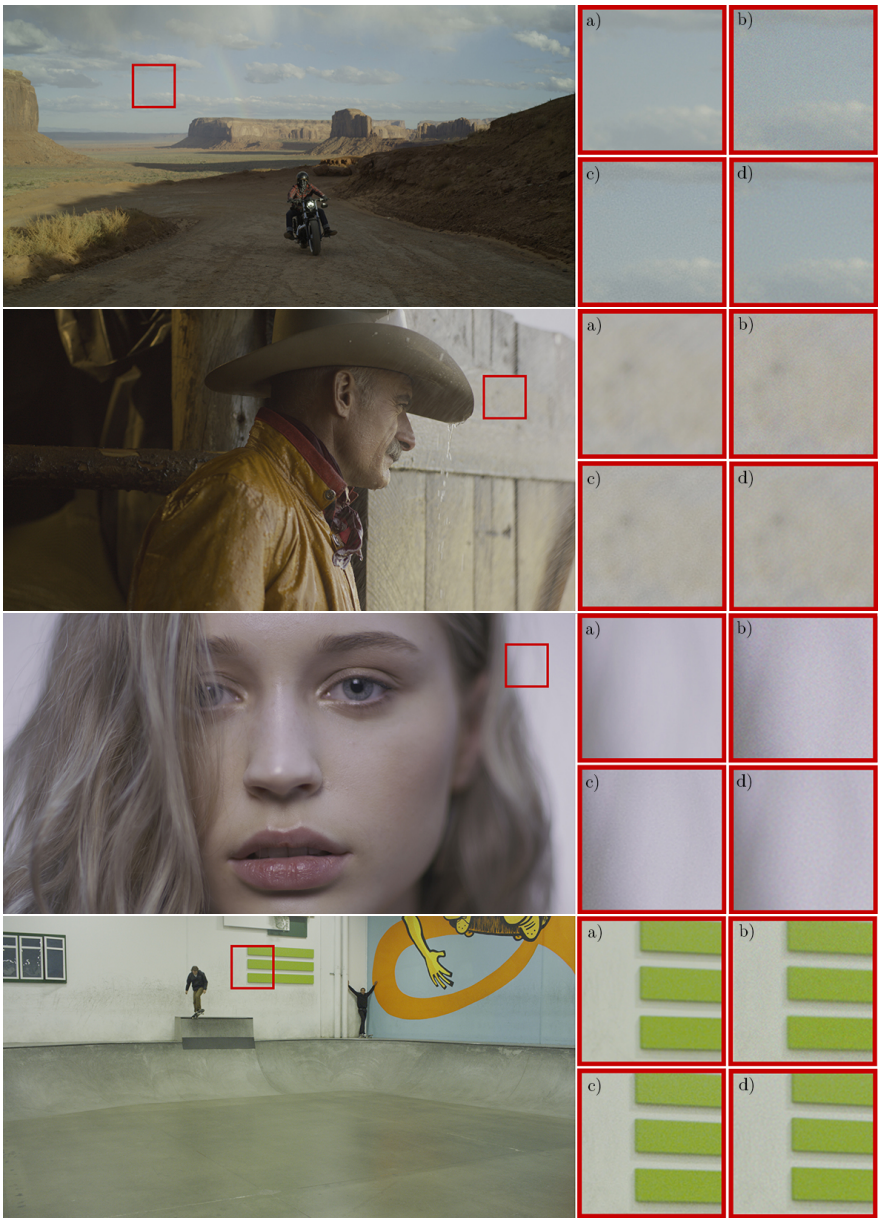

Figure 6.5. Left: frame without noise. Right: zoomed-in detail, a) original, b) film grain emulation by Newson et al. [54], c) film grain emulation by DaVinci Resolve 14, d) proposed retinal noise emulation.

methods are applied using the parameter values just specified. It is however worth noting that, in order to fully distinguish the difference between the methods, the temporal dimension is crucial.

For the evaluation, we used a room with dim ambient illuminance. Observers were instructed to sit approximately one meter away from the screen. A two-alternative forced-choice comparison (2AFC) technique was used: each observer was shown two consecutive videos on the screen, each of them obtained from one of the methods. The observers were asked to choose the most visually appealing video from the pair compared. Thirteen observers took part in the experiments, one of them being a cinema professional from a major postproduction house.

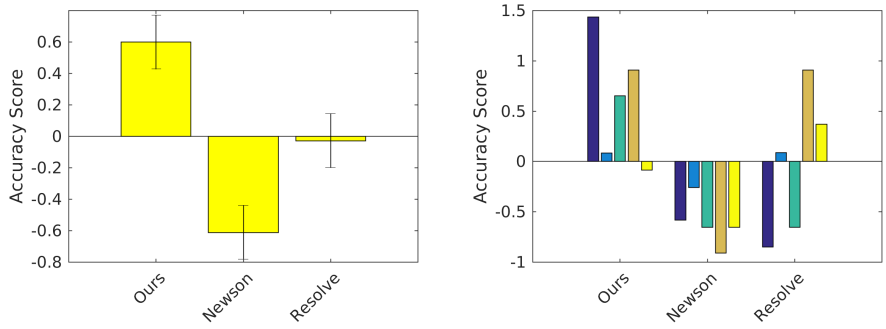

Figure 6.6. Accuracy scores of competing methods for adding texture: 13 observers took part in each experiment and 5 videos were used. Left: average. Right: scores per video.

The analysis of the psychophysical experiment is presented in Fig. 6.6. To compute accuracy scores from the raw psychophysical data, we use the same approach as in [50] (Chapter 5), which is based on Thurstone's law of comparative judgment.

As it can be seen in Fig. 6.6, our method is preferred over the other two. This result is also consistent for each video separately in all the cases except one. The individual preference of the movie professional that took part in the experiment follows the same trend

of the whole group of observers.

6.4 Retinal noise emulation for improving compressed video quality

We test here the suitability of retinal noise emulation to improve the quality of compressed video, or, more precisely, to mask the image degradation inherent to the compression process (i.e. compression artifacts) in order to prevent users from perceiving it. To evaluate this masking effect, we have designed a set of experiments that simulates a video on demand (VoD) service where the content is provided via HTTP/TCP-based adaptive bit rate streaming (ABR) techniques [2]. Under this paradigm, content is encoded at different quality levels associated with unequal bit rates in accordance to a given quality ladder and segmented. These segments are stored in a server or set of servers (Content Delivery Network (CDN)) and are provided to the client upon request [21]. The client selects throughout the streaming session the segments that best suit the system state (channel available bandwidth, terminal capabilities, quality control policies...) to optimize the quality provided to the user [2].

The results of the experiments will indicate down to what point, if retinal grain is added, service providers will be able to decrease the encoding bit rates included in the quality ladders, and therefore the objective quality of the encoded sequences, without the users noticing. To be able to remove the effect of coding and so isolate that of retinal noise, sequences with and without noise are used in the experiments.

All the procedures and selections related to the subjective tests fulfill the guidelines included in Recommendations ITU-R BT.500-13 [66], ITU-T P.910 [67] and ITU-T P.913 [68].

6.4.1 Test material

The test material is made up of four 10-second-long 4K (4096x2160p) Source sequences (SRCs) acquired at 24 fps. They all use a 4:2:2 chroma subsampling at 12 bits. The SRCs were selected from a public dataset made available by Blackmagic including representative, varied, and habitual contents for users [6]. The number, duration, and characteristics of the source content were selected in accordance to Rec. ITU-T P.913 [68]. Table 6.1 includes the characteristics of the SRCs in terms of the average spatial and temporal complexity considering the spatial information (SI) and the temporal information (TI) indicators [38, 67].

Table 6.1. SRCs' spatial and temporal complexity

	Balloon	Nature	Bugs	Closeup
SI	30.85	66.05	23.83	11.47
TI	28.64	44.08	43.21	43.21

Retinal noise was later on added to the 'clean' SRCs to obtain 'noisy' versions of them. The procedure is described next.

Addition of retinal noise

The proofing observer, a motion picture specialist, was seated two picture heights away from a Sony PVM-250 25" Full HD OLED display (Rec709, gamma 2.4 calibrated [3]). The display was driven by a Blackmagic 4K video card via DaVinci Resolve. A Tangent Element color correction panel was used as a control interface. The clips were shown in their native resolution (4096x2160) but were cropped to fit onto the Full HD display. The Tangent Element panel provided controls to shift the crop to different areas of the image. These controls could be operated with some small delay while the clip was playing.

First, proofing to select the proper noise shape parameters was conducted. Eight different noise shapes were generated for a one-second excerpt of a single video clip at three different intensity levels.

Each of these shape sets were placed on the Resolve timeline and viewed through the OLED display in a dark surround. The pan and tilt functions were used to determine the areas of the clip where the noise was most prominent and observations were recorded about each of the clips' appearances. It was found that several of the generated noise shapes added textures to the image which appeared gritty and blocky, while others introduced a finer grain that was more pleasing to the eye. The most ideal grain shape of those generated in the first round was selected to be $\sigma_c = 0.5, \sigma_s = 1$, which differs from the values obtained in the experiment discussed in Section 6.3.3 because now the tests are performed on a higher resolution monitor.

This noise was then added at varying intensity levels (0.025, 0.05, 0.075) to all of the tested clips (1 second excerpts). For this test, the clips were arranged on the Resolve timeline along with the "clean" version of the clip with no noise added. These different intensities were then compared in the dark surround viewing environment. In many cases, the clean versions themselves had a considerable amount of camera noise upon capture, which led to an unpleasant static on the clips. However, when some of the retinal noise was added, this camera noise was to some degree obscured and an overall more pleasing image was produced. In general, the lowest noise setting was selected for these images (0.025). In other cases of brighter images, it was found that a higher intensity of noise (0.05) seemed to improve the visual texture and appearance of the test clips, particularly in high-frequency areas.

