Postgraduate Study Report DC-PSR-2004-06

Human Perception and Computer Graphics

Martin Čadík

Supervisor: *Pavel Slavík*

January 2004

Department of Computer Science and Engineering
Faculty of Electrical Engineering
Czech Technical University
Karlovo nám. 13
121 35 Prague 2
Czech Republic

email: cadikm@fel.cvut.cz
WWW: www.cgg.cvut.cz/~cadikm

This report was prepared as a part of the project

Perceptually Based Acceleration of Rendering

This research has been partly supported by the Ministry of Education, Youth and Sports of the Czech Republic under research program No. Y04/98: 212300014 (Research in the area of information technologies and communications).

. . . . . .
Martin Čadík
postgraduate student

. . . . . .
Pavel Slavík
supervisor

Contents

Human Perception and Computer Graphics

Martin Čadík
cadikm@fel.cvut.cz
Department of Computer Science and Engineering
Faculty of Electrical Engineering
Czech Technical University
Karlovo nám. 13
121 35 Prague 2
Czech Republic

Abstract

This report concerns human perception and its applications to the domain of computer graphics. Having in mind human perception limitations, we can design a perceptually optimized approach to virtually any issue of contemporary computer graphics. Such a perceptually optimized approach enable us either to visualize information more effectively and consequently to grasp important ideas and information from the depiction at a glance, or to save computational time or improve the quality of results by removing perceptually non-important parts of visual simulation. Initially, we outline the anatomy of human visual system (HVS) and characteristics of human perception. Consecutively, we summarize the usage of HVS knowledge in computer graphics, we point out the bottlenecks of contemporary methods and we give the suggestions for future research. Specifically, we cover the issues of the image quality testing, the image comparison, and the acceleration of visual simulations and rendering. Finally, we present an experimental study on comparing image-processing operators.

Keywords

human perception, human visual system, computer graphics, image quality, vision models, image comparison, acceleration of rendering

1 Introduction

> *"The goal of computer graphics is not to control light, but to control our perception of light. Light is merely a carrier of the information we gather by perception."*
>
> (Jack Tumblin, James A. Ferwerda)

Outputs of computer graphics are intended to be observed by human subjects. As human vision has several limitations, the knowledge of the human visual system (HVS) and of the human perception can be utilized to improve the performance of various computer graphics algorithms. In the field of computer graphics the knowledge of the human visual system usually takes the form of the computational models of human vision. Such a model can be incorporated at various areas of computer graphics.

One of the areas where the incorporation of human vision models is extremely beneficial is the image quality assessment and the image comparison. Image quality assessment and comparison metrics play an important role in various computer graphics applications. They can be used to monitor image quality for quality control systems, they can be employed to benchmark image processing algorithms, and they can be embedded into an image processing system to optimize the algorithms and the parameter settings. It is well known [49], that classical comparison metrics like Root Mean Square (RMS) error are not sufficient when applied to the comparison of images, because they poorly predict the differences between the images as perceived by the human observer. To solve the problem properly the visual differences predictors have evolved. The main part of visual differences predictors is typically a model of early vision, so that they perform well when comparing visually very near images. However their performance when comparing quite different images with respect to the contained information is poor. The predictor capable to incorporate such a behaviour would be valuable in the image database retrievals, to evaluation of the perceptual impact of different rendering algorithms, to analysis of the effect of various acceleration techniques, etc.

Alhough the progress in the computer hardware still persists, realistic image simulations and even computer generated animations are yet far away to be computed interactively. However, no matter how carefully we compute the displayed image, perception determines how much and how accurately we will understand what we see. Some visible difference predictors have been successfully applied to the fields of realistic image synthesis and synthetic animations, but most of current work in computer graphics does not consider any perceptual certainty. The search for means of utilization of human perception in rendering algorithms, realistic image synthesis, and computer animation, must go on.

The other problem is that much of the known HVS data has been obtained from specific psychophysical experiments which have been conducted in specialised laboratory environments under reductionistic conditions. These experiments were designed to examine a single dimension of human vision, however, evidence exists [39] to indicate that features of the HVS do not operate individually, but rather number of functions overlap and should be examined as a whole rather than in isolation. There is a strong need for the models of human vision currently used in image synthesis computations to be validated to demonstrate their performance is comparable to the actual performance of the HVS.

The goal of this report is to present specific properties of human visual system that are, or could be employed in computational models. Known visible differences predictors, their drawbacks and advantages are to be summarized. Consecutively, the overview of utilization of these predictors to the computer graphics rendering is to be given. Finally, the way to overcome drawbacks of contemporary approaches to the image comparison is to be outlined.

1.1 Organization of the Report

The report is organized as follows. Section 2 covers the physical structure and the perceptual behaviour of the human visual system. Section 3 summarizes the applications of human perception to computer graphics, namely to the field of image quality measurement. Section 4 describes usage of human perception in order to improve the rendering. Section 5 describes our contribution to the image comparison field. Section 6 concludes the report. Section 7 is the list of references and Section 8 gives the abstract of author's prospective dissertation thesis. The appendix describes several widely used principles of method design and data analysis.

In this report, we do not cover the tone mapping field, which is concerned with the problem of mapping a bright scale of image luminances onto a narrow scale of display device in such a way that the perceived displayed image can be thought of as producing the same mental image as the original image. See the SIGGRAPH'03 Course #19 [10] for an overview. Furthermore, we

do not consider perceptual optimizations of 3D graphics and modeling, i.e. perceptual criteria for level of detail, and adaptive mesh subdivision and simplification, see outlines by Reddy [42] and the recent SIGGRAPH'03 Course #03 [13] respectively for a summary and references.

2 Human Visual System

Since the majority of perceptually based approaches to computer graphics is inspired by properties of the human visual system, we will describe the HVS first. We will begin with the physical structure that is quite well estabilished and that can help us to understand rather complex characteristics of the perceptual behaviour. This part of the report was largely acquired from the excellent book on vision science by S. Palmer [39].

2.1 Physical Structure

In this section we will describe basic visual anatomy and physiology. This can give us insights into the kinds of information that can be coded by visual mechanisms.

2.1.1 The Human Eye

Humans have two eyes, which are approximately spherical in shape except for a bulge at the front. Located at about the horizontal midline of the head, they sit in nearly hemispherical holes in the skull, called the eye sockets. Each eye is moved by the coordinated use of six small, strong muscles, called the *extraocular muscles*, which are controlled by specific areas in the brain. Eye movements are necessary for scanning different regions of the visual field without having to turn the entire head and for focusing on objects at different distances.

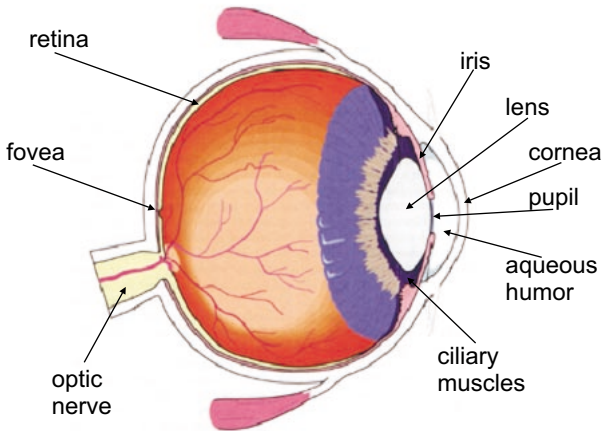

Figure 1: A cross section of the human eye. (After Kolb et al. [24].)

The eyes have two important optical functions: to gather light reflected from surfaces in the world and to focus it in a clear image on the back of the eye. There are many parts of the eye that accomplish different optical functions, see Figure 1. First, light enters the *cornea*, a transparent bulge on the front of the eye behind which is a cavity filled with a clear liquid, called the *aqueous humor*. Next, light passes through the *pupil*, a variably sized opening in the opaque *iris*, which gives the eye its external color. Just behind the iris, light passes through the *lens*, whose shape is controlled by *ciliary muscles*. The len's optical properties can be altered by changing its shape, a process called *accommodation*. The photon then travels through the clear *vitreous humor* that fills the central chamber of the eye. Finally, it reaches the *retina*,

the curved surface at the back of the eye. The retina is densely covered with over 100 million light-sensitive *photoreceptors*, which convert light into neural activity.

The information about the light striking the retina is transmitted to the primary visual cortex in the occipital lobe at the back of the head, see Figure 2. Some estimates put the percentage of cortex involved with visual function at more than 50% in the macaque monkey, athough it is probably slightly lower in humans. The complete visual system includes much of the brain as well as the eyes, and the whole eye-brain system must function properly for the organism to extract reliable information about the environment.

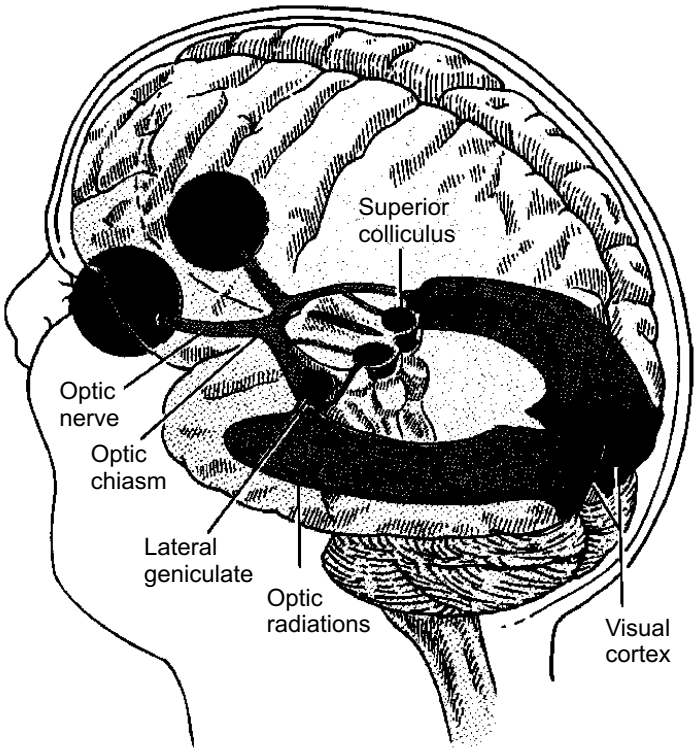

Figure 2: The human visual system. (After Palmer [39].)

2.1.2 The Retina

After the optics of the eye have done their job, the next critical function of the eye is to convert light into neural activity so that the brain can process the optical information. In the visual system, this function is carried out by *photoreceptors* in the retina: photoreceptors are specialized retinal cells that are stimulated by light energy. There are two distinct classes of photoreceptor cells: *rods* and *cones*. Rods are more numerous (about 120 million), extremely sensitive to light, and located everywhere in the retina except at its very center. They are used exclusively for vision at very low light levels (called *scotopic conditions*). Cones are less abundant (about 8 million), much less sensitive to light, and heavily concentrated in the center of the retina, although some are found scattered throughout the periphery. They are responsible for our visual experiences under most normal lighting conditions (called *photopic conditions*) and for all our experiences of colour. There is a small region, called the *fovea*, right at the center of the retina that contains nothing but densely packed cones. The visual angle covered by the fovea is only about 2 degrees. Another region exists where the axons of the ganglion

cells leave the eye at the optic nerve. This region is called the *optic disk* (also known as the *blind spot*) and it contains no receptor cells at all. However, we do not experience blindness there, except under very special circumstances.

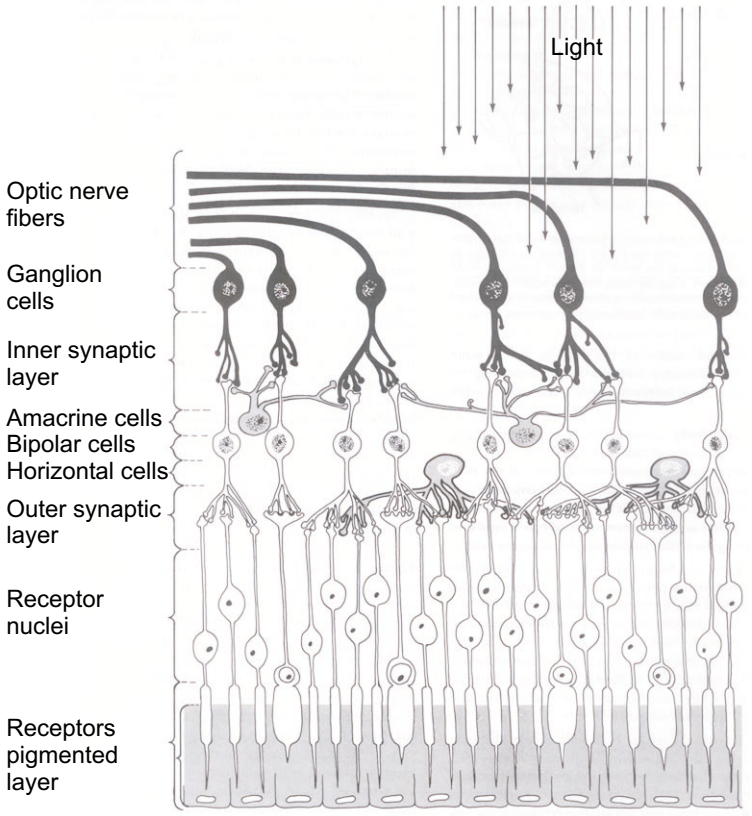

Figure 3: The human retina. (After Palmer [39].)

Once the optical information is coded into neural responses, some initial processing is accomplished within the retina itself by several other types of neurons, including the *horizontal*, *bipolar*, *amacrine*, and *ganglion cells*, all of which integrate responses from many nearby cells, see Figure 3. The inputs of the retinal ganglion cells are arranged in an antagonistic, concentric pattern composed of a centre and a surround region (the area of the retina which the ganglion cell receives input from is called the *receptive field*). The ganglion cell is continually emitting a background signal; however when light strikes the photoreceptors in one region, this stimulates an increased response from the retinal ganglion cell (*on-response*), whereas light falling on the other region will generate a reduced response (*off-response*). There are two distinct types of ganglion cells, the *on-center cells*, where the centre region is stimulated by an on-response, and the *off-center cells*, where the centre region is stimulated by an off-response, see Figure 4.

The axons of the ganglion cells carry information out of the eye through the *optic nerve* to the *optic chiasm*. Here the fibres from the nasal side of the fovea in each eye cross over to the opposite side of the brain while the others remain on the same side. The result is that the mapping from external visual fields to the cortex is completely crossed – all of the information from the left half of the visual field goes to the right half of the brain, while all the information from the right visual field goes to the left half of the brain. From the optic chiasm, there are two separate pathways into the brain on each side. The smaller one (only a few percent) goes to the *superior colliculus*, a nucleus in the brain stem. This visual center seems to process primarily information about where things are in the world and to be involved in the control of

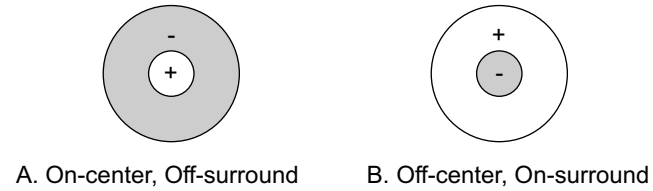

Figure 4: Receptive field structure of ganglion cells. On-center, off-surround cells (A) fire to light onset and stop at offset in their excitatory center, but they stop firing to light onset and begin firing at offset in their inhibitory surround. Off-center, on-surround cells (B) exhibit the opposite characteristics. (After Palmer [39].)

eye movements. The larger pathway goes first to the *lateral geniculate nucleus* (or *LGN*) of the thalamus and then to the *occipital cortex* (or *primary visual cortex*).

2.1.3 Visual Cortex

The human cortex is divided into two halves, *cerebral hemispheres*, that are approximately symmetrical. As a result of many neuropsychological studies, it is now well estabilished that the occipital lobe is the primary cortical receiving area for visual information. Although it would be a gross overstatement to say that vision scientists understand how visual cortex works, they are at least beginning to get some glimmerings of what the assorted pieces might be and how they might fit together.

The first steps in cortical processing of visual information take place in the *striate cortex*, sometimes called *primary visual cortex* or area *V1*. This is the largest part of the occipital lobe and it seems likely that the most complex visual processing occurs there. Striate cortex receives its input from the LGN on the same side of the brain, so the visual input of striate cortex, like that of LGN, is completely crossed. Both sides are activated by the thin central vertical strip, measuring about 1 degree of visual field. The cells that are sensitive to this strip in one side of the brain are connected to the corresponding cells on the other side of the brain through the *corpus callosum*, the large fiber tract that allows communication between the two cerebral hemispheres. The mapping from retina to striate cortex is topological in that nearby regions on the retina project to nearby regions in striate cortex. The central area of the visual field, which falls on or near the fovea, receives proportionally much greater representation in the cortex than the periphery does. This is called the *cortical magnification factor*.

The inferior temporal centers in the lower (ventral) system seem to be involved in *identifying objects*, whereas the parietal centers in the upper (dorsal) system seem to be involved in *locating objects*. These two pathways are often called the *"what" system* and the *"where" system*, respectively. It seems almost inevitable that these two different kinds of information must get together somewhere in the brain so that the "what-where" connection can be made, but it is not yet known where this happens.

It is now abundantly clear that a great deal of visual processing takes place in parallel across different areas, each region projecting fibers to several other areas but by no means to all of them. The connections are generally bidirectional; that is, if area X projects to area Y, then Y projects back to X as well.

The Physiological Pathways Hypothesis

One possible relation between anatomical structure and physiological function has begun to emerge during the last decade. The hypothesis is that there are separate neural *pathways* for processing information about different visual properties such as color, shape, depth, and motion. Livingstone and Hubel [28] proposed that these four types of information are processed in different neural pathways from the retina onward. They report evidence that color, form, motion, and stereoscopic depth information are processed in distinct subregions of visual cortical areas V1 and V2, as indicated schematically in Figure 5.

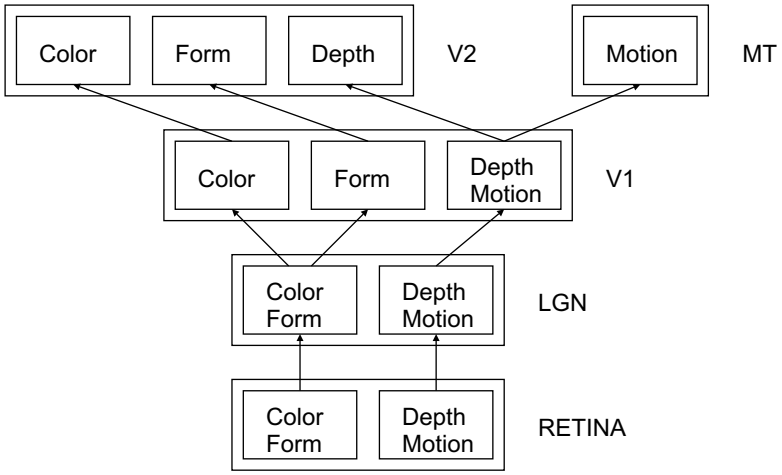

Figure 5: Schematic diagram of the visual pathways hypothesis.

These areas then project to distinct higher-level areas of cortex: movement and stereoscopic depth information to area *V5* (also called *MT*, Medial Temporal cortex), color to area *V4*, and form through several intermediate centers (including V4) to area *IT* (InferoTemporal cortex), where cells have been found that respond selectively to faces, hands, and other highly complex stimuli. From these areas, the form and color pathways may project to the ventral "what" system for object identification and the depth and motion pathways to the dorsal "where" system for object localization.

2.2 Visual Perception

Visual perception is the process of acquiring knowledge about environmental objects and events by extracting information from the light they emit or reflect. Visual perception concerns the *acquisition of knowledge* – this means that vision is fundamentally a cognitive activity, distinct from purely optical processes such as photographic ones. There are indeed important similarities between eyes and cameras in terms of optical phenomena, but there are no similarities whatever in terms of perceptual phenomena – cameras have no perceptual capabilities at all. The knowledge achieved by visual perception concerns *objects and events in the environment*, perception is not merely about an observer's subjective visual experiences. Visual knowledge about the environment is obtained by *extracting information*, that implies information processing approach to the vision. Finally the information that is processed in visual perception comes from the light that is *emitted or reflected by objects*, optical information is the foundation of all vision.

2.2.1 Adaptation

Visual perception changes over time as it adapts to particular conditions. When we enter a darkened room on a bright afternoon, for instance, we cannot see much. After 20 minutes, however, we can see everything surprisingly well. This increase in sensitivity to light is called *dark adaptation*. Adaptation is a very general phenomenon in visual perception – visual experience may become less intense as a result of prolonged exposure to a wide variety of different kinds of stimulation: color, orientation, size, motion, etc. These changes in visual experience show that visual perception is not always a clear window onto reality because we have different visual experiences of the same physical environment at different stages of adaptation.

2.2.2 Ambiguous figures

To provide us with information, vision is an *interpretive process* that somehow transforms complex, moving, two-dimensional patterns of light at the back of the eyes into stable perceptions of three-dimensional space. The objects we perceive are actually interpretations based on the structure of images rather than direct registrations of physical reality. Potent demonstrations of the interpretive nature of vision come from *ambiguous figures*, single images that can give rise to two or more distinct perceptions, see for an example Figure 6. The interpretations of these ambiguous figures are *mutually exclusive*. We perceive just one of them at a time: a duck or a rabbit, not both. This is consistent with the idea that perception involves the construction of an interpretive model because only one such a model can be fit to the sensory data at one time. If perception was completely determined by the light stimulating the eye, there would be no ambiguous figures because each pattern of stimulation would map onto a unique percept.

Figure 6: Ambiguous figures. Figure on the left can be seen as a duck (facing left) or a rabbit (facing right). Figure on the right can be seen as a saxophonist (facing right) or a face of a woman.

2.2.3 Visual Completion

People's perceptions actually correspond to the *models* that their visual systems have constructed rather than to the sensory stimulation on which the models are based. That is why perceptions can be illusory and ambiguous despite the nonillusory and unambiguous status of the raw optical images in which they are based. Perceptual models must be closely coupled to the information in the projected image of the world and must provide reasonably accurate interpretations of this information.

Perhaps the most convincing evidence that visual perception involves the construction of environmental models comes from the fact that our perceptions include portions of surfaces that we cannot actually see. This perceptual filling in of parts of objects that are hidden from

view is called *visual completion*. It happens automatically and effortlessly whenever we perceive the environment. Visual perception also includes information about *self-occluded surfaces*, those surfaces of an object thar are entirely hidden from view by its own visible surfaces.

2.2.4 Impossible Objects

Impossible ojects are two-dimensional line drawings that initially give the clear perception of coherent three-dimensional objects but are physically impossible, see Figure 7. Such demonstrations support the idea that vision actively constructs environmental models rather than simply registering what is present. If visual perception were merely an infallible reflection of the world, a physically impossible object simply could not be perceived. The kinds of errors that are evident in perceiving impossible objects seem to indicate that at least some visual processes work initially at a local level and only later fit the results into a global framework.

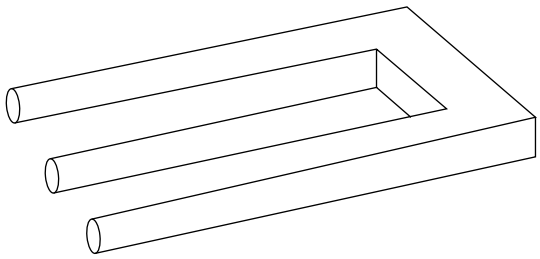

Figure 7: Impossible object. The drawing in this figure produce perception of coherent three-dimensional object, but it is physically impossible.

2.2.5 Classification

Our perceptual constructions go even further than completing unseen surfaces. They include information about the meaning or functional significance of objects and situations. Beeing able to *classify* objects as members of known categories allows us to respond to them in appropriate ways because it gives us access to vast amounts of information that we have stored from previous experiences with similar object. Previous experience with members of a given category allows us to predict with reasonable certainty what new members of the same class will do. As a consequence, we can deal with most new objects at the more abstract level of their category, even though we have never seen that particular object before.

2.2.6 Attention and Consciousness

The visible environment contains much more information than anyone can fully perceive. We must therefore be *selective* in what we attend to, and what we select will depend a great deal on our needs, goals, plans, and desires. Perception is not therefore an entirely stimulus-driven process – perceptions are not determined solely by the nature of the optical information present in sensory stimulation. Our perceptions are also influenced by *cognitive constraints* – higher-level goals, plans, and expectations. We look at different things in our surroundings depending on what we are trying to accomplish, and we may perceive them differently as a result.

One of the functions of *attention* is to bring visual information to consciousness. Certain properties of objects do not seem to be experienced consciously unless they are attended, yet unattended objects are often processed fully enough outside of consciousness to attract our attention. Once the object is attended, we become conscious of its detailed properties and

are able to identify it and discern its meaning in the present situation. In general, lower levels of perception do not seem to be accessible to, or modifiable by, conscious knowledge and expectations, whereas higher levels do. However not much is yet known about the role of consciousness in perception.

2.3 Image Processing Theories

Several theories for description of the nature of human image processing have evolved over the years. These theories compete because their functional implications are quite different and no one is universally held. We will describe the two most common of them: line–edge detection theory, and spatial frequency theory, although yet another approaches exist (connectionistic theory, neural networks, scale-space, etc.). Finally we will give an overview of contemporary theoretical hypothesis about the architecture of human image processing. This hypothesis integrates in some respect these two fundamentally different theories.

2.3.1 Line and Edge Detection Theory

Hubel and Wiesel [22] were the first to successfully apply the receptive field mapping techniques (described in section 2.1.2) to striate cortex. Their investigations revealed that there were several different kinds of cortical cells that had different receptive field characteristics. They classified them into three types: *simple cells, complex cells,* and *hypercomplex cells*. For simple cells, the responses to complex stimuli can be predicted from their responses to individual spots of light. A simple cell's response to a larger, more complex pattern of stimulation can therefore be roughly predicted by summing its responses to the set of small spots of light that compose it. There were identified several different subtypes of simple cells. The vast majority have an elongated structure, firing most vigorously to a line or an edge at a specific retinal position and orientation. Simple cells that have an area of excitation on one side and an area of inhibition on the other, respond to a luminance edge in the proper orientation and are called *edge detectors*. Simple cells that have receptive fields with a central elongated region that is either excitatory or inhibitory, with an antagonistic field on both sides of it, respond maximally to bright or dark lines and are called *line detectors* or *bar detectors*.

The view of image processing that has emerged from these findings is that an early step in spatial image processing is to find the lines and edges in the image. Higher-level properties, such as shapes and orientations of objects, might then be constructed by putting together the many local edges and lines that have been identified by their detector cells in V1. Whether or not this is the correct view is still an open question, but it has dominated thinking about the initial stages of visual processing for several decades.

About 75% of the cells in striate cortex are *complex cells*. Similar to simple cells, complex cells have elongated receptive fields but they differ from simple cells in several important respects: complex cells are highly *nonlinear*, they tend to be highly responsive to *moving* lines or edges anywhere within their receptive field. Complex cells are not very sensitive to the *position* of certain stimuli and they tend to have somewhat *larger receptive fields* than simple cells. Complex cells are thought to receive input from several simple cells whose receptive fields have the same orientation but different positions.

The third type of striate cell is the *hypercomplex cell*. The most striking characteristic of hypercomplex cells is that extending a line or edge beyond a certain length causes them to fire less vigorously than they do to a shorter line or edge. For this reason, they are often called *end-stopped cells*. Recent quantitative studies suggest that the degree of "end-stopping" is a continuum rather than an all-or-none phenomenon.

2.3.2 Spatial Frequency Theory

The spatial frequency theory dominates psychophysical theories of early spatial vision because it is able to explain a large number of important and surprising results from psychophysical experiments, not only in adult vision, but in infant vision as well. This theory is based on an atomistic assumption: the representation of any image, no matter how complex, is an assemblage of many primitive spatial "atoms". The primitives of spatial frequency theory are spatially extended patterns called *sinusoidal gratings*: two-dimensional patterns whose luminance varies according to a sine wave over one spatial dimension and is constant over the perpendicular dimension. Each primitive sinusoidal grating can be characterized completely by four parameters: its *spatial frequency, orientation, amplitude,* and *phase*. Spatial frequency is usually specified in terms of the number of light/dark cycles per degree of visual angle. The orientation of the grating is specified in degrees counterclockwise from vertical. Phase is specified in degrees, such that the grating whose positive-going inflection point is at the reference point is said to have a phase 0° (called sine phase).

There is a good formal mathematical reason for choosing sinusoidal gratings as primitives: *Fourier analysis*. Fourier analysis is a method, by which any two-dimensional luminance image can be analyzed into the sum of a set of sinusoidal gratings that differ in spatial frequency, orientation, amplitude, and phase. The Fourier analysis of an image consists of two parts: the power spectrum and the phase spectrum. The *power spectrum* specifies the amplitude of each grating at a particular spatial frequency and orientation, whereas the *phase spectrum* specifies the phase of each grating at a particular spatial frequency and orientation. If all of these gratings at the proper phases and amplitudes were added up, they would exactly recreate the original image. Thus, Fourier analysis provides a very general method of decomposing complex images into primitive components, since it has been proven to work for any image. Fourier analysis is also capable of being "inverted" through a process called *Fourier synthesis* so that the original image can be reconstructed from its power and phase spectra.

Spatial Frequency Channels

The spatial frequency theory proposes that early visual processing can be understood in terms of a large number of overlapping *psychophysical channels* that are selectively tuned to different ranges of spatial frequencies and orientations. Thanks to many psychophysical studies of people's detection and discrimination of grating stimuli [1], there is now a great deal of evidence to support this view.

The standard measurement of human sensitivity to gratings at different frequencies is called the *contrast sensitivity function (CSF)*. It is determined by finding the lowest contrast at which the observer can just barely detect the difference between a sinusoidal grating and a uniform gray field, that is, the threshold at which a very low-contrast grating stops looking like a uniform gray field and starts to look striped. This threshold is measured for gratings at many different spatial frequencies from low to high. The results can be summarized in a graph in which the contrast sensitivity at threshold is plotted as a function of spatial frequency, as shown in Figure 8.