Figure 6.7 shows screenshots of the four sequences, with the optimal amount (in terms of image appearance) of retinal grain added to them.

Generation of test sequences

We have tested 30 different combinations of encoding and network parameter values. Each of these combinations, called Hypothetical Reference Circuit (HRC) [67], is applied to all source sequences, re-

Figure 6.7. Screenshots of the four sequences used in the subjective assessment, with optimal amount (in terms of image appearance) of retinal grain added. Each image shows a zoomed-in region (marked with a red square) in two versions, with (top) and without (bottom) the retinal grain. From left to right, top to bottom: "Balloon", "Nature", "Bugs", and "Closeup".

sulting in a set of Processed Video Sequences (PVSs), one per SRC and HRC, that are presented to the users for evaluation. The combinations considered in the tests are included in Tables 6.2 and 6.3. HRCs are 4:2:0 and have a color depth of 10 bits to match common broadcast conditions. Furthermore, they preserve the framerate of the sources: 24 fps. Five of them (named $\mathrm{HRC}^{\mathrm{C}}_{i,1}$, where $i = \{1, \ldots, 5\}$) are anchor points derived directly from the quality ladders included in Apple' HLS technical note [1]. To decrease the gap between consecutive quality levels and so enable a finer-grained analysis, two additional HRCs (named $\mathrm{HRC}^{\mathrm{C}}_{i,j}$, where $j = \{2, 3\}$) were created from each anchor point $\mathrm{HRC}^{\mathrm{C}}_{i,1}$. The spatial resolutions of these 15 HRCs, marked with the superscript "C" for clean, nearly

Table 6.2. HRCs used to create the test sequences presented to the observers

HRC	Resolution	bit rate (Mbps)	Retinal Noise
$HRC^{C}_{1,1}$	2160p	22.2	No
$HRC^{N}_{1,1}$	2160p	22.2	Yes
$HRC^{C}_{1,2}$	2160p	18.4	No
$HRC^{N}_{1,2}$	2160p	18.4	Yes
$HRC^{C}_{1,3}$	2160p	14.6	No
$HRC^{N}_{1,3}$	2160p	14.6	Yes
$HRC^{C}_{2,1}$	1440p	10.7	No
$HRC^{N}_{2,1}$	1440p	10.7	Yes
$HRC^{C}_{2,2}$	1440p	9.7	No
$HRC^{N}_{2,2}$	1440p	9.7	Yes
$HRC^{C}_{2,3}$	1440p	8.7	No
$HRC^{N}_{2,3}$	1440p	8.7	Yes
$HRC^{C}_{3,1}$	1080p	7.8	No
$HRC^{N}_{3,1}$	1080p	7.8	Yes
$HRC^{C}_{3,2}$	1080p	6.7	No
$HRC^{N}_{3,2}$	1080p	6.7	Yes
$HRC^{C}_{3,3}$	1080p	5.6	No
$HRC^{N}_{3,3}$	1080p	5.6	Yes

follow a geometric progression with ratio $1/\sqrt{2}$. So, each picture resolution is close to half the previous one. The other half of the set of HRCs, marked with the superscript "N" for noise, share the same characteristics as the original set of HRCs and, in addition, they include retinal noise.

All HRCs are H.264/AVC encoded. Bit rates for the anchor HRCs were obtained from the H.264/AVC ladder whenever the associated resolution was included there and extrapolated accordingly to the ladder rule for the remaining resolutions. The bit rates for the rest of

Table 6.3. HRCs used to create the test sequences presented to the observers

HRC	Resolution	bit rate (Mbps)	Retinal Noise
$HRC^{C}_{4,1}$	720p	4.5	No
$HRC^{N}_{4,1}$	720p	4.5	Yes
$HRC^{C}_{4,2}$	720p	3.7	No
$HRC^{N}_{4,2}$	720p	3.7	Yes
$HRC^{C}_{4,3}$	720p	2.9	No
$HRC^{N}_{4,3}$	720p	2.9	Yes
$HRC^{C}_{5,1}$	540p	2.0	No
$HRC^{N}_{5,1}$	540p	2.0	Yes
$HRC^{C}_{5,2}$	540p	1.7	No
$HRC^{N}_{5,2}$	540p	1.7	Yes
$HRC^{C}_{5,3}$	540p	1.4	No
$HRC^{N}_{5,3}$	540p	1.4	Yes

the HRCs were obtained by linearly interpolating between the values of the anchor HRCs.

Finally, we generated a total of 120 Processed Video Sequences (PVS's). As mentioned before, the aspect ratios of the HRCs, and therefore of the PVS's, were 16:9 (∼1.78:1), following the SMPTE ST 2036-1 standard [78] and Recommendation ITU-R BT.2020 [65]. As the aspect ratio of the SRCs is 256:135 (∼1.9:1), according to the Digital Cinema System Specification of the Digital Cinema Initiatives (DCI) [16], the PVS's resolutions required a minor cinema-to-broadcast format adaptation that was conducted using a bicubic filter [10].

6.4.2 Environment and equipment

The test room was set to simulate home viewing conditions. Furthermore, the brightness was controlled according to recommended

values [67]: 24.4 Lx in front of the subjects, 16.5 Lx to their left, 85.4 Lx to their right, 71.7 Lx above them, and 20.1 Lx behind them.

The device used in the tests was a TV set with a 43-inch screen and a 3840x2160 pixel resolution (Samsung UE43NU7475). The viewing distance was set to twice the height of the screen, in accordance to Rec. ITU-T P.913 [68].

6.4.3 Methodology

Before starting the experiments, the test designer read the guidelines of the tests to the observers. Next, subjects were trained by showing them examples of the best and worse quality levels they should expect for sequences with and without retinal noise (four extra PVS's created from a fifth content according to $\mathrm{HRC}^{\mathrm{C}}_{1,1}$, $\mathrm{HRC}^{\mathrm{N}}_{1,1}$, $\mathrm{HRC}^{\mathrm{C}}_{5,3}$ and $\mathrm{HRC}^{\mathrm{N}}_{5,3}$). In this way, subjects were more aware of the scale of qualities that they would encounter and rate the sequences accordingly.

During the experiments, all the PVS's, that is, every combination of video sources -SRC's- and encoding and network conditions to be tested -HRC's-, including the reference sequence, were sequentially and randomly presented to the subjects. Each PVS was presented once to each subject. The order of presentation of the PVS's was different for each pair of observers and was set randomly, in accordance with Rec. ITU-T P.910 [67]. The whole session was slightly shorter than 30 minutes, as recommended by ITU-R BT.500-13. The test method followed in the tests is the Absolute Category Rating with Hidden Reference (ACR-HR) proposed in Rec. ITU-T P.910 [67], where subjects have five possible answers to choose from: "Excellent", "Good", "Fair", "Poor", and "Bad". The subjects were asked to assess each PVS right after its visualization. To help it, a four-second grey sequence was included between consecutive PVS's, also as stated in Rec. ITU-T P.910 [67].

There were 18 observers (6 women and 12 men) in the experiment, all of them having normal or corrected vision, aged between 20 and

30 years. The number of subjects is sufficiently significant, as stated in Rec. ITU-R BT.910 [67]. The observers were rewarded for their participation in the tests, and a maximum of two observers were allowed in each test session. Due to the characteristics of the play-out system, the assessment was conducted sequentially on the two demi-sets of PVS's: first the PVS's including retinal noise (named $\mathrm{HRC}_{i,j}^{\mathrm{N}}$), called 'noisy' PVS's, and then the 'clean' PVS's (named $\mathrm{HRC}_{i,j}^{\mathrm{C}}$). No observers were rejected after the screening of the subjective results.

6.4.4 Test Results

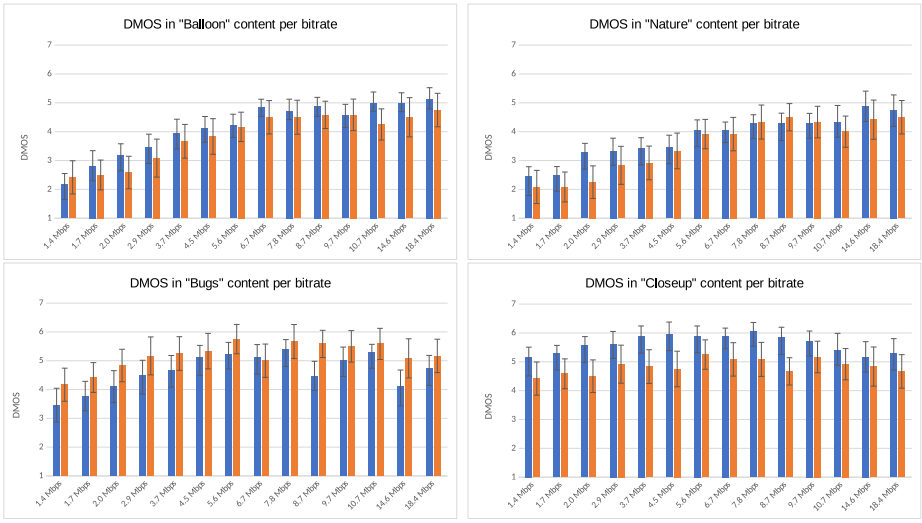

Figure 6.8. DMOS per content and bit rate. From left to right, top to bottom: "Balloon", "Nature", "Bugs", and "Closeup". Orange bars are used for 'clean' content and blue bars indicate content with retinal grain.

Figure 6.8 depicts the results in terms of the evolution per content of the differential mean opinion score (DMOS) versus the encoding

bit rate. The DMOS is defined as follows:

$$\mathrm{DMOS(PVS)} = \mathrm{MOS(PVS)} - \mathrm{MOS(REF)} + 5 \tag{6.10}$$

where MOS is the Mean Opinion Score computed for a given content (PVS or reference sequence). Therefore, the better the image quality of the sequence presented to the user (i.e. the more it looks like the reference one), the greater it will be the MOS of that sequence, and so the resulting DMOS.

The DMOS values have been computed per user for each one of the 'clean' and 'noisy' sets of PVS's with respect to the scores given to their corresponding references, $\mathrm{HRC}^{\mathrm{C}}_{1,1}$ and $\mathrm{HRC}^{\mathrm{N}}_{1,1}$, as it is usually done in the literature. Each bar includes its 95% confidence intervals.

We can easily distinguish two trends in the figures per content: that of sequences "Balloon" and "Nature" and that of sequences "Bugs" and "Closeup". The results of the assessment of the first two sequences show a clear and steady decrease in the quality perceived by observers with the reduction of bit rate on average. However, the results for the other two sequences do not show any clear connection between the perceived quality and the bit rate. This is an interesting outcome of this exploratory analysis on the addition of retinal noise, as this discrepancy stems from the different nature of the video contents. On the one hand, all the elements in every picture in the sequences "Balloon" and "Nature" are in focus. On the other hand, a significant part of every picture in the sequences "Bugs" and "Closeup" is out of focus due to the limited depth of field used in their acquisition. As this experiment has been carried out on visual information in a bit rate limited scenario, it means that the last two sequences were always better treated by the compression and decompression system, as more bits could be devoted for the encoding of the in-focus part of the picture. Therefore, the selection of test material in subsequent experiments should consider the depth of field information in addition to their spatial and temporal information. Moreover, it is important to note the type of content presented. Since the sequence "Bugs" did not show any benefit of being treated

with noise, there is a possibility that documentary content may be less susceptible to being processed by adding noise. Thus, utmost care should be placed on content selection where texture noise could be added.

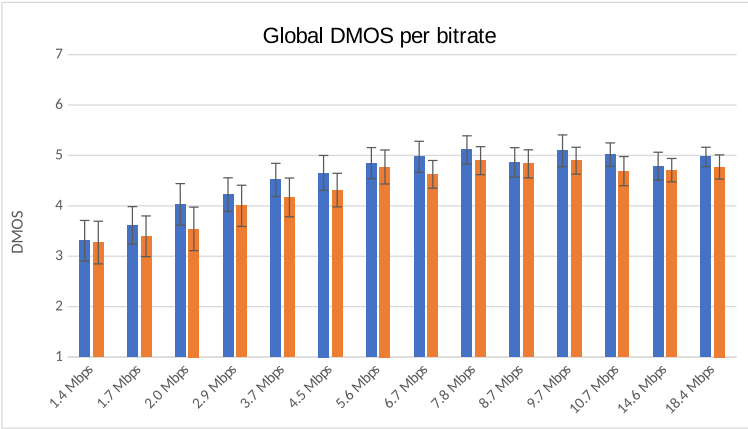

Figure 6.9. Global DMOS per bit rate. Orange bars are used for "clean" content and blue bars indicate content with retinal grain.

The analysis of the aggregated DMOS, presented in Figure 6.9, shows a reduced but significant advantage on the addition of retinal noise in the outcome of the assessment. Even if the confidence intervals overlap, the DMOS of each "noisy" PVS is always higher than the one of its "clean" counterpart, the advantage is more evident for resolutions up to high definition. We consider the results significant since there is a clear tendency for the results with noise to be better rated on average. From these results, we conclude that, when the bit rate is reduced, the decrease in perceived image quality is smaller if the video has had retinal grain added to it.

In order to measure properly the benefits of applying our proposal to mask compression artifacts in terms of bit rate saving and quality improvement, we have applied a regression on the noise and clean sets of points. The regression has been performed using a sigmoid function [83] and the least squares method. Figure 6.10 shows the

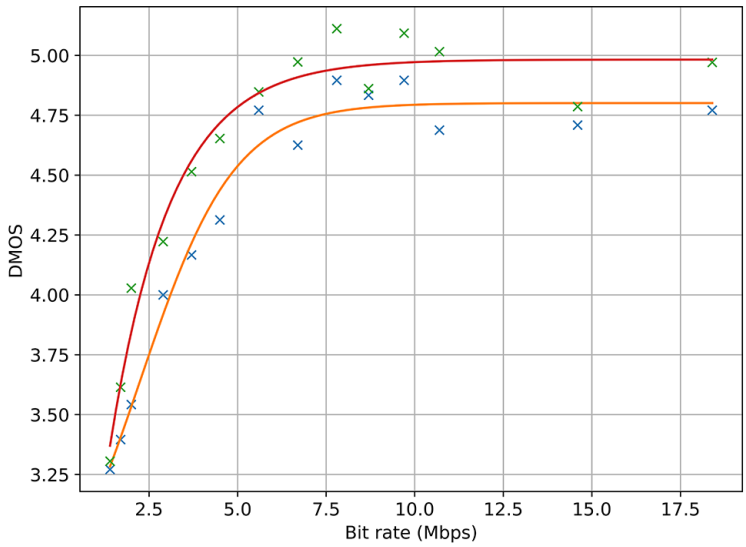

Figure 6.10. Regression on the bit rate-DMOS values. Green crosses represent "noise" scores, blue crosses are "clean" scores, the red line represents the regression on "noise" values and the orange line regression on "clean" values.

result of the process. The green and blue crosses represent the average user scores for noisy and clean contents, respectively. The red and yellow lines are their respective regression.

First, one can verify the conclusions drawn above: the addition of noise represents an enhancement of quality over the clean signal for all considered bit rates. Moreover, this result can also be seen from another point of view: for a given quality level, using the version that includes the emulation of retinal noise leads to significant savings in bandwidth. Both gains have been analyzed quantitatively. So, the former has been measured in terms of BD-DMOS by computing the area between the lines (represented in purple in Figure 6.11) by means of a vertical integration [25, 71]. Results point at a DMOS average improvement of 0.2. Regarding bit rate savings, they have been measured in terms of BD-Rate by computing the area between

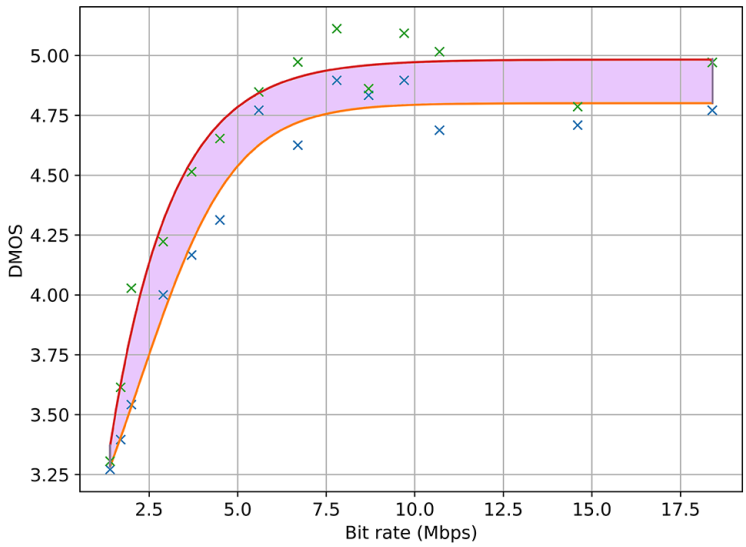

Figure 6.11. Regression on the bit rate-DMOS values including the area between the lines where the BD-DMOS is computed by vertical integration. This plot highlights the fact that, at any given bit rate, the addition of emulated retinal noise improves perceived image quality (the DMOS value is higher for the sequence with retinal noise).

the lines (represented in purple in Figure 6.12) through a horizontal integration [25, 71]. Results indicate that the application of the retinal noise emulation method allows for a significant improvement in coding efficiency, with average bit rate savings of over 22.5%. Nevertheless, let us point out that our experiments suggest a potential weakness of our approach in that the applicability of the method may depend on the specific video sequence that is dealt with, because as remarked earlier there is the possibility that for some kind of content, like documentary footage, a cinema-like appearance is not preferable for the viewer. However, a mere previous analysis of the content could determine the suitability of the inclusion of retinal noise.

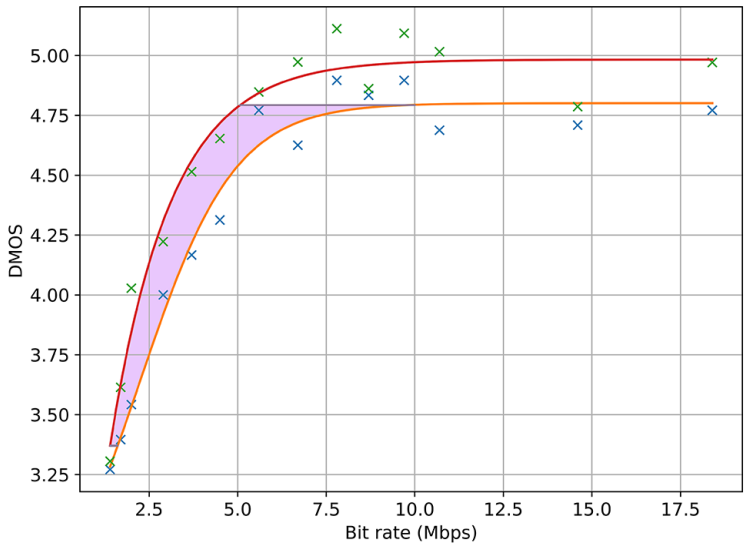

Figure 6.12. Regression on the bit rate-DMOS values including the area between the lines where the BD-Rate is computed by horizontal integration. This plot highlights the fact that, for any given perceived quality level (DMOS value), the encoding of the sequence with emulated retinal noise is more efficient (the required bit rate is lower than the one necessary to attain the same DMOS value with the clean sequence).