The CSF shows that people are most sensitive to intermediate spatial frequencies at about 4–5 cycles per degree of visual angle. If the CSF is measured under low-light (scotopic) conditions in humans, sensitivity to all frequencies drops dramatically, especially at the highest frequencies. This means that at night, when just the rods are operating, human vision lacks the high acuity that it has in daylight.

Several measurements on human subjects have shown the selective adaptation of channels. The extended exposure to the grating caused the subject's visual system to *adapt*, that is, to become less sensitive after the prolonged viewing experience (see section 2.2.1), but only near

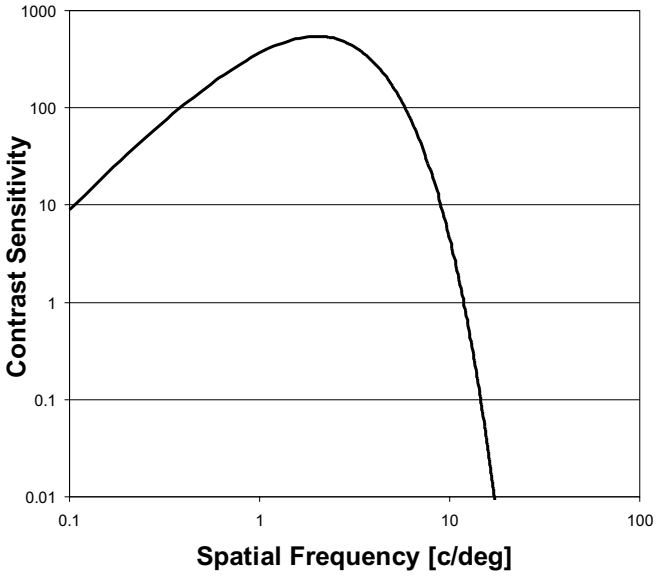

Figure 8: Contrast sensitivity function.

the particular spatial frequency and orientation of the adapting grating. Just as gratings of a particular spatial frequency and orientation produce specific adaptation effects, they also produce specific *aftereffects*.

Local Spatial Frequency Theory

Psychophysical channels are hypothetical mechanisms inferred from behavioral measures rather than directly observed biological mechanisms. If these channels are real, however, they must be implemented somewhere in the visual system. To face this problem there arises second theory about the function of the cells in striate cortex. There is now substantial evidence that these cells may be performing a *local spatial frequency analysis* of incoming images. A local, piecewise, spatial frequency analysis can be accomplished through many small patches of sinusoidal gratings that "fade out" with distance from the center of the receptive field. This sort of receptive field structure, called a *Gabor function*, is constructed by multiplying a global sinusoidal grating by a bell-shaped Gaussian envelope, see the multiscale transforms introduction in the Appendix B.

The degree of frequency tuning in cortical cells seems to fall along a continuum; some are very sharply tuned and others quite broadly tuned. Cells that are tuned to high spatial frequencies have narrower tuning than do cells that are tuned to low spatial frequencies. There is a similar continuum in the degree of orientation tuning; some cells respond only to gratings that are very close to their "favorite" orientation, whereas others respond almost equally to gratings in any orientation. Moreover, cells that are broadly tuned for spatial frequency are also broadly tuned for orientation, and cells that are narrowly tuned for spatial frequency are also narrowly tuned for orientation.

Although the evidence that simple and complex cells in area V1 may be doing a local spatial frequency analysis of input images is impressive, this conclusion is not universally held. Nevertheless, local spatial frequency theory must be counted a very serious alternative to the line and edge detection theory.

Architecture of Image Processing Hypothesis

We have reviewed two theories about the function of the cortical cells. The psychophysical view is that these cells describe images in a piecewise, local Fourier analysis. However, the edge detection theory claims that these same cells are actually the physiological implementation of edge detection mechanisms at different spatial scales. These different views are perhaps not as incompatible as they appear.

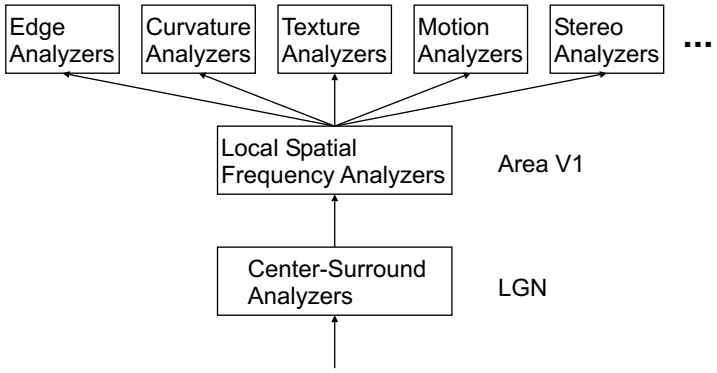

Figure 9: A theoretical hypothesis about the architecture of image processing.

The hypothesis is that local spatial frequency theory and edge detection theory may be appropriate for different levels of the visual system, as diagrammed in Figure 9: center/surround cells in retina and LGN provide input to local spatial frequency analyzers in area V1 of visual cortex, which then project their output to a variety of different modules that compute edges, surface curvatures, textures, stereopsis, and so on, at later stages. According to this view, edge detectors would then be constructed from the output of local spatial frequency analyzers by coding for the output pattern that is characteristic of luminance edges.

3 Perceptually Based Image and Video Quality Metrics

In recent years a lot of effort has been given to the research of incorporation of human perception into the computer graphics and image processing methods. Thanks to this we have seen a big progress in several areas, e.g. in the field of image and video quality assessment. The goal of an objective image or video quality assessment is to develop quantitative measures that can automatically predict perceived image quality [54]. An objective image quality metric can play an important role in a broad range of applications, such as image acquisition, compression, communication, displaying, printing, restoration, enhancement, analysis and watermarking.

In this section we will first outline the general framework of quality metrics. Then we will summarize the perceptually driven image and video quality metrics that are or could be used in computer graphics. Finally, we will outline merits and shortcomings of these techniques.

3.1 General Framework of Perceptual Quality Metrics

A great variety of models has been proposed in the literature. For many of these models, common computational parts can be identified [55]. These parts are: preprocessing, CSF filtering, channel decomposition, error normalization and masking, and finally the error pooling, see Figure 10.

- The *pre-processing* stage may perform alignment, transformations of color spaces, calibration for display devices, point spread function filtering, and light adaptation.

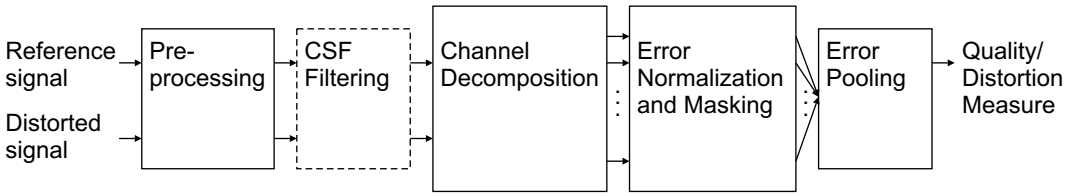

Figure 10: Block diagram of typical discrimination/quality metric.

- *CSF* may be implemented before the channel decomposition using linear filters that approximate the frequency responses of the CSF. Some metrics, on the other side, implement CSF as weighting factors for channels after the channel decomposition.

- *Channel decomposition* is used to model the frequency selective channels in the HVS. The channels serve to separate the visual stimulus into different spatial and temporal *subbands*. During this phase, quality metrics differ mostly in the chosen filters.

- *Error normalization and masking* is typically implemented within each channel. Most models implement masking in the form of a gain-control mechanism that weights the error signal in a channel by a space-varying *visibility threshold* for that channel. The visibility threshold adjustment at a point is calculated based on the energy of the signal in the neighbourhood of that point, as well as the HVS sensitivity for that channel in the absence of masking effects.

- *Error pooling* is the process of combining the error signals in different channels into a single distortion/quality interpretation. The typical implementation uses *Minkowski summation* (also called L_p-norm) on the two sets of channels to compute the model response r:

$$r = \Big(\sum_l \sum_k |e_{l,k}|^\beta \Big)^{1/\beta},$$

  where $e_{l,k}$ is the normalized and masked error of the k-th coefficient in the l-th channel, and β is a constant with a value between 1 and 4.

3.2 Perceptual Image Quality Metrics

Objective image quality metrics serve primarily to assessment of the difference between two images, an original image and a distorted image. They can be classified according to the availability of an original image, with which the distorted image is to be compared. Most existing approaches are known as *full-reference*, meaning that a complete reference image is assumed to be known. In many practical applications, however, the reference image is not available, and a *no-reference* or "blind" quality assessment approach is desirable. In a third type of method, the reference image is only partially available, in the form of a set of extracted features made available as side information to help evaluate the quality of the distorted image. This is referred to as *reduced-reference* quality assessment.

Image quality metrics could be employed not only to image comparison, but also to **acceleration of rendering algorithms**, **perception-guided rendering of animations**, etc., as one may see in Chapter 4. Since the visible differences predictor by Daly is extensively applied in the context of this report, it will be described more thoroughly.

3.2.1 Pixel-Based Metrics

The *mean squared error* (MSE) and the *peak signal-to-noise ratio* (PSNR) are the most popular difference metrics in image and video processing. The MSE is the mean of the squared differences between the gray-level values of pixels in two pictures I and $\bar{I}$:

$$MSE = \frac{1}{XY} \sum_x \sum_y [I(x,y) - \bar{I}(x,y)]^2,$$

for pictures of size $X \times Y$. The average difference per pixel is thus given by the root mean squared error $RMSE = \sqrt{MSE}$.

The PSNR in decibels is defined as:

$$PSNR = 10 \log \frac{m^2}{MSE},$$

where m is the maximum value that a pixel can take (e.g. 255 for 8-bit images). *Color PSNR* is a version of the PSNR that accounts for colors, using perceptually uniform differences. Both MSE and PSNR are well-defined only for luminance information, there is no agreement on the computation of color values.

3.2.2 Model After Mannos and Sakrison

The first perception based image quality metric for luminance images was developed by Mannos and Sakrison [31]. Computation of the proposed model begins by normalizing all the luminance values L_{ij} by the mean luminance L_m. The nonlinearity in perception is accounted for by taking the cubed root of each normalised luminance. A Fast Fourier Transform is computed of the resulting values, and the magnitude of the resulting transform at frequencies in the horizontal and vertical directions (u, v), (where u and v are expressed in terms of cycles per visual degree) is denoted $f_{uv}(||\sqrt[3]{\frac{L}{L_m}}||)$. The magnitudes f_{uv} are then filtered with the CSF $A_M(u,v) = A_M(r)$, where $r = u^2 + v^2$ to account for spatial frequency sensitivity to produce the array of values g_{uv} :

$$A_M(r(u,v)) = 2.6 * [0.0192 + 0.144\sqrt{r}] \exp[-(0.144\sqrt{r})^{1.1}],$$

$$g_{uv} = f_{uv}(||(\frac{L}{L_m})^{0.333}||) * A_M(r(u,v)).$$

Finally, the distance between the two images is computed by finding the Mean Square Error of the values g_{uv} for each of the two images:

$$M(X;Y) = \frac{1}{N} \sum_{all\ u,v} (g_{X,uv} - g_{Y,uv})^2.$$

This technique therefore measures similarity in Fourier amplitude between images. It was shown to correlate quite well with subjective ranking data. Despite its simplicity, this metric was one of of the first works in engineering to recognize the importance of applying vision science to image processing.

3.2.3 Model After Gervais

Another simple model was adapted from a study of confusion between letters of the alphabet [15] by Rushmeier et al. [43]. The model includes the effect of phase as well as magnitude in the frequency space representation of the image. The luminances are normalised by dividing by the mean luminance. An FFT is computed producing an array of phases $p_{uv}(||\frac{L}{L_m}||)$ and

magnitudes $f_{uv}(||\frac{L}{L_m}||)$. The magnitudes are then filtered with an anisotropic CSF filter function constructed by fitting splines to psychophysical data presented by Campbell et al. [5], producing the filtered values $g_{uv}(||\frac{L}{L_m}||)$. The distance between images is then computed using:

$$M(X;Y) = \frac{1}{N} \sum_{all\ u,v} (((\log g_{X,uv} + 1) - (\log g_{Y,uv} + 1))(1 + p_{X,uv} - p_{Y,uv}))^2.$$

Since the Gervais model include phase (i. e. pixel position) information, its performance suffers due to subjectively minor registration problems between images. However, in situations where geometric alignment is not a problem, or is of critical importance for some other reason, this model may actually outperform the others.

3.2.4 Visible Differences Predictor

The Visible Differences Predictor [8] (VDP) is one of the best-known image distortion metrics. The VDP model interprets early vision behavior, from retinal contrast sensitivity to spatial masking. Figure 11 shows the use of the VDP, which consists of three main stages: components for *calibration* of the input images, a *human visual system (HVS) model* and a method for *displaying* the HVS visible *differences*. The input to the algorithm includes two images and parameters for viewing conditions, whereas the output is a map describing the visible differences between them (see Figure 22 on the page 31 for an example). The output map defines the probability of detecting the differences between the two images as a function of their location in the images. This metric, *probability of detection*, provides a description of the threshold behavior of vision but does not discriminate among different suprathreshold visual errors.

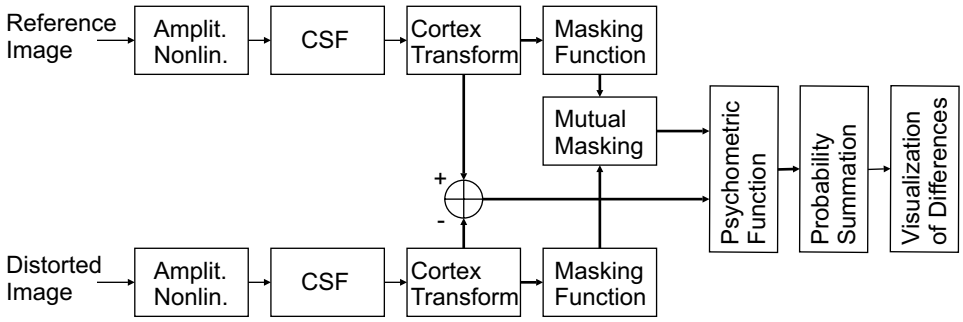

Figure 11: Block diagram of the Visible Differences Predictor (heavy lines indicate parallel processing).

Calibration

Firstly, the input images represented by unitless digital numbers are calibrated. The calibrating input parameters are: the *viewing distance* for which the VDP will make its visual prediction, and the *physical pixel spacings*, which along the viewing distance map the visual frequencies expressed in cycles per degree (c/deg) to frequencies expressed digitally as a fraction of the Nyquist frequency.

Human Visual System Model

The human visual system model is the key element of the VDP. It concatenates on the lower-order processing of the visual system, such as the optics, retina, lateral geniculate nucleus, and striate cortex. The HVS model consists of a number of processes that limit visual sensitivity. Three main sensitivity variations are accounted for, namely, as a function of *light*

level, *spatial frequency*, and *signal content*. Sensitivity S is defined as the inverse of the contrast C_T required to produce a threshold response, $S = 1/C_T$, where contrast is defined as $C = (L_{max} - L_{mean})/L_{mean}$, where L_{max} and L_{mean} refer to the maximum and mean luminances.