6.5 Conclusions and future work

We have presented a method for adding texture to digital cinema that is inspired by processes in the visual system and produces results that look natural and visually pleasing even for challenging scenes. The computational complexity of the method is very low, appearing suitable for a real-time implementation in order to be used on the set. The method has three parameters whose default values produce satisfactory results in a variety of scenarios, and that can be modified for artistic purposes, in order to achieve different looks. A psychophysical validation was conducted, showing that the proposed method

outperforms algorithms from the state of the art in the academic literature and in the industry.

The retinal noise emulation method can also improve the quality of compressed video by masking compression artifacts. The aim was to help concealing distortions due to compression and thus allowing to maintain image quality while reducing the bit rate or improving image quality while maintaining the bit rate fixed. The proposed method has been validated through subjective assessment on 4K professional cinema sequences, where the amount of retinal noise was selected by a motion picture specialist based solely on aesthetic preference. The experiment has shown that the proposed method can yield very impressive savings in bit rate. A special effort has been made to maintain the rigorousness and reproducibility of the subjective tests carried out. As future work, we intend to explore the impact of the type of content in the usefulness of the addition of retinal grain noise to mask compression artifacts.

Our results point to a novel and, we believe, very promising avenue of research in computer vision which is the connection between vision models of retinal grain, perceived image quality (a very active area of interest in computer vision because the main challenges remain unsolved), image compression algorithms and image compression as performed by the visual system: the connection here is even more explicit since the classic work of Olshausen and Field [57], that allows to link convolutional neural netowrks (CNNs) trained for compression with the receptive fields that are actually measured in the human visual system. Ongoing work involves training CNNs for compression on natural images with and without retinal grain.

6.6 Supplementary material

The videos used for the experiments and some supplementary images can be found at: `http://ip4ec.upf.edu/RetinalNoise`

7 Perceived image quality

This chapter starts by emphasizing the importance of accurate quality assessment and describing some widely-used image quality metrics. Then, we explain in detail a physiology-based quality metric, INRF-IQ, that takes into account various perception phenomena present in the HVS. Finally, we propose different strategies to optimize the INRF-IQ model parameters, so the resulting metric is competitive with state-of-the-art methods. The chapter is ended with a description of the experiments performed to test the proposed metric optimization. This work has been done in collaboration with Adrián Martín, from Universitat Pompeu Fabra.

This work, and more specifically the Pytorch implementation of the optimization method, and the Matlab implementation of the INRF transformation, has been developed in collaboration with Adrián Martín, from Universitat Pompeu Fabra.

7.1 Motivation

Quality evaluation is of crucial importance in the image and video processing field. It has numerous practical applications, and it also plays an important role in the development, optimization, and test-

ing of algorithms. Many tasks in image processing require validation by comparing the result with the original data, e.g. image denoising, image deblurring, etc. Subjective evaluation, consisting in measuring image quality by human beings, is costly and time-consuming. Therefore, the goal of objective quality assessment is to develop quantitative measures to automatically predict perceived quality in a way that is consistent with subjective human evaluation.

In the context of cinema and broadcasting, a quality metric can be useful for many applications: as an automatic video streaming quality measure, for video coding, image compression, etc.

7.2 Related work on image quality estimators

Image quality (IQ) methods can be divided into three categories, depending on the amount of information available about the original reference image: *full-reference* methods, for which an original reference image is available, make a comparison between the distorted and the reference image; *reduced-reference* methods compare some characteristics of the distorted and reference image since the complete reference image is not available; and *no-reference* methods (also called blind models) operate solely on the distorted image.

The vast majority of image quality approaches consist in full-reference methods and in this chapter we will focus on this type of methods. A simple solution, and most widely used metric to estimate image quality is the peak signal-to-noise ratio (PSNR), based on the mean square error (MSE), which at the same time is also a very popular metric. These methods are simple to calculate, and they have a clear physical meaning, however, they are not very well correlated with perceived visual quality (see Fig. 7.1).

Therefore, in the last decades, the goal of IQ research has been to improve these metrics and develop more sophisticated methods, using models that mimic the early stages of the visual system, or sometimes

Figure 7.1. Original and 11 distorted images, from the CSIQ database. All distorted images have a PSNR of 22.5 dB, however, there exists a large variation in perceived quality between images.

applying information theory models. We will describe a selection of them more in detail. Full-reference methods can be divided into:

- The *model-based methods* incorporate different models (e.g. HVS models, information theory models) to address the quality perception problem. Some methods of this type are:
  - The Normalized Laplacian Pyramid Distance (NLPD) [39] is based on transformations present in the early visual system: local luminance subtraction and local gain control, obtained from a decomposition of images using a Laplacian pyramid. The quality of a distorted image, relative to its original reference image, is the root mean square error in this "normalized Laplacian" domain.
  - The Structural Similarity Index (SSIM) [86] is based on the hypothesis that the HVS is highly adapted for extracting structural information from the viewing field. The structural similarity measurement is obtained from three comparisons: luminance, contrast, and structure, which are considered relatively independent components. The comparison is made into local patterns of pixel intensities that have been normalized for luminance and contrast.
  - The Feature Similarity Index (FSIM) [100] is based on the assumption that HVS understands an image according to its low-level features. The primary feature employed by FSIM is the phase congruency, which measures the significance of a local structure, but it is contrast invariant. Therefore, the secondary feature is the image gradient magnitude, which encodes contrast information. These features are complementary since they reflect different aspects of the HVS.
  - The Visual Information Fidelity Measure (VIF) [76] measures the mutual information between the perceived reference and distorted images. It is based on natural scene

statistics (NSS), and models of the image degradation process and the human visual system (HVS).

- The *learning-based methods* use supervised machine learning methods to learn a metric from a set of training images and their corresponding distances. Some methods of this type are:

  - The Learned Perceptual Image Patch Similarity (LPIPS) Metric [101] is based on the hypothesis that perceptual similarity is a consequence of visual representations, and not a special function all of its own. They found that internal activations of networks trained on high-level image classification tasks, do correspond well to human perceptual judgments, outperforming widely-used metrics such as PSNR, SSIM [86], or FSIM [100].
  - The Deep Image Structure and Texture Similarity (DISTS) Metric [17] uses a variant of the VGG convolutional neural network to construct a function that combines structure and texture similarity measurements between corresponding feature maps of the reference and distorted images. The parameters of the network are optimized to match human ratings. This measure is explicitly designed to be robust to texture resampling and modest geometric transformations, and it correlates well with human perceptual scores, both on conventional image quality databases, as well as on texture databases.
  - PerceptNet [32] is a convolutional neural network where the architecture reflects the structure and various stages in HVS: a cascade of canonical linear filters + divisive normalization layers simulate the retina-LGN-V1 cortex pathway. This domain replicates the image representation at the end of the primary visual cortex (V1). Then, the network is trained to maximize the Pearson correlation between the mean opinion score (MOS) and the l_2-distance

of the reference and distorted images in this perceptual domain. The TID2008 dataset is used for training the model. The performance of PerceptNet is similar to other neural networks, but the number of parameters is several orders of magnitude less.

7.3 INRF as an image quality metric

This section summarizes some definitions and results that can be found in [5]. The INRF quality metric, named INRF-IQ, is based on physiology and HVS perception knowledge, and it assumes a non-linear response of neurons, explained below.

In the vision science literature, most vision models assume a linear response of the receptive field (the portion of the visual field where light can evoke a sensory neural response) of neurons. However, this assumption conflicts with some HVS properties:

- According to the visual adaptation phenomena, the spatial receptive field properties of neurons are modified depending on the input. Therefore, the receptive field cannot be a fixed, constant property of a neuron.

- The visual system is nonlinear, therefore it has no basis functions, while a linear receptive field presupposes a set of basis functions for the visual system.

- A linear receptive field followed by an output nonlinearity model is questioned by recent works in vision science, that describe neuron response as being highly nonlinear.

The *intrinsically nonlinear receptive field* (INRF) is a model for a single-neuron receptive field response, whose general formula is [5]:

$$\mathrm{INRF}(x) = \sum_i m_i I(y_i) - \lambda \sum_i w_i \sigma[I(y_i) - \sum_j g(y_j - x) I(y_j)] \quad (7.1)$$

where m_i stands for $m(x, y_i)$ and w_i for $w(x, y_i)$, and σ represents a non-linearity.

The model is based on some knowledge about dendritic cells: Some dendritic branches act as non-linear units, while a single non-linearity σ is not enough to model dendritic computations. Moreover, there is feedback from the neuron soma to the dendrites. In the INRF model some dendrites are linear and their contributions are summed with weights m_i, and some other dendrites are nonlinear and their contributions are summed with weights w_i. The feedback from the soma is reflected in the shifting term of the non-linearity σ, expressed by the term $\sum_j g(y_j - x)I(y_j)$. The final neural output can be obtained by applying a non-linearity (such as rectification, divisive normalization, etc.) to the INRF values.

Based on this neural response model, a perceptual metric is produced, named INRF-IQ. For this metric, the transformations m, w, and g in Eq. 7.1 are Gaussian kernels with standard deviations σ_μ, σ_w, and σ_g, respectively. The metric is obtained as follows: given an image I, and its distorted version I_D, the INRF transformation is applied to both of the images, obtaining O and O_D, and then, the root mean square error between the processed images is computed:

$$\text{dist}(I, I_D) = \sqrt{\text{MSE}(O, O_D)} \tag{7.2}$$

This metric has five parameters: the standard deviation σ_μ of the Gaussian kernel m, the standard deviation σ_w of the Gaussian kernel w, the standard deviation σ_g of the Gaussian kernel g, and the weighting parameter λ. Using the optimized parameters for a brightness perception experiment ($\sigma_\mu = 10$, $\sigma_w = 30$, $\lambda = 5$, and the non-linearity σ being $\sigma(z) = z^{0,5}$ when $z \geq 0$ and $\sigma(z) = z^{0,7}$ when $z < 0$, as explained in [5]), the metric INRF-IQ reaches a Pearson correlation value with the MOS of 74% on the database TID2013, which is comparable to a state-of-the-art deep learning perceptual metric LPIPS [101] with a correlation of 76%, see Table 7.1.