The variations in sensitivity as a function of light level are simulated by **amplitude nonlinearity**. Each input luminance L_{ij} is transformed by simplified version of the retinal response to an "amplitude non-linearity value" b_{ij} defined as: $b_{ij} = L_{ij}/(L_{ij} + 12.6L_{ij}^{0.63})$, where the constants 12.6 and 0.63 apply when luminance is expressed in cd/m^2. For this model the adaptation level for an image pixel is solely determined from that pixel.

The variations as a function of spatial frequency are modeled by the **contrast sensitivity function**, implemented as a filtering process. A Fast Fourier transform (FFT) is applied to the values b_{ij}. The resulting magnitudes, $f_{uv}(b)$ are filtered by a CSF which is a function of the image size in degrees and light adaptation level L_m. The resulting contrast sensitivity filter $A_D(r(u,v))$ is given by:

$$A_D(r(u,v)) = (0.008/r^{1.5} + 1)^{-0.2} 1.42\sqrt{r}\exp(-0.3\sqrt{r})\sqrt{(1 + 0.06\exp(0.3\sqrt{r})},$$

where $r = u^2 + v^2$.

The variations in sensitivity due to a signal content are reffered to as *masking*. Masking effects are modeled by the **detection mechanism**, which is the most complicated element of the VDP. It consists of four subcomponents: image channeling, spatial masking, psychometric function, and probability summation. *Image channeling* involves a decomposition similar to the *cortex transform* introduced by Watson [56]. Cortex transform is a multi-resolution pyramid (see Appendix A) that simulates the spatial-frequency and orientation tuning of simple cells in the primary visual cortex (see Section 2.3). During the image channeling stage, the input image is fanned out from one channel to 31 channels or bands as follows. Each channel is associated with one cortex filter which consists of a *radial filter* (*dom*, difference of mesa filter) and an *orientational filter* (*fan* filter). The total number of radial filters is six resulting in five frequency bands and one base band. Each of these bands except for the base band is further fanned out into six channels of different orientation, see Figure 12. Thus five frequency bands times six orientations per bands plus one base band results in 31 channels.

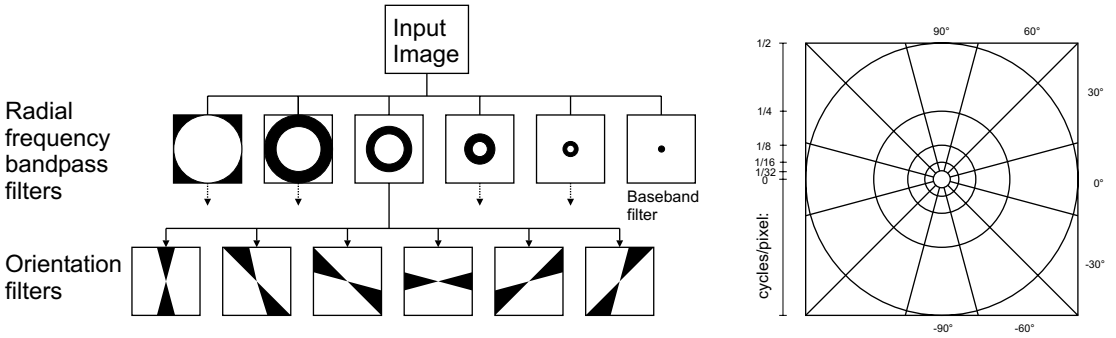

Figure 12: Cortex transform. On the left: organization of the filter bank. On the right: decomposition of the image frequency plane into the radial and orientation selectivity channels.

Spatial masking reduces the detectability of a given stimulus through the simultaneous presence of an additional suprathreshold stimulus. The masking depends on several factors, such as mutual masking, learning effects, the nature of masking signal, etc. Due to visual masking, threshold values can be elevated. This is accounted for after the transformation of all channels back to the spatial domain. For every channel and for every pixel, the elevation of the

detection threshold is calculated based on the mask contrast for that channel and that pixel. Mutual masking can be considered by taking the minimal threshold elevation value for the corresponding channels and pixels of the two input images.

Psychometric function estimates the probability of detecting the differences for a given channel. The applied psychometric function describes the increase in the probability of detection as the signal contrast increases. Once the detection probabilities have been computed for each band of the filter hierarchy, the probability images are combined into single image by pooling together probability contributions from all bands as a function of position.

Difference Visualization

There are two ways for visualizing the VDP output. The first technique is the *free-field difference map*, where the visible difference predictions appear on a uniform field with a gray value near the system mean. The second method, the *in-context difference map*, is the mapping of the output probabilities in color on the reference image. It is assumed that the difference can be perceived for a given pixel when the probability value is greater than 0.75, which is a standard threshold value for discrimination tasks.

3.2.5 Perceptual Distortion Measure by Teo and Heeger

Teo and Heeger [49] presented a perceptual distortion measure based on the so-called *normalisation model* – the nonlinear model of early phases of human vision. The model fits empirical measurements of the response properties of neurons in the primary visual cortext (see Section 2.1.3), and the psychophysics of spatial pattern detection. In the primary visual cortex, a so-called *contrast gain control* mechanism keeps natural responses within the permissible dynamic range while at the same time retaining global pattern information. In the metric, contrast gain control is realized by an excitatory nonlinearity that is inhibited divisively by a pool of responses from other neurons. The channel decomposition process uses quadrature steerable filters with six orientation levels and four spatial resolutions. The distortion measure is computed from the resulting normalized responses by a simple squared-error norm to produce the difference map, similar to one produced by the VDP. Masking is modeled through contrast normalization and response saturation.

Authors in the paper [49] thoroughly demonstrate that the proposed measure is far better than the MSE and illustrate the usefulness of the model in measuring perceptual distortion in real images.

3.2.6 Visual Discrimination Model

The Sarnoff Visual Discrimination Model [29] (VDM) is another image discrimination metric. The overall structure of the model is outlined in Figure 13. The VDM operates in the spatial domain. First, the inputs are convolved with an approximation of the point spread function of the eye's optics. The signals are then re-sampled to reflect the photoreceptor sampling in the retina. A Laplacian pyramid [4] (see Appendix A) is used to decompose the images into seven resolutions (each resolution is one-half of the immediately higher one), followed by band-limited contrast calculations. A set of orientation filters implemented through steerable filters of Freeman and Adelson [14] is then applied for orientation selectivity in four orientations. The CSF is modeled by normalizing the output of each frequency-selective channel by the base-sensitivity for that channel. Masking is implemented through a sigmoid non-linearity, after which the errors are convolved with disk-shaped kernels at each level. Finally, a distance measure or JND map is computed as the L_p-norm (Minkowski summation) of the masked responses.

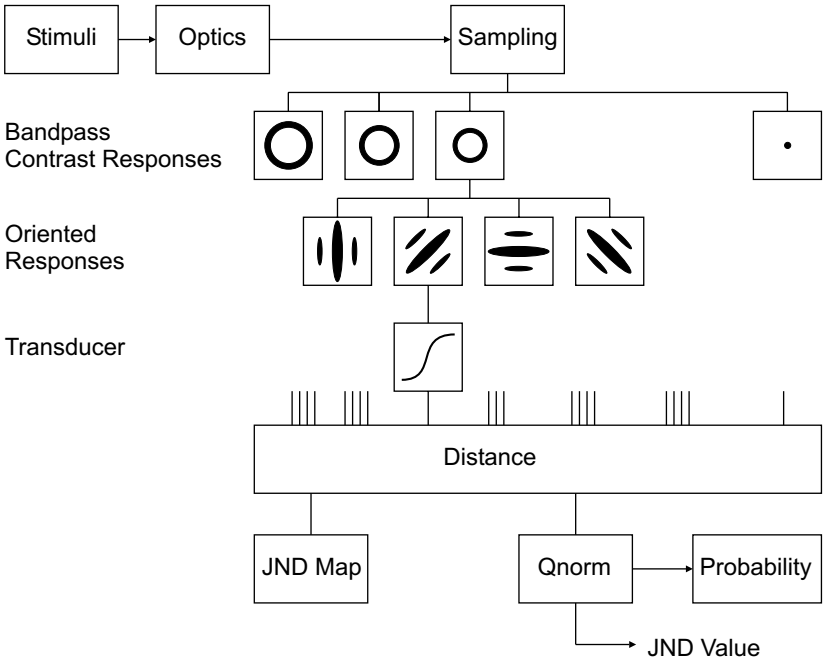

Figure 13: Block diagram of the Visual Discrimination Model.

The VDM is one of the few models that take into account the eccentricity of the images in the observer's visual field. It was later modified to the Sarnoff JND metric for color video [57]. The complexity of Sarnoff VDM is $O(N)$, because the VDM operates in the spatial domain and avoids expensive FFT and FFT^{-1} transformations which take up to 40% of the execution time in the Daly VDP.

3.2.7 Gabor Pyramid Model of the HVS

Taylor et al. [48] presented a Gabor pyramid-based model of the human visual system (HVS) for image quality assessment. Their model departs from previous approaches in three ways:

- a physiologically and psychophysically plausible Gabor pyramid is used to model a receptive field decomposition
- psychophysical experiments are involved to directly assess the percept to be modeled
- the discrimination performance is modeled by using discrimination thresholds instead of detection thresholds.

A number of physiological studies have confirmed the hypothesis that mammalian visual systems contain neurons whose receptive fields closely resemble Gabor patches [48]. Because of the physiological and psychological plausibility of Gabor decomposition, the proposed model involves a *Gabor pyramid*.

The model accepts two grayscale images as inputs and generates a probability map as output. A block diagram of the model is shown in Figure 14. A multiresolution decomposition is performed on each image to generate a number of channels, each containing the response of an ensemble of visual receptors. The receptors are modeled by Gabor functions of varying frequency and orientation. The multiresolution pyramid is built by lowpass filtering and decimating the original image, see Appendix A. Each output image for a particular pyramid level is called *base image*. The base image for each pyramid level is convolved with even and odd

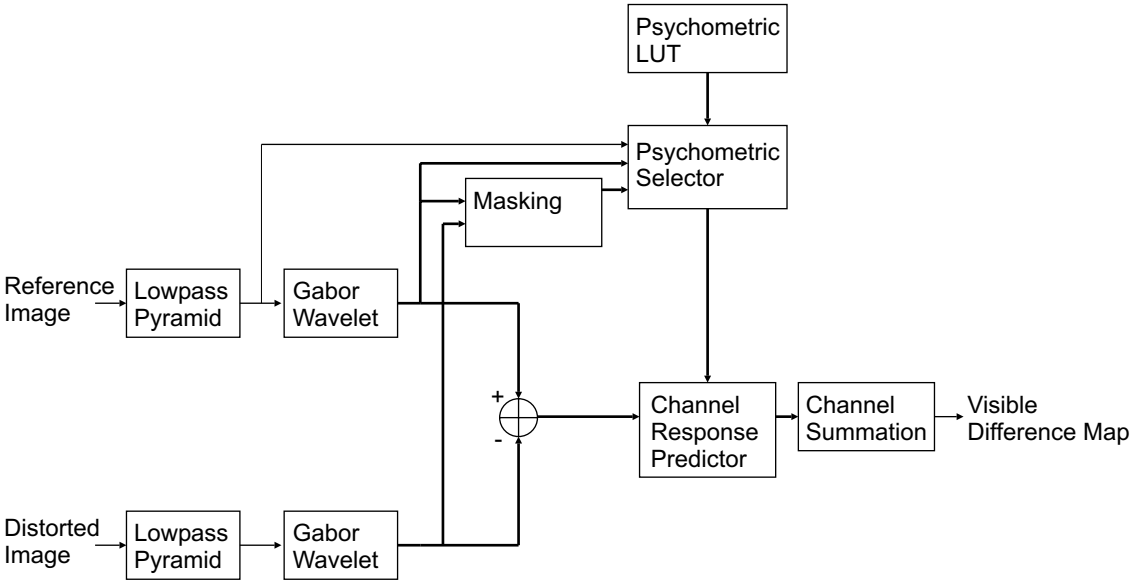

Figure 14: Block diagram of the Gabor Pyramid Model.

symmetric Gabor wavelets at eight orientations. The square root of the sum of the squares of the resulting even-odd image pairs describes the response of an ensemble of neurons tuned to a particular spatial frequency and orientation. These images are called the *channel images*.

The Psychometric Look Up Table (LUT) consists of a family of psychometric functions that have been empirically determined by psychophysical experiments described below. The Psychometric Selector selects the appropriate psychometric function from the family of psychometric functions in the Psychometric LUT. A higher pyramid level base image determines the adaptation level, and the channel image determines the frequency, orientation, and reference contrast levels used to select the appropriate psychometric function. The difference between the contrast images for each channel is then applied to the appropriate psychometric function to produce a separate probability map for each channel. All of the probability maps from the different channels are combined using probability summation.

Two psychophysical experiments were conducted to determine the parameters of the model. The first experiment tested the visual system's sensitivity to Gabor patches as a function of spatial frequency, orientation, and average luminance. The second experiment compared the relation between detection and discrimination thresholds.

3.2.8 Wavelet Visible Difference Predictor

Bradley's [3] wavelet visible difference predictor (WVDP) is largely based on previously mentioned (see Section 3.2.4) visible differences predictor, but has a number of modifications that make it more amenable to potential integration into a wavelet based image comparison scheme. These modifications include the use of a *separable wavelet transform* instead of the cortex transform, the application of a *wavelet contrast sensitivity function,* and a simplified definition of sub-band contrast that allows prediction of noise visibility directly from wavelet coefficients, see Figure 15.

Another wavelet based metric has been proposed by Lai and Kuo [26]. Their metric is based on the Haar Wavelet and the masking model can account for channel interactions as well as suprathreshold effects.

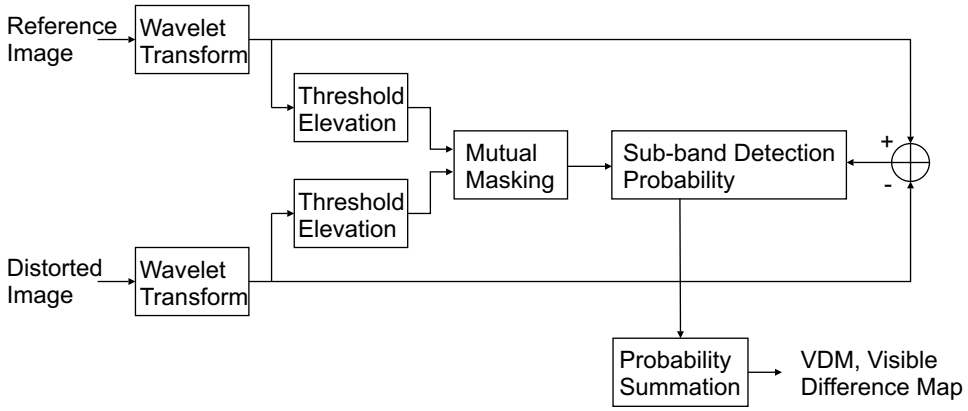

Figure 15: Block diagram of the Wavelet Visible Differences Predictor.