In the following sections, we will explain how the parameter values are optimized in order to maximize the Pearson correlation between

	TID2013
PSNR	0,570
SSIM	0,650
LPIPS	0,760
INRF(Ours)	0,740

Table 7.1. Numbers indicate Pearson correlation with Mean Opinion Scores (MOS) in TID2013 database for different image quality metrics: PSNR, SSIM [86], LPIPS [101], INRF-IQ. The parameters used for the INRF transformation are the optimized parameters for brightness perception. Adapted table from [5].

the INRF distance and the Mean Opinion Score in the TID2008 [60] image quality database.

7.4 Experimental results

For the optimization and validation processes of the proposed metric INRF-IQ, different databases have been used. A detailed description of them can be found in Appendix I.

The parameters of the INRF quality metric have been optimized using the database TID2008 [60], to achieve the highest correlation between the mean opinion scores (MOS) of observers and the INRF perceptual distance following Eq. 7.1 and 7.2. Different methods have been used for optimizing the model parameters: interior point method (implemented in the *fmincon* Matlab function), grid search, genetic algorithm, and a neural network optimization approach, as it is explained more in detail in Section 7.5.

The INRF quality metric is implemented as follows: in a preprocessing stage, the image signal is converted into a more appropriate color space for the HVS. The CIELAB space, intended as a perceptual color space, will be used, so the INRF transformation will be applied to the luminance channel of this color space. The INRF

metric has been extended to color (explained in Section 7.7), however, the obtained correlation between the mean opinion score and the INRF distance is lower compared to the standard implementation correlation. Even though in certain scenarios (INRF model with fixed parameters, see [5]) a cascade of INRF transformations can lead to better results with respect to a single INRF transformation, in our proposed approach a single INRF outperforms the stack implementation. Different non-linearities have been tested for the function σ: a piece-wise exponential function, a hyperbolic tangent, and an arctangent σ non-linearity, which produced the best correlation.

Therefore, the final INRF transformation consists of a single INRF transformation applied to the luminance channel in the CIELAB color space, using an arctangent non-linearity σ, and with parameters $\sigma_\mu = 1.74$, $\sigma_w = 25$, $\sigma_g = 1$, and λ=3 (obtained with a grid search optimization approach to achieve the highest correlation between the INRF distance and the MOS in TID2008 database).

Using this transformation and calculating the INRF-IQ distance as indicated in Eq. 7.2, the following Spearman rank correlation coefficient (SRCC) values are obtained (see Table 7.2):

We compared the INRF metric against a set of full-reference image quality methods, including eight knowledge-driven models and three data-driven CNN-based models. Results, reported in Table 7.2, show that INRF performs favorably in comparison to some CNN-based models (e.g. LPIPS [101], and DISTS [17]), and some widely-used classic methods (NLPD [39]). Overall, the best performances across all three databases are obtained with PerceptNet [32], GMSD [93], and our proposed INRF metric.

As it is explained in [18], image quality metrics are usually tested by computing their agreement with standardized image quality datasets (i.e. TID2008, TID2013, LIVE, or CSIQ), consisting of artificially distorted images. It is important to test image quality metrics in other databases, to avoid over-fitting to the standard types of distorsions. Therefore, we have tested our method in a series of image generation/restoration databases, and we have compared the perfor-

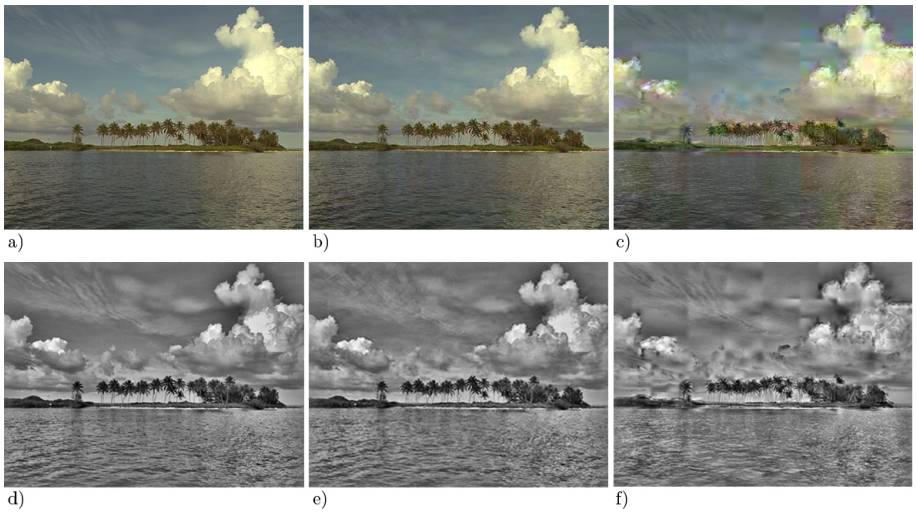

Figure 7.2. Visualization of INRF transformation. a) Original image, b) distorted image with comfort noise level 3, c) distorted image with comfort noise level 5, d), e), and f) INRF transformations of the corresponding above images (scaled for visualization).

mance of INRF-IQ with other metrics on these databases (details in Section 7.6).

7.5 Optimization process details

The optimization of the model parameters has been done using different approaches, detailed below. The list of parameters to be optimized is: the standard deviation σ_μ of the Gaussian kernel m, the standard deviation σ_w of the Gaussian kernel w, the standard deviation σ_g of the Gaussian kernel g, and the weighting parameter λ. Overall, the highest correlation has been obtained using a grid search optimization approach.

	LIVE	CSIQ	TID2013	(mean)
MS-SSIM	0,951	0,886	0,782	(0,873)
CW-SSIM	0,781	0,738	0,680	(0,733)
VIF	**0,963**	0,911	0,676	(0,850)
NLPD	0,938	0,937	0,800	(0,892)
GMSD	0,960	**0,950**	**0,804**	**(0,905)**
MAD	0,960	**0,941**	0,773	(0,891)
FSIM	**0,963**	0,916	**0,802**	(0,894)
VSI	0,950	0,923	0,793	(0,889)
LPIPS	0,932	0,837	0,616	(0,795)
DISTS	0,942	0,905	0,764	(0,870)
PerceptNet	**0,98**	**0,96**	**0,87**	**(0,93)**
INRF(Ours)	0,947	0,952	**0,802**	**(0,900)**

Table 7.2. Numbers indicate Spearman rank correlation coefficients (SRCC). The INRF metric is compared against a set of full-reference image quality methods: MS-SSIM [87], CW-SSIM [88], VIF [76], NLPD [39], GMSD [93], MAD [40], FSIM [100], VSI [99], LPIPS [101], DISTS [17], and PerceptNet [32]. Adapted table from [18].

7.5.1 Interior point method

First, the optimization has been done using the *fmincon* function in Matlab. This function uses the interior point method to solve the optimization problem. The initial values for the parameters to be optimized have been selected randomly. Different parameter initializations have been tried.

7.5.2 Grid search

The search space is defined as a grid of parameter values, and every position in the grid is evaluated. The optimal parameters are those which generate the highest correlation. The grid values are: $\{1.74\}$ for σ_μ, $\{1.5, 3, 4.5\}$ for λ, $\{10, 50, 150\}$ for σ_w, and $\{1\}$ for σ_g.

7.5.3 Pytorch implementation

We code a class INRF as `torch.nn.module` to define a specific instance of the model to be trained. The forward method of the class receives an image as input and outputs the result from applying Eq. 7.1 on this image. In order to do so, Gaussian kernels as a function of the corresponding standard deviations are generated before being convolved with the data. The main difference between a "standard neural network" designed as a `torch.nn.module` arises from the fact that we do not learn the kernel weights individually but we just learn the standard deviation of each Gaussian kernel, therefore dramatically reducing the number of parameters. The INRF module is trained to maximize the Pearson correlation between the MOS and the mean square error between the reference and the distorted images after applying the INRF transformation. For this purpose, we build a specific data-loader in which the data are 3-tuples that contain a reference and a distorted image and their associated MOS. The training is performed in batches of 25 tuples from which a single value of training loss is obtained. The INRF module implements a one-step pass of Eq. 7.1. For adding more "layers", a `torch.nn.Sequential` module can be used to concatenate an arbitrary number of INRF modules, so the input of each layer is the output from the previous one after the application of INRF.

7.6 Experimental results in not-standardized image quality databases

As it can be observed in Table 7.3, the proposed metric shows a limited performance when generalizing to alternative datasets containing other types of distortions: denoising, rendering, texture synthesis, etc. For a further description of these databases see Appendix I.

	Denoising FLT	Deblurring Liu13	Super-res. Ma17	Dehazing SHRQ	Rendering Tian19	Texture synt. SynTEX
PSNR	0,183	0,803	0,592	**0,74**	**0,536**	0,114
SSIM	0,355	0,777	0,624	0,692	0,23	0,62
MS-SSIM	0,246	0,898	0,795	0,687	0,396	0,469
VIF	0,169	0,864	0,831	0,667	0,259	0,448
CW-SSIM	0,101	0,742	0,706	0,698	0,522	0,496
FSIM	0,182	0,897	**0,864**	0,605	**0,622**	0,134
GMSD	**0,555**	**0,921**	0,747	0,695	0,476	0,093
VSI	0,389	0,918	**0,851**	0,663	0,479	0,006
NLPD	0,528	**0,92**	0,71	0,696	0,531	0,123
PieAPP	**0,629**	0,786	0,771	0,725	0,298	**0,709**
LPIPS	0,457	0,867	0,788	**0,777**	0,311	**0,663**
DISTS	**0,636**	**0,941**	**0,878**	**0,789**	**0,671**	**0,92**
INRF(Ours)	0,1526	0,8681	0,7365	0,401	0,3772	0,4429

Table 7.3. Numbers indicate Spearman rank correlation coefficients (SRCC). Comparison of INRF-IQ to other image quality metrics: PSNR, SSIM [86], MS-SSIM [87], VIF [76], CW-SSIM [88], FSIM [100], GMSD [93], VSI [99], NLPD [39], PieAPP [62], LPIPS [101], and DISTS [17] in not-standardized image quality databases. Adapted table from [18].

7.7 Extension to color image quality assessment

Although the INRF-IQ metric is designed to be applied in the luminance component of color images, the chrominance information can be incorporated by means of a simple extension of INRF-IQ, and we call this extension INRF_C-IQ. Since many distortions affect the color information of images, better performance is expected if the chrominance information is incorporated in INRF for color image quality assessment. Two alternative extensions of the model have been implemented and tested in the TID2008 database:

- One INRF_C-IQ implementation consists in applying the INRF transformation to each *RGB* channel separately (with differ-

ent parameters for each channel), and then calculate the MSE between the transformed reference and distorted images. The optimization of the parameters is done iteratively: first, the optimal parameters are found for the R-channel, then the optimal G-channel parameters are found (using the previously found optimal R-channel parameters), and finally the optimal B-channel parameters are found (using the previously found optimal R and G-channel parameters).

- An alternative implementation of INRF$_C$-IQ is to apply the INRF transformation to each RGB channel, using the optimal parameters obtained in Section 7.4. Then, the INRF-IQ distance is calculated following the formula in Eq. 7.2 for each RGB channel separately. Finally, the INRF$_C$-IQ distance will be obtained by applying the following formula:

$$\text{dist}(I, I_D) = \alpha \cdot \text{dist}_R(I, I_D) + \beta \cdot \text{dist}_G(I, I_D) + \delta \cdot \text{dist}_B(I, I_D) \tag{7.3}$$

  where dist_R, dist_G, and dist_B are the INRF-IQ distances calculated in each RGB channels respectively, and the parameters α, β, and α are obtained by maximizing the correlation between the INRF$_C$-IQ distance and the mean opinion score in the TID2008 database.

Since the correlation obtained using these two implementations was lower than the correlation obtained with the standard INRF-IQ implementation (applied to the luminance channel of CIELAB color space), they were discarded.

7.8 Discussion and future work

The proposed metric with the optimal parameters obtained as described in Section 7.4 achieves significant results in benchmark image quality databases (i.e. TID2008, TID2013, LIVE, or CSIQ), see

Table 7.2, outperforming some neural network metrics (e.g. LPIPS [101], and DISTS [17]), but with a number of parameters several orders of magnitude less (4 parameters). However, it shows a limited performance when generalizing to alternative datasets containing other types of distortions: denoising, rendering, texture synthesis, etc, see Table 7.3. This behaviour can be explained by analyzing each database individually. For instance, the texture synthesis database synTEX [29], consists of a set of synthesized texture images. Some metrics that measure texture similarity, such as DISTS [17] outperform our proposed metric, as the INRF metric is a spatial dependent metric. Other databases that contain spatial distortions, such as Tian19 [82], are also a challenge to spatial dependent metrics.

As future work, we propose study more in detail and create alternative implementations of INRF-IQ to deal with color information.

8

Conclusions and future work

In this thesis, we have focused on developing different image processing algorithms that can be useful for movie creators as artistic tools. Therefore, the proposed algorithms have low computational cost, so they can be used on-set to produce real-time results, allowing cinematographers to experiment with them. We have described the transition from film cinematography to its digital format to be able to understand the current necessities of digital cinema. We have also studied the camera inner processes to redefine these problems in terms of digital camera knowledge.

In Chapter 5, different methods for style transfer have been analyzed. It has been observed that statistical-based methods are simple, have low computational cost, and are effective in terms of producing results free of artifacts. The election of an uncorrelated color space is important for statistical-based methods, so PCA has been applied to images in order to choose an adequate color space. Most of these types of methods are designed to transfer the style between still images, so an extension of them is proposed to transfer the style to video. A psychophysical validation has been conducted, showing that the proposed method outperforms algorithms from the state of the art in the academic literature and is comparable to the methods in the industry.

In Chapter 6, the addition of texture to images has been proposed with two complementary purposes: for aesthetical reasons, and for coding efficiency purposes. A "retinal noise" model is proposed, inspired by processes of the visual system, which produces visually pleasing results. The method is preferred over other methods that emulate film grain, as psychophysical experiments show. Also, the perceived quality of images is increased by the addition of retinal noise, allowing for a significant improvement in coding efficiency, with average bit rate savings of over 22.5%.

Quality evaluation is of crucial importance in the image and video processing field. Therefore, in Chapter 7, an image quality metric is presented, and its optimization process is explained. The vision model behind the metric, unlike most models in the vision science literature, assumes a non-linear response of neurons. This metric achieves significant results in benchmark image quality databases (i.e. TID2008, TID2013, LIVE, or CSIQ), outperforming some neural network metrics (e.g. LPIPS [101], and DISTS [17]), but with a number of parameters several orders of magnitude less (4 parameters).

8.1 Future work

Our current implementation in Chapter 5 can be improved by incorporating a number of extensions, like allowing for foreground/background segmentation in order to overcome challenging transformation where a global approach is not able to represent the characteristics of the images. Also, the use of keyframes can be helpful to select the most suitable image for video style transfer.

The retinal noise model proposed in Chapter 6 can be improved by exploring the impact of the type of content in the usefulness of the addition of retinal grain noise to mask compression artifacts, as it has been observed a large variability between different videos.

In Chapter 7, the proposed quality metric obtains significant results in standard quality databases, while in other types of databases

its performance is limited. These results can be analyzed more in detail in order to improve the metric. An interesting line of research is to develop alternative implementations of INRF-IQ to deal with color information.

9

Appendix I

This appendix includes a compilation of some databases used for image quality assessment. An *image quality database* generally contains a set of reference and distorted images, along with the corresponding average ratings for each distorted image, obtained from quality-rating studies. Some of the databases used for optimization and validation of our proposed metric are:

- *TID2008* and *TID2013* databases were proposed by Ponomarenko et al. [59, 60]. The first one contains 25 reference images and 1700 distorted images (25 reference images $\times$ 17 types of distortions $\times$ 4 levels of distortions). The types of distortions in this database are: additive Gaussian noise, additive noise in color components is more intensive than additive noise in the luminance component, spatially correlated noise, masked noise, high-frequency noise, impulse noise, quantization noise, gaussian blur, image denoising, JPEG compression, JPEG2000 compression, JPEG transmission errors, JPEG2000 transmission errors, non-eccentricity pattern noise, local block-wise distortions of different intensity, mean shift (intensity shift), and contrast change. TID2013 database contains 24 types of distortions, 17 are common with TID2008, and the additional types

are: change of color saturation, multiplicative Gaussian noise, comfort noise, lossy compression of noisy images, image color, chromatic aberrations, and sparse sampling and reconstruction. This database contains 25 reference images and 3000 distorted images (25 reference images trusted× 24 types of distortions × 5 levels of distortions). The ratings were collected from 838 observers for the first database, and 971 for TID2013 database.

- *LIVE Image Quality Database* [77] contains 29 reference images, and 779 distorted images. There are five distortion types: JPEG compression, JPEG2000 compression, additive Gaussian white noise, Gaussian blurring, and JPEG2000 with bit errors. The ratings were collected from 29 observers.

- *Categorical Subjective Image Quality (CSIQ) Database* [40], contains 30 reference images, and 866 distorted images. There are 6 types of distortions: JPEG compression, JPEG2000 compression, additive Gaussian white noise, additive Gaussian pink noise, Gaussian blurring, and global contrast decrements. The ratings were obtained from 35 subjects.

- The *FLT denoising* database [20] is created from distorted images by white Gaussian noise addition. It contains 75 reference images, and 300 denoised images, with a special emphasis on images with low contrast and noise-like texture. The denoised images have been obtained by filtering the distorted images with BM3D filters with four different thresholds. The authors describe the FLT database as more complex than earlier datasets, as noise-like textures can visually mask the noise affecting the perceived image quality.

- The *Liu13 deblurring* database [42] is created from synthetically motion-blurred images. It contains 40 reference images, and 1200 deblurred images, resulting from 5 different deblurring algorithms.

- The *Ma17 super-resolution* database [46] contains 30 reference images, and 1620 resulting images from applying nine super-resolution methods to low-resolution images (generated from the reference images).

- The *SHRQ dehazing* database [49] is created from 75 synthetic hazy images (regular and aerial images). It contains 75 reference images and 600 dehazed images from 8 dehazing algorithms. In this case, the quality of the dehazed image is rated by comparing it with the hazy image and the reference image (unlike the standard way of rating, where only the distorted and the reference images are shown to observers).

- The *Tian19 rendering* database [82] is created from depth-image-based rendered (DIBR) images, obtained from seven different DIBR algorithms. The process of DIBR synthesis consists in generating novel views of a scene from original texture images and associated depth information. Usually, DIBR algorithms incorporate inpainting techniques to fill the disocclusion holes that appear when generating novel views, and these inpainted areas can be a challenge to image quality metrics. This database contains 10 reference images and 90 synthesized images.