3.2.9 Multistage Perceptual Quality Assessment Model

Multistage perceptual quality assessment model [38] (MPQA) was proposed to compare original and lossy compressed digital angiogram images. As shown in Figure 16, the MPQA model includes amplitude nonlinearity, octave bandwidth spatial frequency decompositions into six orientations using Watson's cortex transformation [56], and contrast masking based on CSF modeling and region classification from the decomposed images. A perceptual distortion visibility map (PDVM) is produced via a distance computation and summation of efforts across different spatial frequency bands. A perceptual quality rating (PQR) is then calculated from the PDVM converting fidelity to quality, and transformed into a one to five scale, PQR_{1-5}.

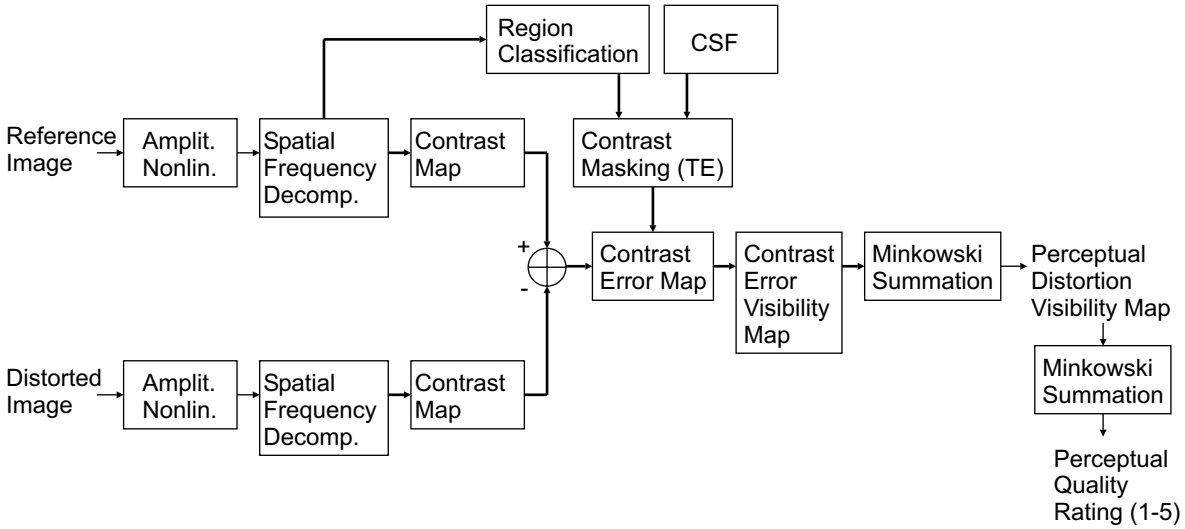

Figure 16: Block diagram of the MPQA model.

As one may notice, the MPQA model is based on previously mentioned quality assessment models (mainly on the VDP, see Section 3.2.4), however it differs in the inclusion of contrast masking as a function of *background uncertainty*. The human eye can tolerate larger errors in high uncertainty simuli (i.e. in textured areas) than in low uncertainty stimuli of the same contrast (i.e. along edges). The spatially decomposed images are therefore classified into *flat*, *edge*, and *texture* regions to consider the relationship between stimulus and background uncertainty. Flat regions are the areas with lower contrast than the base threshold contrast

given by CSF. Edge areas are detected using a Sobel edge detector. Remaining regions are classified as texture regions. The threshold values are then elevated accordingly to the area type.

3.2.10 Metro: measuring error on simplified surfaces

Metro [6] is a tool that allows one to compare the difference between a pair of *surfaces* (e.g. a triangulated mesh and its simplified representation) by adopting a surface sampling approach. It has been designated as a highly general tool, and it does no assumption on the particular approach used to build the mesh representation. It returns both numerical results (meshes areas and volumes, maximum and mean error, etc.) and visual results, by coloring the input surface according to the approximation error.

Metro evaluates the difference between two meshes S_1 and S_2, on the basis of the *approximation error* measure. The approximation error between two meshes is defined as the distance between corresponding sections of the meshes. Given a point p and a surface S, the distance $e(p, S)$ is defined as:

$$e(p, S) = \min_{p' \in S} d(p, p'),$$

where $d()$ is the Euclidean distance between two points in E^3. The one-sided distance between two surfaces S_1, S_2 is then defined as:

$$E(S_1, S_2) = \max_{p \in S_1} e(p, S_2).$$

Given a set of uniformly sampled distances, the mean distance E_m between two surfaces is defined as the surface integral of the distance divided by the area of S_1:

$$E_m(S_1, S_2) = \frac{1}{|S_1|} \int_{S_1} e(p, S_2) ds.$$

The error is evaluated by scan converting the first mesh faces with a user-specified sampling step, and computing a point-to-surface distance for each scan-converted point. The mean and maximum distances between meshes are returned.

Although Metro does not explicitly utilize any human visual system properties in the computation, it has been used in several psychophysical experiments. Metro v.2 is available as public domain software at the *Visual Computing Group* web site of the CNUCE and IEI, C.N.R. Institutes at Pisa (`http://miles.cnuce.cnr.it/cg/metro.html`).

3.3 Video Quality Metrics

Assessment of video quality in terms of artifacts visible to the human observer is becoming very important in various applications dealing with digital video encoding, transmission, compression techniques, and computer graphics. Subjective video quality measurement is costly and time-consuming, and requires many human viewers to obtain statistically meaningful results. Several video quality metrics have been developed to face this problem.

Same as perceptual image quality metrics, the perceptual video quality metrics can save time by elimination of subjective testing. Moreover these metrics can help a lot when optimizing the storage space or download times of video clips. Applications of video quality metrics include:

- video encoder tuning and optimization,
- video security and watermarking,
- video quality monitoring.

In this section we will describe various artifacts that can occur in a video sequence. Consecutively, we will give a brief overview on perceptually based video quality metrics.

3.3.1 Artifacts

We can distinguish a variety of artifacts in a video seguence [57]. Some of them may be caused by the compression algorithm, while the others occur as a consequence of transmission errors and various video conversions.

- *Blockiness* or the *blocking effect* refers to a block pattern in the compressed sequence. It is due to the independent quantization of individual blocks (usually 8×8 pixels) in block-based DCT coding schemes.
- *Blur* is characterised by the loss of fine detail and the smearing of edges in the video. It is typically caused by a high-frequency attenuation at some stage of the recording or encoding process. Wavelet-based encoders also cause blurry artifacts.
- *Flickering* appears when a scene has high texture contrast. Texture blocks are compressed with variyng quantization factors over time, which results in a visible flickering effect.
- *Color bleeding* is the smearing of the color between areas of strongly differing chrominance. It results from the suppression of high-frequency coefficients of the chroma components. Due to chroma subsampling, color bleeding extends over an entire block.
- *Aliasing* can be noticed when the content of the scene is above the Nyquist rate, either spatially or temporally.
- *Mosquito noise* is a temporal artifact seen mainly in smoothly textured regions as luminance/chrominance fluctuations around high-contrast edges or moving objects. It is a consequence of the varied coding of the same area of a scene in consecutive frames of a sequence.
- When transporting media over noisy channels *packet loss* or *packet delay* can occur. Such losses or delays can affect both the semantics and the syntax of the media stream.

A survey of video coding distortions can be found in an article by Yuen and Wu [58].

3.3.2 VQM by Lukas and Budrikis

The first video quality metric was developed by Lukas and Budrikis [30]. It is based on a spatio-temporal model of the contrast sensitivity function using an excitatory and an inhibitory path. The two paths are combined in a nonlinear way, enabling the model to adapt to changes in the level of background luminance. Masking is also incorporated in the model by means of a weighting function derived from the spatial and temporal activity in the reference sequence. In the final stage of the metric, an L_p-norm (Minkowski summation) of the masked error signal is computed over blocks in the frame whose size is chosen such that each block covers the size of the foveal field of vision. The resulting distortion measure was shown to outperform MSE as a predictor of perceived quality.

3.3.3 ST-CIELAB

Tong et al. [50] proposed a single-channel video quality metric called *ST-CIELAB* (spatio-temporal CIELAB). ST-CIELAB is an extension of the spatial CIELAB (S-CIELAB) image quality metric [59]. Both are backward compatible to the CIELAB standard, i.e. they reduce to CIE $L^*a^*b^*$ for uniform color fields. The ST-CIELAB metric is based on a spatial, temporal, and chromatic model of human contrast sensitivity. The contrast sensitivity is modeled in an *opponent color space*, where the color information is encoded as white-black, red-green and blue-yellow color difference signals. After the CSF modeling, the data are transformed to CIE $L^*a^*b^*$ space, whose difference formula is used for pooling.

3.3.4 Moving Picture Quality Metric

The *Moving Picture Quality Metric* (MPQM) proposed by Lambrecht [51] is a multichannel video quality model. It is based on a local contrast definition and Gabor-related (see Apendix B) filters for the spatial decomposition, two temporal mechanisms, as well as a spatio-temporal contrast sensitivity function and a simple intra-channel model for contrast masking. A color version of the MPQM based on an *opponent color space* was presented as well as a variety of appplications and extensions of the MPQM, e.g. for assessing the quality of certain image features such as contours, textures, and blocking artifacts, or the study of motion rendition.

Due to the MPQM's purely frequency-domain implementation of the spatio-temporal filtering process and the resulting huge memory requirements, it is not practical for measuring the quality of sequences with a duration of more than a few seconds, however. The *Normalization Video Fidelity Metric* (NVFM) [27] avoids this shortcoming by using a steerable pyramid transform for spatial filtering and discrete time-domain filter approximations of the temporal mechanisms. It is a spatio-temporal extension of Teo and Heeger's image distortion metric (see Section 3.2.5) and implements inter-channel masking through an early model of contrast gain control.

3.3.5 Perceptual Distortion Metric

Perceptual Distortion Metric (PDM), presented by Winkler [57], is a metric for both the digital color images and video. It is based on a contrast gain model of the HVS, which takes into account color perception, the multi-channel architecture of temporal and spatial mechanisms, spatio-temporal contrast sensitivity, pattern masking and channel interactions, see Figure 17.

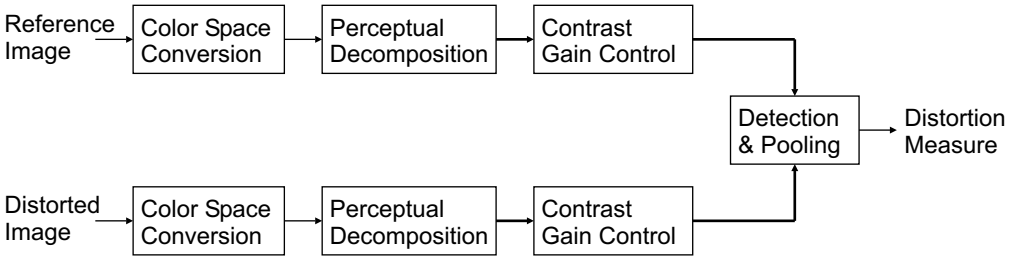

Figure 17: Block diagram of the PDM model.

The metric requires both the reference sequence and the distorted sequence as inputs. Both these sequences are converted to the *opponent color space*, where the color information is encoded as white-black, red-green and blue-yellow color difference signals. After the conversion, each of the resulting three color components is subjected to a spatio-temporal filter bank decomposition. Temporal mechanisms are simulated by the temporal low-pass filter and by the

band-pass filter. The decomposition in the spatial domain is carried out by means of the *steerable pyramid transform* [23]. This transform has the advantage of being rotation-invariant and self-inverting while minimizing the amount of aliasing in the subbands. In the case of the PDM, five subband levels with four orientation bands each plus one low-pass band are computed. These perceptual channels are weighted according to contrast sensitivity and subsequently undergo *contrast gain control* for pattern masking. Finally, the sensor differences are combined into a distortion measure. The PDM was shown to accurately fit to psychophysical contrast sensitivity and contrast masking data and to behave consistently with human observation.

3.3.6 Non-Perceptual Metrics

Metrics based on multi-channel vision models are the most general and potentially the most accurate ones. However, quality metrics need not necessarily rely on sophisticated general models of the HVS; they can for example exploit a priori knowledge about the compression algorithm and the pertinent types of artifacts using ad-hoc techniques or specialized vision models. While such metrics are not as versatile, they normally perform well in a given application area. Their main advantage lies in the fact that they often permit a computationally more efficient implementation. However, even some more universal non-perceptual metrics exist, two examples of such metrics are the *Universal Image Quality Index* [53], and the recenty published *Multidimensional IQM using Singular Value Decomposition* [44], see the thesis by S. Winkler [57] for an overview.

3.4 Conclusions and Future Research

As we have seen, significant number of human vision models that could be applied to computer graphics issues exists. However, particularly in the field of image comparison and image and video quality assessment, there is still a lot of work to do.

- Most of the above mentioned approaches incorporate only a few of many factors that influence human perception, i.e. typically just the luminance adaptation and the contrast sensitivity. Nevertheless, the perception of *color* by human visual system is a very important task, for example it helps us to define the shapes in a perceived scene. Since many of mentioned approaches have only been applied to binary or greyscale images, little consideration has been given to the analysis of full-color images. Even in the field of image segmentation, the development of segmentation algorithms based on color is still incipient [7].

- Almost all of mentioned models are severely sensitive to shifts, dilatations, rotations, etc., inbetween the pair of input images. However, this does not correspond to the human practice, where such a differences are not held so strictly.

- As most of perceptual image quality metrics are relative (full-reference) – they take the reference image and the examined image as inputs, there is a lack of *absolute* (no-reference) perceptual picture quality metrics.

- Contemporary perceptual image metrics typically consider the image as a whole. Therefore, there is a lot of work in comparison of images with *regions of interest* (ROI), i.e. the JPEG 2000 [47] ROI compressed images.

- What does the RMS error mean for 2D comparison, the METRO measure is in 3D. However, not as much attention as to 2D perceptual comparison was given to perceptual comparison of 3D models. Appropriate 3D error metric should involve HVS properties.

- We suppose that more work could be done to extent or replace existing perceptual comparison techniques, to produce a more perceptually robust solution even for *not* strictly photorealistic images.
- Nowadays, there are no agreed-upon standards for measuring the *realism* of computer-generated images [12]. Sometimes physical accuracy is used as a criterion, at other times perceptual criteria are applied, and under many conditions an ad-hoc "looks-good" standard is used.
- Since the amount of contemporary quality metrics is considerable, evaluation and comparison of them becomes an important issue. However, only a few *comparative studies* exist that have investigated the prediction accuracy of metrics in relation to others.
- So far, no *general-purpose metric* has been found that is able to replace subjective testing.

4 Perceptually Accelerated Rendering

Having in mind various approaches to modeling of human visual system, we can now explore how human perception can be utilized to improve the performance of computer graphics rendering methods. Principally, we can distinguish two types of approaches to acceleration of rendering. The first possibility is to embed characteristics of human visual system directly into algorithms, while the second option is to apply them on rendered images to drive the computation effectively. This is an area where perceptually based quality metrics, summarized in Section 3, though usually in a simplified form, take a place.

4.1 Embedding HVS Characteristics Directly Into Algorithms

A straightforward way to improve rendering computations is to directly embed some of simple perceptually-based error metrics to the light interactions between surfaces. Such a metric is evaluated densely during the computation so that it should be very simple and effective.