- The *SynTEX texture synthesis* database [29] is created from synthesized texture images. It contains 30 reference images, and 150 synthesized texture images, from 5 different texture synthesis algorithms.

Bibliography

[1] Apple (2018). HLS Authoring Specification for Apple Devices. `https://developer.apple.com/documentation/http_live_streaming/hls_authoring_specification_for_apple_devices`.

[2] Bentaleb, A., Taani, B., Begen, A. C., Timmerer, C., and Zimmermann, R. (2019). A survey on bitrate adaptation schemes for streaming media over http. *IEEE Communications Surveys Tutorials*, 21(1):562–585.

[3] Bertalmío, M. (2014). *Image processing for cinema*. CRC Press.

[4] Bertalmío, M. (2019). *Vision models for high dynamic range and wide colour gamut imaging: techniques and applications*. Academic Press.

[5] Bertalmío, M., Gomez-Villa, A., Martín, A., Vazquez-Corral, J., Kane, D., and Malo, J. (2020). Evidence for the intrinsically nonlinear nature of receptive fields in vision. *Scientific reports*, 10(1):1–15.

[6] Blackmagic (2019). Blackmagic pocket cinema camera 4k gallery. https://www.blackmagicdesign.com/products/blackmagicpocketcinemacamera/gallery.

[7] Bonneel, N., Sunkavalli, K., Paris, S., and Pfister, H. (2013). Example-based video color grading. *ACM Transactions on Graphics*, 32(4).

[8] Brady, N. and Field, D. J. (2000). Local contrast in natural images: Normalisation and coding efficiency. *Perception*, 29(9):1041–1055.

[9] Brown, M. and Süsstrunk, S. (2011). Multi-spectral SIFT for scene category recognition. In *CVPR 2011*, pages 177–184. IEEE.

[10] Cámara, M., Díaz, C., Casal, J., Ruano, J., and García, N. (2019). Perceptually equivalent resolution in handheld devices for streaming bandwidth saving. *IEEE Signal Processing Letters*, 26(6):878–882.

[11] Carandini, M. and Heeger, D. J. (2012). Normalization as a canonical neural computation. *Nature Reviews Neuroscience*, 13(1):51.

[12] Cisco (2019). Visual Networking Index: Forecast and Methodology, 2017-2022. https://www.cisco.com/c/en/us/solutions/collateral/service-provider/visual-networking-index-vni/white-paper-c11-741490.pdf.

[13] Coote, J. (1993). *The Illustrated History of Colour Photography*. Fountain Press Ltd.

[14] Cyriac, P., Kane, D., and Bertalmío, M. (2015). Perceptual dynamic range for in-camera image processing. In *British Machine Vision Conference (BMVC)*. The British Machine Vision Association.

[15] Dai, J., Au, O. C., Pang, C., Yang, W., and Zou, F. (2010). Film grain noise removal and synthesis in video coding. In *IEEE International Conference on Acoustics, Speech and Signal Processing*, pages 890–893.

[16] DCI (2018). Digital cinema system specification.

[17] Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2020). Image quality assessment: Unifying structure and texture similarity. *CoRR*, abs/2004.07728.

[18] Ding, K., Ma, K., Wang, S., and Simoncelli, E. P. (2021). Comparison of full-reference image quality models for optimization of image processing systems. *International Journal of Computer Vision*, (4):1258–1281.

[19] Ebner, F. and Fairchild, M. D. (1998). Development and testing of a color space (IPT) with improved hue uniformity. In *Color and imaging conference*, number 1, pages 8–13. Society for Imaging Science and Technology.

[20] Egiazarian, K. O., Ponomarenko, M., Lukin, V. V., and Ieremeiev, O. (2017). Statistical evaluation of visual quality metrics for image denoising. *CoRR*, abs/1711.00693.

[21] Fan, Q., Li, X., Wang, S., Fu, S., Zhang, X., and Wang, Y. (2019). Na-caching: An adaptive content management approach based on deep reinforcement learning. *IEEE Access*, 7:152014–152022.

[22] Ferradans, S., Bertalmio, M., Provenzi, E., and Caselles, V. (2011). An analysis of visual adaptation and contrast perception for tone mapping. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 33(10):2002–2012.

[23] Ferradans, S., Papadakis, N., Rabin, J., Peyré, G., and Aujol, J.-F. (2013). Regularized discrete optimal transport. In *International*

Conference on Scale Space and Variational Methods in Computer Vision, pages 428–439. Springer.

[24] Frigo, O., Sabater, N., Demoulin, V., and Hellier, P. (2014). Optimal transportation for example-guided color transfer. In *Asian Conference on Computer Vision (ACCV)*, pages 655–670. Springer.

[25] G. Bjøntegaard (2001). Calculation of average PSNR differences between RD curves. *VCEG-M33*.

[26] Gatys, L. A., Ecker, A. S., and Bethge, M. (2016). Image style transfer using convolutional neural networks. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 2414–2423.

[27] Gil Rodríguez, R. et al. (2018). *Digital camera colour processing pipeline for high dynamic range imaging and colour stabilisation for cinema*. PhD thesis, Universitat Pompeu Fabra.

[28] Gil Rodríguez, R., Vazquez-Corral, J., and Bertalmío, M. (2020). Color matching images with unknown non-linear encodings. *IEEE Transactions on Image Processing*, 29:4435–4444.

[29] Golestaneh, S. A., Subedar, M. M., and Karam, L. J. (2015). The effect of texture granularity on texture synthesis quality. In *Applications of Digital Image Processing XXXVIII*, volume 9599, page 959912. International Society for Optics and Photonics.

[30] Grogan, M. and Dahyot, R. (2019). L2 divergence for robust colour transfer. *Computer Vision and Image Understanding*, 181:39–49.

[31] HaCohen, Y., Shechtman, E., Goldman, D. B., and Lischinski, D. (2011). Non-rigid dense correspondence with applications for image enhancement. *ACM transactions on graphics (TOG)*, 30(4):1–10.

[32] Hepburn, A., Laparra, V., Malo, J., McConville, R., and Santos-Rodriguez, R. (2020). Perceptnet: A human visual system inspired neural network for estimating perceptual distance. In *2020 IEEE International Conference on Image Processing (ICIP)*, pages 121–125. IEEE.

[33] Hordley, S. (2006). Scene illuminant estimation: Past, present, and future. *Color Research & Application*, 31:303 – 314.

[34] Huang, J. and Mumford, D. (1999). Statistics of natural images and models. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 541–547. IEEE.

[35] Kandel, E. R., Schwartz, J. H., Jessell, T. M., Siegelbaum, S., Hudspeth, A. J., and Mack, S. (2000). *Principles of neural science*, volume 4. McGraw-hill New York.

[36] Kotera, H. (2005). A scene-referred color transfer for pleasant imaging on display. In *IEEE International Conference on Image Processing 2005*, volume 2, pages II–5. IEEE.

[37] Kurihara, T., Manabe, Y., Aoki, N., and Kobayashi, H. (2011). Digital image improvement by adding noise: an example by a professional photographer. *Journal of Imaging Science and Technology*, 55(3):30503–1.

[38] L. Janowski and M. Pinson (2015). The accuracy of subjects in a quality experiment: a theoretical subject model. *IEEE Transactions on Multimedia*, 17(12):2210–2224.

[39] Laparra, V., Ballé, J., Berardino, A., and Simoncelli, E. (2016). Perceptual image quality assessment using a normalized laplacian pyramid. *Electronic Imaging*, 2016:1–6.

[40] Larson, E. C. and Chandler, D. M. (2010). Most apparent distortion: full-reference image quality assessment and the role of strategy. *Journal of electronic imaging*, 19(1):011006.

[41] Li, Y., Liu, M.-Y., Li, X., Yang, M.-H., and Kautz, J. (2018). A closed-form solution to photorealistic image stylization. In *Proceedings of the European Conference on Computer Vision (ECCV)*, pages 453–468.

[42] Liu, Y., Wang, J., Cho, S., Finkelstein, A., and Rusinkiewicz, S. (2013). A no-reference metric for evaluating the quality of motion deblurring. *ACM Transactions on Graphics (TOG)*, 32.

[43] Loertscher, M., Weibel, D., Spiegel, S., Flueckiger, B., Mennel, P., Mast, F., and Iseli, C. (2016). As film goes byte: The change from analog to digital film perception. *Psychology of aesthetics, creativity, and the arts*, 10:458–471.

[44] Luan, F., Paris, S., Shechtman, E., and Bala, K. (2017). Deep photo style transfer. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 4990–4998.

[45] Lukac, R. (2008). *Single-Sensor Imaging: Methods and Applications for Digital Cameras*. CRC Press, 1 edition.

[46] Ma, C., Yang, C.-Y., Yang, X., and Yang, M.-H. (2017). Learning a no-reference quality metric for single-image super-resolution. *Computer Vision and Image Understanding*, 158:1–16.

[47] McKernan, B. (2005). *Digital Cinema: The Revolution in Cinematography, Post-Production, and Distribution*. McGraw-Hill Education.

[48] Menon, D. and Calvagno, G. (2011). Color image demosaicking: An overview. *Signal Processing: Image Communication*, 26(8-9):518–533.

[49] Min, X., Zhai, G., Gu, K., Zhu, Y., Zhou, J., Guo, G., Yang, X., Guan, X., and Zhang, W. (2019). Quality evaluation of image dehazing methods using synthetic hazy images. *IEEE Transactions on Multimedia*, 21(9):2319–2333.

[50] Morovic, J. (1998). *To develop a universal gamut mapping algorithm*. PhD thesis, University of Derby, UK.

[51] Mudrová, M. and Procházka, A. (2005). Principal component analysis in image processing. In *Proceedings of the MATLAB technical computing conference*.