4.1.1 Perceptually-Driven Radiosity

Gibson and Hubbold [16] proposed a *perception-driven hierarchical radiosity* algorithm in which a tone mapping operator and the perceptually uniform color space CIE $L^*u^*v^*$ are used to decide when to stop the hierarchy refinement. Links between patches are not further refined once the difference between successive levels of elements becomes unlikely to be detected perceptually. A similar error metric was applied to measure the perceptual impact of the energy transfer between two interacting patches, and to decide upon the number of shadow rays that should be used in a visibility test for these patches.

Martin et al. [33] proposed a similar strategy as Gibson and Hubbold. They use an oracle of patch refinement that operates directly in the image space and tries to improve the radiosity-based image quality for a given view.

4.2 Perceptual Metrics Operating on Rendered Images

All the techniques discussed in previous section used perceptual error metrics at the atomic level, causing a significant overhead on procedures that are repeated densely. This imposes severe limitations on the complexity of human spatial vision models, which in practice are restricted to models of brightness and contrast perception [36]. Recently, more complete vision models have been used in rendering to develop higher level perceptual error metrics which operate on the complete images.

However, the computation of complex vision models is typically very time-consuming. We must therefore carefully consider whether the savings in computation that we obtain can compensate this cost. For an extensive discussion of this issue in the context of various radiosity methods, look at the thesis by J. Přikryl [40].

4.2.1 Perceptually Based Adaptive Sampling Algorithm by Bolin and Meyer

Bolin and Meyer [2] developed a perceptually based approach for selecting image samples. They used the Sarnoff *Visual Discrimination Model* (described in Section 3.2.6) that has been extended to handle color and has been simplified to run efficiently, see Figure 18.

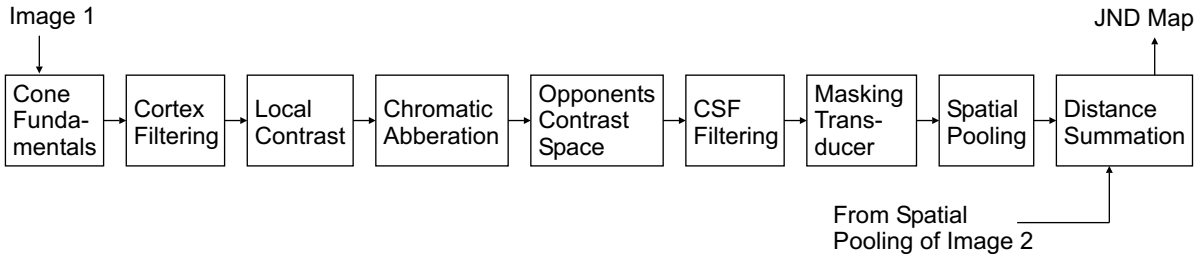

Figure 18: Block diagram of the vision model proposed by Bolin and Meyer.

The resulting new image quality model was inserted into an image synthesis program by first modifying the rendering algorithm so that it computed a wavelet representation. In addition to allowing image quality to be determined as the image was generated, the wavelet representation made it possible to use statistical information about the spatial frequency distribution of natural images to estimate values that were yet to be taken. The image quality model was also used to decide when enough samples had been taken across the entire image, providing a visual stopping condition.

4.2.2 Perceptually Based Physical Error Metric for Realistic Image Synthesis by Ramasubramanian et al.

As we mentioned before, in some cases the cost of recomputing the vision model may cancel the savings gained by employing the perceptual error metric to speed up the rendering algorithm. To combat this Ramasubramanian et al. [41] introduced a threshold model that handles the luminance-dependent processing and spatially-dependent processing *independently*, allowing the expensive spatially-dependent component to be precomputed, see Figure 19.

The model for a given image produces a *threshold map* that predicts the maximum tolerable luminance error. The luminance-dependent processing computes a starting threshold map ΔL_{tvi} for the luminance distribution using the threshold-vs-intensity (TVI) function. The spatially-dependent processing computes a map containing elevation factors $F_{spatial}$ for the spatial pattern using the CSF and masking function. From these two maps the final threshold map $\Delta L_T(x, y)$ is derived as:

$$\Delta L_T(x, y) = \Delta L_{tvi}(x, y) \times F_{spatial}(x, y).$$

The employed spatial decomposition is the same as in the visual discrimination model by Lubin (see Section 3.2.6), however the visual masking model ignore spatial orientation channels.

The computed threshold map is used to predict the sensitivity of the HVS to noise in the indirect ligting component of global illumination simulation. This enables a reduction in the number of samples needed in areas of an image with high frequency texture patterns, geometric details, and direct lighting variations, giving a significant speedup in computation.

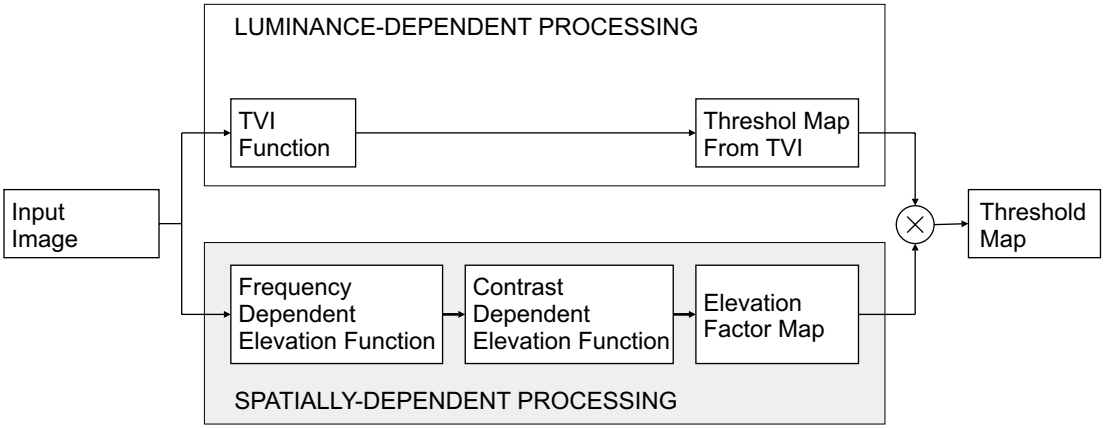

Figure 19: Block diagram of the threshold model by Ramasubramanian et al.

4.3 VDP Applications

Visible differences predictor by Daly (described in Section 3.2.4) was shown [34, 32] to be valuable for such tasks as adaptive meshing performance, accuracy of shadow reconstruction, convergence of the solution of illumination for indirect lighting, and so on. In this section we will outline some of these applications.

4.3.1 Perceptual Convergence of Global Illumination Algorithms

Global illumination algorithms have different performance at different stages of computation. Myszkowski [34] used the VDP to provide the quantitative measures of *perceptual convergence* by predicting the perceivable differences between the intermediate and final images. Following view-independent algorithms were investigated: deterministic direct lighting (DDL), shooting iteration hierarchical radiosity (SHR), and density estimation photon tracing (DEPT) – direct (DDEPT) and indirect (IDEPT) [34]. The performance of these basic techniques was measured using the VDP.

The reported results show that the perceptual convergence of the indirect lighting solution for the SHR technique is slower than the IDEPT approach. The SHR technique shows better performance for simple scenes only. Moreover, at initial stages of computation, the DEPT technique provides the best performance, and rapidly gives meaningful feedback to the user. At later stages, the DDL+IDEPT hybrid shows faster perceptual convergence to the final image. These results confirm that the examined algorithms have a different performance at different stages of computation.

4.3.2 Hybrid Approach to Global Illumination

Based on the results described in previous section Volevich et al. [52] proposed a hybrid technique that uses DDEPT, IDEPT and DDL algoritms. The proposed technique aims to minimize the perceived differences between the intermediate and final images as a function of time by switching to the best candidate algorithm at every stage of computation. The switchover points between the sequentially executed algorithms could not be measured on-line using the VDP because of its enormous computational demands.

To overcome this problem a robust heuristic was proposed: for the sake of simplicity only two switchover points T_1 and T_2 were used. The whole technique is summarized as follows:

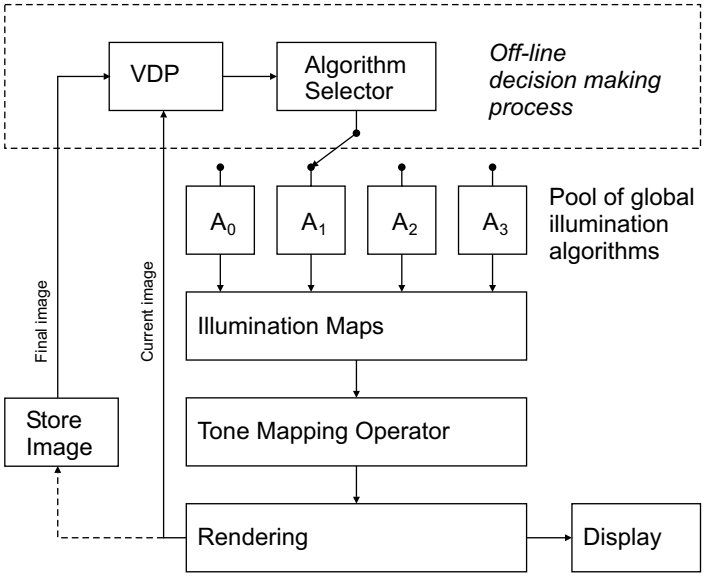

Figure 20: Block diagram of the experimental setting for evaluation of the image quality progression.

- The DEPT algorithm is used in the time interval $[0, T_1]$.
- The DDL algorithm is switched to at time T_1.
- When the DDL computation have completed, the IDEPT algorithm is switched to refine the indirect solution.

As the T_2 point is obtained automatically by completing the DDL algorithm computation, the only thing to assess is the T_1 switchpoint. An experimental setting that was used to solve this problem is outlined in Figure 20. Using this approach, the following solution was proposed: first, the DEPT is run for time $T_\alpha = 0.1\ T_{i0}$, where T_{i0} is the computation time of the first iteration of the DDL. Than, the RMS error $\tilde{E}$ of the indirect lighting simulation is estimated. Finally, the required computation time T_{thr} to reach the accuracy level $E_{thr} \approx 15\%$ is estimated as $T_{thr} = T_\alpha \frac{\tilde{E}^2}{E_{thr}^2}$, and the T_1 is set to $T_1 = \min(T_{thr}, T_{i0})$. The heuristic was shown to provide stable progressive refinement of rendered image quality.

4.3.3 Stopping Conditions for Global Illumination Computation

In algorithms that produce intermediate results rapidly (i.e. the hybrid approach in previous section) we are forced to use a HVS model off-line. However, for applications which require substantial computation time, embedding advanced HVS model might be profitable [36]. Myszkowski [34] used the VDP to decide the stopping conditions for global illumination computations. Based on experimental practice it is assumed that the computation can be stopped if the VDP does not report significant differences between intermediate images. The problem is how to select an appropriate intermediate image which should be compared against the current image to get robust stopping conditions.

The following solution to this problem was proposed: let the current image obtained after the computation time T is $\Im_T$ and the VDP response for a pair of images $\Im_T$ and $\Im_{\alpha T}$ is $VDP(\Im_T, \Im_{\alpha T})$, where $0 < \alpha < 1$. Then we should find an α to get reasonable match between $VDP(\Im_T, \Im_{\alpha T})$ and $VDP(\Im_C, \Im_T)$, where $\Im_C$ is an image for fully converged solution. It was experimentally found that $\alpha = 0.5$ provides a tight upper bound on the estimate of $VDP(\Im_C,$

$\Im_T$). Therefore, if $VDP(\Im_T, \Im_{0.5\ T})$ is less than a specified threshold then we can stop the computation.

4.4 Perception-driven Rendering of Animations

In this section we briefly summarize the Animation Quality Metric (AQM). The AQM is a perceptual animation quality metric that was proposed and successfully applied to acceleration of rendering of high-quality walkthrough animations. That is why we describe it here, although it can certainly be used also to the assessment of video quality, which is concerned in Section 3.3.

4.4.1 Animation Quality Metric

Animation Quality Metric is a metric of animated sequence quality, which is specifically tuned for *synthetic animation sequences* [36, 37]. Two comparison animation sequences are provided as input to the AQM. For every pair of input frames the probability map P_{Map} of perceiving the differences between these frames is generated as output. P_{Map} provides for all pixels the probability values, which are calibrated in such a way that 1 Just Noticeable Differences (JND) unit corresponds to a 75% probability that an observer can perceive the difference between the corresponding image regions. The AQM takes into account the following characteristics of the Human Visual System: the Weber's law [39], spatio-velocity Contrast Sensitivity Function [9], and visual masking. The spatio-velocity CSF requires the velocity value for every pixel, which is aquired from the Pixel Flow (PF). The PF is computed for the previous and following frames along the animation path in respect to the input frame, see Figure 21.

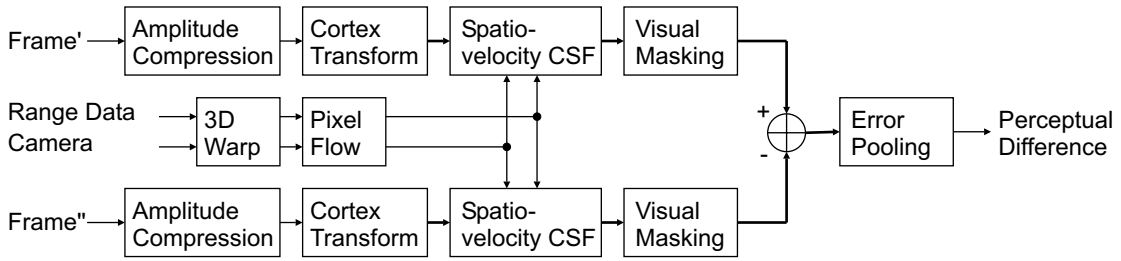

Figure 21: Block diagram of the Animation Quality Metric.

The AQM was used to steer the global illumination computation in dynamic environments for high-quality animation rendering. The global illumination solution was based on stochastic photon tracing and density estimation techniques. A locally operating energy-based error metric was used to prevent photon processing in the temporal domain for the scene regions in which lighting distribution changes rapidly. A perception-based error metric computed by the AQM was used to keep noise inherent in stochastic methods below the sensitivity level of the human observer. As a result a perceptually-consistent quality across all animation frames was obtained. Furthermore, the computation cost was reduced compared to the traditional approaches operating solely in the spatial domain.

4.5 Conclusions

In this chapter we have focused on embedding the characteristics of the HVS directly into various rendering algorithms to improve their efficiency. We have outlined various examples of such an applications. Particularly the global illumination computations may gain much by focusing computation on those scene features which can be perceived by human observer. Nevertheless, we are still unable to perform global illumination computations at interactive

frame rates. We suppose that the informed incorporation of human perception knowledge may further help to achieve this appealing, but so difficult goal. There are several promising directions of further research in the field:

- Creation of special vision models for shapes and materials.
- Finding the way how to differentiate between acceptable and disturbing rendering errors.
- Development of task based metrics tuned to a specific rendering algorithm.

5 Our Effort: Comparing Image-Processing Operators by Means of the Visible Differences Predictor

Utilization of non-photorealistic techniques (NPR), see Section 5.1 for an overview, is beneficial in many cases, in comparison with the use of traditional rendering methods in computer graphics. The observer's sensation is often straighter, clearer, or even more valuable. There exists plenty of various NPR techniques in computer graphics, however application of one technique to the specific problem is not necessarily as providential as usage of another one. There arises a strong need to classify NPR techniques with respect to their applicability, to find a mechanism able to compare NPR techniques automatically. Such a mechanism in not yet available and the search for it will be a long term goal. In this section we present our first steps towards the solution to this problem – the comparison of 2D-based NPR techniques using Daly's Visible Differences Predictor, described in Section 3.2.4.

Figure 22: Comparison of two input images using the VDP. From the left: images obtained by *Cutout* and *Paint Daubs* techniques, right: colour visualization of the map of detection probabilities.

5.1 Non-Photorealistic Computer Graphics

Contemporary computer graphics methods simulate synthetic scenes with ever-increasing realism and complexity. With this ability comes a new problem of depicting and visualizing these complex scenes in a way that communicates as effectively as possible. Thus, over the past decade, a new area of endeavor has grew up – the *non-photorealism*, (or NPR [46, 20]), dealing with the computer generation of images and animations that, generally speaking, appear to be made in part "by hand". Such images often resemble those that, for example, architects, industrial artists, or scientific illustrators produce to communicate more or less specific information, see Figure 23. They are characterized by their use of randomness, ambiguity, or arbitrariness rather than completeness and adherence to the portrayed objects' properties. Recently, there has been published significant amount of new NPR techniques from artistic screening

methods for printing images using microdots with meaningful shapes that might deliver their own message; to techniques for rendering images in pen-and-ink, watercolor, or engraved etchings styles; to procedures for lighting and even distorting three-dimensional models in order to clarify shapes or direct a wiever's attention.

Several studies have shown that NPR methods allow us to emphasize or omit details in order to communicate information more effectively [19]. For example, sketch rendered images (one group of NPR techniques) of architectural scenes have shown better result in appreciation between architects and clients, compared with that obtained through realistic rendering [45].

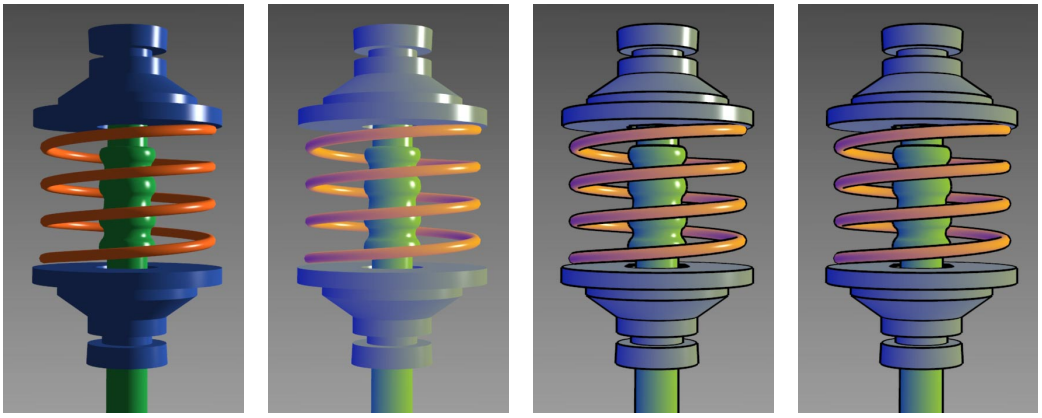

Figure 23: Comparison of traditional computer graphics techniques and the technique of Gooch et al. [18].

5.2 Motivation

Not as much attention as to comparison of classical rendering methods was given to the comparison of output images of NPR techniques. The existence of a general tool that would measure the actual difference between two images with respect to their information content would be strategic both for researchers, in the design of new rendering algorithms, and for users, to allow them to compare the results of different NPR algorithms and to choose the NPR method that best preserves the image semantics.

We assume the way towards the solution to this problem has two stages. The first stage is the "low-level" perception stage that we partly address in this section by means of the Daly's Visible Differences Predictor. Second stage is the "semantic-level" perception where the phenomena like the meaning of a scene, semantics or context are treated.

5.3 Comparison of Image-Processing Operators by the VDP

Since the visualization by means of NPR techniques is commonly not used in cases where we want to obtain image as close to the reality as possible, our problem has two stages. As we already declared before, the first stage is the "low-level" perception stage. In this section we describe the insight in this stage via comparison by means of the VDP method.

The goal is to have a tool that will determine the most suitable technique for the given class of objects. Suitability in our case means the lowest information loss typical for particular NPR method. In such a way we can substantially improve visual communication with computer systems.

5.3.1 Input scenes

We have compared all the techniques mentioned below on several typical input images. These images included a natural photograph of a tree, a computer-generated bust, a classical radiosity scene (cornell box), simple raytraced scene, and several other similar images, see Figure 24. The radiosity scene contained soft shadows, while the raytraced scene encompassed only sharp-edged shadows.

The images were of several resolutions around 640×480 pixels. The diagonal of the images displayed on the CRT was about 0.2 meters, and we assumed, that images were observed from the distance of one meter.

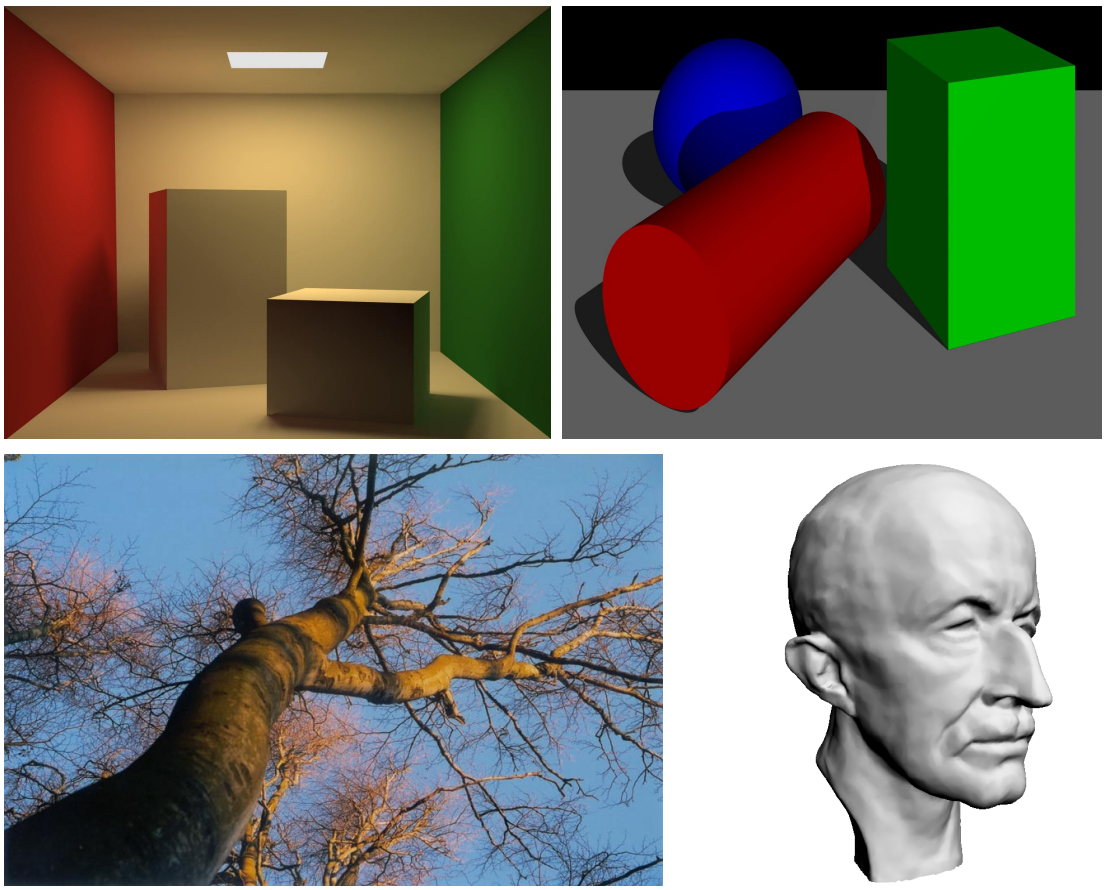

Figure 24: Input Images. Top left: Cornell box scene, top right: ray-tracing scene, bottom left: photograph of a tree, bottom right: bust image.

5.3.2 Tested techniques

There exist notable amount and variation of current techniques in non-photorealistic rendering, see Section 5.1 for an overview. With respect to make our results reproducible, we have used image based techniques ordinary available with the program Adobe Photoshop 6.0, although other possibilities are open [21]. We have investigated following 27 techniques divided into 7 groups:

brush Strokes: Angled Strokes, Crosshatch, Ink Outlines, Spatter,
sketch: Bas Relief, Graphic Pen, Chalk & Charcoal, Charcoal, Note Paper, Photocopy,
artistic: Colored Pencil, Cutout, Dry Brush, Paint Daubs, Poster Edges, Smudge Stick, Sponge, Underpainting, Watercolor,

stylize: Diffuse, Emboss, Find Edges,
other: Crystallize, Add Noise (12.5%), Pointillize, Sharpen, Smart Blur,
and certainly the unchanged **original** picture.

5.3.3 Comparison of the techniques

We compared the output image of every technique with the output image of every other technique, so we obtained $V = \frac{n!}{(n-k)!} = \frac{28!}{(28-2)!} = 756$ variations of difference maps for each input image (27 techniques plus the original image). For each difference map i we computed the difference value $D_{0.75}(i)$.

For all of the techniques we obtainded 28 difference values. These difference values were treated as a discrete difference fuction $D_{0.75}(i)$, where the variable i stands for a technique ordered as in the section 5.3.2. These functions were plotted in planar and 3D graphs and examined for correlations.

We quantified the differences between two difference functions $D_{0.75}(i), D_{0.75}(j)$ by the absolute metrics

$$\rho_{0.75}(x, y) = max|D_{0.75}(i) - D_{0.75}(j)|,$$

where i, j denote technique i or j respectively.

5.4 Results

All of the results were depicted in 3D graphs, see examples of such graphs in Figure 26. Series of investigated techniques ordered as in Section 5.3.2 are drawn on the X and the Y axes, while the probability of difference detection between input images $D_{0.75}$ is displayed as the elevation on the Z axis.

Figure 25: "Point-based" techniques applied on the *Tree* photograph. From the left: *Add Noise, Pointilize* and *Sponge* technique.

Slices of these 3D graphs were used to depict data as ordinary planar charts and used to inspect the results. On the horizontal axis the investigated techniques are depicted as in the 3D graph and similarly you may notice the gaps on depicted functions that separate groups of NPR techniques.

5.4.1 Absolute values of differences

The difference of the original image and the image produced by some image-based NPR technique is typically considerable. Therefore the difference is easy to detect for an observer and the absolute value of difference between images $D_{0.75}$ is high. This is perfectly true for *Ink Outlines* technique. The probabilities of the difference detection always exceed 82%, see the

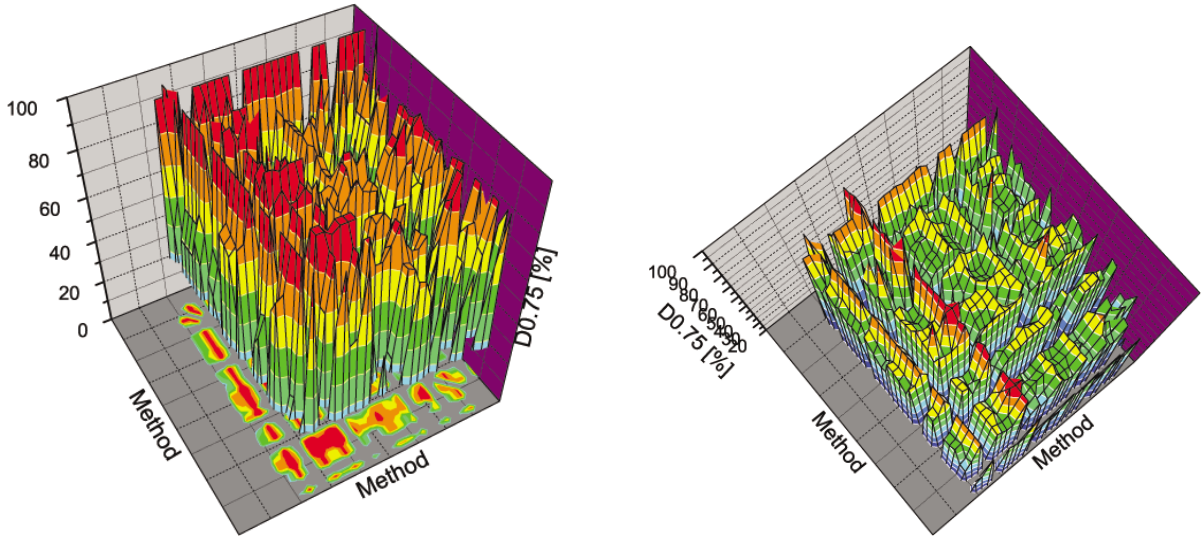

Figure 26: Differences $D_{0.75}$ between all images produced by all inspected NPR techniques depicted as the 3D graph, for the *Cornell box* (left) and the *Bust* (right) input scene. Note the lines of red peaks – almost constalntly high absolute values for the *Ink Outlines* and *Note Paper* methods.

upper function in the left Figure 27. We have observed such a characteristic for any of input images, because *Ink Outlines* technique changes the input image massively and the difference is therefore straightforward.

Very high difference values were observed also for methods *Bas Relief, Graphic Pen, Note Paper* and *Pointilize*. Arithmetical averages of the difference values $D_{avg} = \frac{1}{28}\sum_{i=1}^{28} D_{0.75}(i)$ for these methods are recorded in the Table 1. For example the $D_{avg} = 90\%$ means for particular technique that in average case a human oberver is able to distinguish 90% of the image area when comparing to the image obtained by another technique.

Ideed the distribution of probabilities of difference detection strongly depends on compared techniques. See Figure 29 where probability maps for two pairs of techniques are depicted. Pixels with absolute value of probability detection lower than 0.5 are displayed green while pixels with higher values are in reds.

For the synthetic images the absolute values of the differences were generally lower than for the photos. This is due to the fact that some of the technique's naturally added distortions are not exerted for the synthetic images.

Technique	$\mathbf{D_{avg}}[\%]$
Ink Outlines	99.996
Bas Relief	99.461
Graphic Pen	96.391
Note Paper	94.753
Pointilize	89.761

Table 1: Average values of differences for selected techniques.

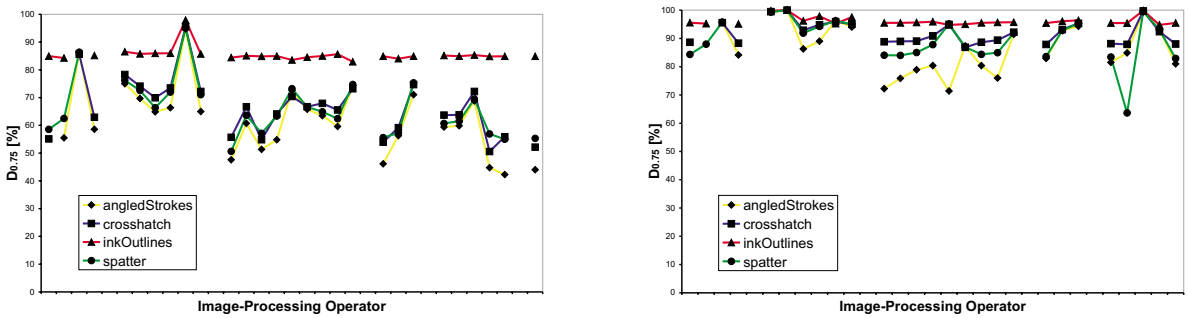

Figure 27: Absolute values of the differences $D_{0.75}$ between images for *Brush Strokes* group of techniques compared with all the other techniques, for the *Bust* (left) input image and the *Tree* (right) input image. Note the upper curve for the *Ink Outlines* technique that does not correlate with the others.