[52] Naccari, M. and Pereira, F. (2011). Advanced H.264/AVC-based perceptual video coding: architecture, tools, and assessment. *IEEE Transactions on Circuits and Systems for Video Technology*, 21(6):766–782.

[53] Nakamura, J. (2017). *Image sensors and signal processing for digital still cameras*. CRC press.

[54] Newson, A., Faraj, N., Galerne, B., and Delon, J. (2016). Realistic film grain rendering. *Image Processing On Line*, 7:165–183.

[55] Norkin, A. and Birkbeck, N. (2018). Film grain synthesis for AV1 video codec. In *2018 Data Compression Conference*, pages 3–12. IEEE.

[56] Oh, B. T., Lei, S.-m., and Kuo, C.-C. J. (2009). Advanced film grain noise extraction and synthesis for high-definition video coding. *IEEE Transactions on Circuits and Systems for Video Technology*, 19(12):1717–1729.

[57] Olshausen, B. A. and Field, D. J. (2000). Vision and the coding of natural images: The human brain may hold the secrets to the best image-compression algorithms. *American Scientist*, 88(3):238–245.

[58] Passaglia, C. L. and Troy, J. B. (2004). Impact of noise on retinal coding of visual signals. *Journal of neurophysiology*, 92(2):1023–1033.

[59] Ponomarenko, N., Ieremeiev, O., Lukin, V., Jin, L., Egiazarian, K., Astola, J., Vozel, B., Chehdi, K., Carli, M., Battisti, F., and Kuo, C. C. J. (2013). A new color image database TID2013: Innovations and results. In *Advanced Concepts for Intelligent Vision Systems*, pages 402–413. Springer International Publishing.

[60] Ponomarenko, N., Lukin, V., Zelensky, A., Egiazarian, K., Carli, M., and Battisti, F. (2009). TID2008 - A Database for Evaluation of Full-Reference Visual Quality Assessment Metrics. *Advances of Modern Radioelectronics*, 10:30–45.

[61] Poynton, C. (2003). *Digital Video and HDTV Algorithms and Interfaces*. Morgan Kaufmann Publishers Inc., 1 edition.

[62] Prashnani, E., Cai, H., Mostofi, Y., and Sen, P. (2018). Pieapp: Perceptual image-error assessment through pairwise preference. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 1808–1817.

[63] Rabin, J. and Papadakis, N. (2015). Non-convex relaxation of optimal transport for color transfer between images. In *International Conference on Geometric Science of Information*, pages 87–95. Springer.

[64] Ramanath, R., Snyder, W., Yoo, Y., and Drew, M. (2005). Color image processing pipeline. *IEEE Signal Processing Magazine*, 22(1):34–43.

[65] Recommendation ITU-R BT.2020-2 (2015). Parameter values for ultra-high definition television systems for production and international programme exchange.

[66] Recommendation ITU-R BT.500-13 (2012). Methodology for the subjective assessment of the quality of television pictures.

[67] Recommendation ITU-T P.910 (2008). Subjective video quality assessment methods for multimedia applications.

[68] Recommendation ITU-T P.913 (2016). Methods for the subjective assessment of video quality, audio quality and audiovisual quality of Internet video and distribution quality television in any environment.

[69] Reinhard, E., Ashikhmin, M., and Shirley, P. (2001). Color transfer between images. *IEEE Computer Graphics and Applications*, 21:34–41.

[70] Reinhard, E., Ward, G., Pattanaik, S., and Debevec, P. (2005). *High Dynamic Range Imaging: Acquisition, Display, and Image-Based Lighting*. Morgan Kaufmann Publishers Inc.

[71] Rouis, K., Larabi, M., and Belhadj Tahar, J. (2018). Perceptually adaptive lagrangian multiplier for hevc guided rate-distortion optimization. *IEEE Access*, 6:33589–33603.

[72] R.S. Allison, K. Brunnström, D.M. Chandler, H. Colett, P. Corriveau, S. Daly, J. Goel, J.Y. Long, L.M. Wilcox, Y. Yaacob, S.-N. Yang and Y. Zhang (2018). Perspectives on the definition of visually lossless quality for mobile and large format displays. *Journal of Electronic Imaging*, 27(5):1–23.

[73] Ruderman, D. L. (1994). The statistics of natural images. *Network: Computation in Neural Systems*, 5(4):517–548.

[74] Ruderman, D. L., Cronin, T. W., and Chiao, C.-C. (1998). Statistics of cone responses to natural images: implications for visual coding. *Journal of the Optical Society of America A*, 15(8):2036–2045.

[75] Shapley, R. and Enroth-Cugell, C. (1984). Visual adaptation and retinal gain controls. *Progress in retinal research*, 3:263–346.

[76] Sheikh, H. and Bovik, A. (2006). Image information and visual quality. *IEEE Transactions on Image Processing*, 15(2):430–444.

[77] Sheikh, H., Sabir, M., and Bovik, A. (2006). A statistical evaluation of recent full reference image quality assessment algorithms. *IEEE Transactions on Image Processing*, 15:3440–3451.

[78] SMPTE ST 2036-1:2014 (2014). Ultra High Definition Television – Image Parameter Values for Program Production.

[79] Sun, H., Rüttiger, L., and Lee, B. B. (2004). The spatiotemporal precision of ganglion cell signals: a comparison of physiological and psychophysical performance with moving gratings. *Vision research*, 44(1):19–33.

[80] Sun, H. and Shi, Y. Q. (2008). *Image and video compression for multimedia engineering: Fundamentals, algorithms, and standards*. CRC press.

[81] Tavakoli, S., Gutierrez, J., and Garcia, N. (2014). Subjective quality study of adaptive streaming of monoscopic and stereoscopic video. *IEEE Journal on Selected Areas in Communications*, 32(4):684–692.

[82] Tian, S., Zhang, L., Morin, L., and Déforges, O. (2019). A benchmark of DIBR synthesized view quality assessment metrics on a new database for immersive media applications. *IEEE Transactions on Multimedia*, 21(5):1235–1247.

[83] Upenik, E., Rerabek, M., and Ebrahimi, T. (2017). On the Performance of Objective Metrics for Omnidirectional Visual Content. In *IEEE 2017 Ninth International Conference on Quality of Multimedia Experience*, pages 1–6.

[84] Vazquez-Corral, J. and Bertalmío, M. (2014). Color stabilization along time and across shots of the same scene, for one or several cameras of unknown specifications. *IEEE Transactions on Image Processing*.

[85] Wan, X., Kobayashi, H., and Aoki, N. (2015). Improvement in perception of image sharpness through the addition of noise and its relationship with memory texture. In *Human Vision and Electronic Imaging XX*, volume 9394, page 93941B.

[86] Wang, Z., Bovik, A., Sheikh, H., and Simoncelli, E. (2004). Image quality assessment: from error visibility to structural similarity. *IEEE Transactions on Image Processing*, 13(4):600–612.

[87] Wang, Z., Simoncelli, E., and Bovik, A. (2003). Multiscale structural similarity for image quality assessment. In *The Thrity-Seventh Asilomar Conference on Signals, Systems & Computers, 2003*, volume 2, pages 1398–1402. IEEE.

[88] Wang, Z. and Simoncelli, E. P. (2005). Translation insensitive image similarity in complex wavelet domain. In *Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing*, volume 2, pages ii–573. IEEE.

[89] Watson, A. B. and Ahumada, A. J. (2005). A standard model for foveal detection of spatial contrast. *Journal of vision*, 5(9):6–6.

[90] Wikipedia (2021). `https://en.wikipedia.org/wiki/Cinematography`.

[91] Williams, C. (1996). *Cinema: The beginnings and the future.* University of Westminster Press.

[92] Xiao, X. and Ma, L. (2006). Color transfer in correlated color space. In *Proceedings of the 2006 ACM international conference on Virtual reality continuum and its applications*, pages 305–309.

[93] Xue, W., Zhang, L., Mou, X., and Bovik, A. C. (2014). Gradient magnitude similarity deviation: A highly efficient perceptual image quality index. *IEEE Transactions on Image Processing*, 23(2):684–695.

[94] Zabaleta, I. and Bertalmío, M. (2018). In-camera, photorealistic style transfer for on-set automatic grading. In *SMPTE 2018 Annual Technical Conference and Exhibition*, pages 1–12.

[95] Zabaleta, I. and Bertalmío, M. (2021). Photorealistic style transfer for video. *Signal Processing: Image Communication*, 95:116240.

[96] Zabaleta, I., Cámara, M., Díaz, C., Canham, T., García, N., and Bertalmío, M. (2020). Retinal noise emulation: A novel artistic tool for cinema that also improves compression efficiency. *IEEE Access*, 8:67263–67276.

[97] Zamir, S. W., Vazquez-Corral, J., and Bertalmío, M. (2017). Automatic, fast and perceptually accurate gamut mapping based on vision science models. In *SMPTE 2017 Annual Technical Conference and Exhibition*, pages 1–14.

[98] Zamir, S. W., Vazquez-Corral, J., and Bertalmio, M. (2019). Vision models for wide color gamut imaging in cinema. *IEEE Transactions on Pattern Analysis and Machine Intelligence*.

[99] Zhang, L., Shen, Y., and Li, H. (2014). VSI: A visual saliency-induced index for perceptual image quality assessment. *IEEE Transactions on Image Processing*, 23(10):4270–4281.

[100] Zhang, L., Zhang, L., Mou, X., and Zhang, D. (2011). FSIM: A feature similarity index for image quality assessment. *IEEE Transactions on Image Processing*, 20(8):2378–2386.

[101] Zhang, R., Isola, P., Efros, A. A., Shechtman, E., and Wang, O. (2018). The unreasonable effectiveness of deep features as a perceptual metric. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 586–595.