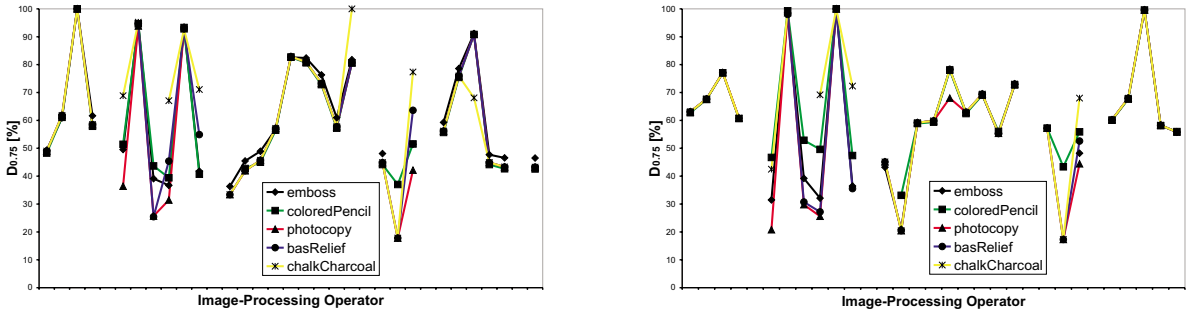

Figure 28: Coherences of techniques *Emboss, Colored Pencil, Photocopy, Bas Relied* and *Charcoal* for the *Cornell box* (left) and *Ray-traced* (right) input image.

5.4.2 Coherences

Consecutively, we have investigated the coherences between the tested techniques to be able to classify them. Coherences between discrete functions $D_{0.75}(i)$ were examined using previously defined absolute metric ρ. This metric was evaluated for each pair of functions $D_{0.75}(i)$, $D_{0.75}(j)$ and was depicted as an ordinary planar graph. Below, we consider as coherent all of the cases when the absolute value of difference ρ does not exceed the 5% threshold.

General Coherences

We observed a strong coherence for following groups of techniques:
1) *Diffuse, Dry Brush, Original, Sharpen, Smart Blur*,
2) *Noise, Pointilize, Sponge*,
3) *Colored Pencil, Crystalize, Paint Daubs, Photocopy, Spatter.*
These coherences were independent on the type of the input image. Average values of ρ for several pairs from these groups are recorded in Tables 2.

Just mentioned groups reflect our vague knowledge of common properties of the given techniques. In the group 1 there are techniques that do not distort the image too much, they are just "improving" the input image in some sense. Group 2 consists of "point-based" techniques,

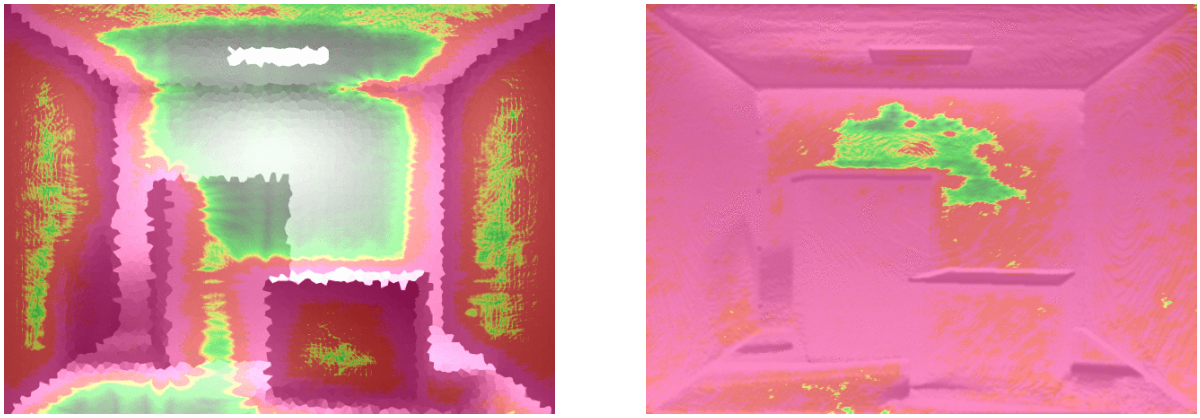

Figure 29: The map of probabilities for comparison of *Crystalize* and *Graphic Pen* (on the left) and for comparison of *Bas Relief* and *GraphicPen* (on the right) techniques visualized in the same colours as in the project [35]. *Cornell box* input scene.

see Figure 25. Finally, in the group 3 there are techniques producing similar "shake" distortion.

Coherences between the groups of techniques

In the scope of the groups of techniques (brush strokes, sketches, artistic effects, stylize, other techniques) we primarily expected good coherences between techniques. There is sometimes noticable correlation of the results, see for instance the chart on the left Figure 27. However, generally we have noticed good coherences only in the group of schetches (after excluding the *Note Paper* method), where the average values of ρ do not exceed 6%. In the other groups there are no conspicuous coherences between apropriate techniques, see the right Figure 27.

Coherences dependent on type of input image

Apart from general-valid coherences described above, we found also the other group of coherent techniques dependent on the type of the input image. We have observed that for synthesised images with uniform-colored faces there is a good coherence between *Emboss, Colored Pencil, Photocopy, Bas Relief,* and *Charcoal* techniques. This is especially distinctive in the case of the *Cornell box* scene, where these techniques produce similar edge enhancement. The graphs in Figure 28 exhibit the coherence between the mentioned techniques. Note the minimal differences in the first and third parts of the charts, that represent very good coherence with the *Brush Strokes* and *Artistic* groups respectively.

5.5 Conclusions

In this section we have described our first steps towards finding a mechanism which would be able to automatically compare NPR images. We have investigated the properties of the images, obtained by various image-processing techniques, using the Visible Differences Predictor.

We have shown that by such a low-level mechanism like the VDP is (from the point of view of the complexity of the human perception of the NPR images) we are able to distinguish some of the naturally vague defined groups of images with similar properties.

	Technique–Technique	$\rho_{\mathbf{avg}}[\%]$
Group 1)	**Original–Smart Blur**	0.5799
	Original–Diffuse	2.639
	Dry Brush–Sharpen	4.949
Group 2)	**Noise–Sponge**	1.025
	Pointilize–Noise	3.897
	Pointilize–Sponge	4.346
Group 3)	**Color Pencil–Photocopy**	3.206
	Spatter–Crystalize	0.9
	Crystalize–Paint Daubs	4.772

Table 2: Average values of ρ for selected pairs of techniques.

Next, we have observed that the absolute values of differenecs are inherently high for most techniques and that these values are generally lower for synthetic images than for the photographs. However, it is evident that the VDP-like mechanisms are just a first stage in the field of comparing of the NPR techniques. NPR techniques are often utilized in such cases where we want to highlight fundamental information content of the image, which principles of "low-level" perception are unable to catch.

In the future, we will carry on psychophysical experiments on human observers in order to validate the presented results. We also intent to compare our results with another algorithms describing the human visual response, and especially we want to interpret our knowledge in the context of the work of Duke et al. [11] to design an algorithm, the result of which will correspond with "semantic sensation" of a human observing a NPR image.

6 Conclusions

In this report, we have summarized physiological and perceptual properties of the human visual system. Human perception is a complex process of obtaining knowledge of visual environment, which has many specific properties. Atlhough not yet fully understood, we have shown that it can be and recently really was successfully applied to various issues of computer graphics. In computer graphics, the perceptual knowledge usually takes the form of human vision models. We have described various HVS models and concentrated on their applications to the image and video quality assessment and comparison. We have outlined merits and shortcomings of these HVS models and we have given several suggestions for the future research. Relevant ideas from the image assessment field were applied to the acceleration of computer graphics rendering. We have given an overviewed of these applications. Finally, we have outlined the way to overcome some drawbacks of contemporary approaches to the image comparison.

7 References

[1] C. Blakemore and F. W. Campbell. On the existence of neurons in the human visual system selectively responsive to the orientation and size of retinal images. *Journal of Physiology*, 203:237– 260, 1969.

[2] M. R. Bolin and G. W. Meyer. A perceptually based adaptive sampling algorithm. *Computer Graphics*, 32(Annual Conference Series):299– 309, 1998.

[3] A. P. Bradley. A wavelet visible difference predictor. *IEEE Transactions on Image Processing*, 8(5):717– 730, 1999.

[4] P. J. Burt and E. H. Adelson. The laplacian pyramid as a compact image code. *IEEE Transactions on Communications*, Com–31:532– 540, 1983.

[5] F. W. Campbell, J. J. Kulikowski, and J. Levinson. The effect of orientation on the visual resolution of gratings. In *Journal of Physiology*, volume 187, pages 427– 436, 1966.

[6] P. Cignoni, C. Rocchini, and R. Scopigno. Metro: measuring error on simplified surfaces. In *Computer Graphics Forum*, 1998.

[7] L. da Fontoura Costa and R. M. Cesar. *Shape analysis and classification: theory and practise*. CRC Press, USA, 2001.

[8] S. Daly. The visible differences predictor: An algorithm for the assessment of image fidelity. In A. B. Watson, editor, *Digital Images and Human Vision*, pages 179– 206, Cambridge, MA, 1993. MIT Press.

[9] S. Daly. Engineering observations from spatiovelocity and spatiotemporal visual models. In *IS&T/SPIE Conference on Human Vision and Electronic Imaging III, SPIE volume 3299*, pages 180–191, January 1998.

[10] P. Debevec, G. Ward, and D. Lemmon. HDRI and image-based lighting. In *Siggraph '03 Course no. 19*, 2003.

[11] D. J. Duke, P. J. Barnard, N. Halper, and M. Mellin. Rendering and affect. *Computer Graphics Forum*, 22(3), 2003.

[12] J. A. Ferwerda. Three varieties of realism in computer graphics. In B. E. Rogowitz and T. N. Pappas, editors, *Proceedings SPIE Human Vision and Electronic Imaging '03*, volume 5007, pages 290– 297, 2003.

[13] J. A. Ferwerda, H. Rushmeier, and B. Watson. Frontiers in perceptually-based image synthesis: Modeling, rendering, display, validation. In *SIGGRAPH '03 Course no. 03*, 2003.

[14] W. T. Freeman and E. H. Adelson. The design and use of steerable filters. *IEEE Trans. Pattern Analysis and Machine Intelligence*, 13(9):891– 906, 1991.

[15] M. J. Gervais, L. O. Harvey, and J. O. Roberts. Identification confusions among letters of the alphabet. In *Journal of Experimental Psychology: Human Perception and Performance*, volume 10(5), pages 655– 666, 1984.

[16] S. Gibson and R. J. Hubbold. Perceptually-driven radiosity. *Computer Graphics Forum*, 16(2):129– 141, 1997.

[17] R. C. Gonzalez and R. E. Woods. *Digital Image Processing*. Prentice-Hall, New Jersey, 2nd edition, 2002.

[18] A. Gooch, B. Gooch, P. Shirley, and E. Cohen. A non-photorealistic lighting model for automatic technical illustration. *Proceedings of SIGGRAPH 98*, pages 447–452, July 1998. ISBN 0-89791-999-8. Held in Orlando, Florida.

[19] A. A. Gooch and P. Willemsen. Evaluating space perception in npr immersive environments. In *Non-Photorealistic Animation and Rendering (NPAR'02)*, France, 2002.

[20] B. Gooch and A. Gooch. *Non-Photorealistic Rendering*. AK Peters, Ltd., Canada, 2001.

[21] N. Halper, T. Isenberg, F. Ritter, B. Freudenberg, O. Meruvia, S. Schlechtweg, and T. Strothotte. Opennpar: A system for developing, programming, and designing non-photorealistic animation and rendering. In *Proceedings of Pacific Graphics '03*, CA, 2003.

[22] D. H. Hubel and T. N. Wiesel. Receptive fields of single neurons in the cat's striate cortex. In *Journal of Physiology*, volume 148, pages 574– 591, London, 1959.

[23] A. Karasaridis and E. Simoncelli. A filter design technique for steerable pyramid image transforms. In *Proceedings of ICASSP-96*, Atlanta, 1996.

[24] H. Kolb, E. Fernandez, and R. Nelson. Webvision – the organization of the vertebrate retina. electronic publication, October 2000.

[25] V. Kruger and G. Sommer. Affine real-time face tracking using a wavelet network. In *Proceedings ICCV'99 Workshop Recognition, Analysis and Tracking of Faces and Gestures in Real-time Systems*, Greece, 1999.

[26] Y. K. Lai and C.-C. J. Kuo. A Haar wavelet approach to compressed image quality measurement. *Journal of Visual Communication and Image Understanding*, 11:17– 40, 2000.

[27] P. Lindh and C. J. van den Branden Lambrecht. Efficient spatio-temporal decomposition for perceptual processing of video sequences. In *Proceedings of the International Conference on Image Processing*, volume 3, pages 331– 334, Lausanne, Switzerland, 1996.

[28] M. S. Livingstone and D. Hubel. Segregation of form, color, movement, and depth: Anatomy, physiology, and perception. In *Science*, volume 240(4853), pages 740– 749, 1988.

[29] J. Lubin. A visual discrimination model for imaging system design and evaluation. In E. Peli, editor, *Vision Models for Target Detection and Recognition*, pages 245– 283, Singapore, 1995. World Scientific.

[30] F. X. J. Lukas and Z. L. Budrikis. Picture quality prediction based on a visual model. *IEEE Transactions on Communications*, 30(7):1679– 1692, 1982.

[31] J. L. Mannos and D. J. Sakrison. The effects of a visual fidelity criterion on the encoding of images. In *IEEE Transactions on Information Theory*, volume IT-20, pages 525– 536, 1974.

[32] W. L. Martens and K. Myszkowski. Psychophysical validation of the visible differences predictor for global illumination applications. *Proceedings of IEEE Visualization '98*, 1998.

[33] I. Martin, X. Pueyo, and D. Tost. An image-space refinement criterion for linear hierarchical radiosity. *Graphics Interface'97*, pages 26– 36, 1997.

[34] K. Myszkowski. The visible differences predictor: applications to global illumination problems. In *Proceedings of the Ninth Eurographics Workshop on Rendering*, pages 223– 236, Wienna, Austria, 1998.

[35] K. Myszkowski. The visible differences predictor: applications to global illumination problems. Web Pages, 1998. `http://www.mpi-sb.mpg.de/resources/vdp`.

[36] K. Myszkowski. *Efficient and Predictive Realistic Image Synthesis*. Warsaw Institute of Technology, 2001.

[37] K. Myszkowski. Perception-based global illumination, rendering, and animation techniques. In A. Chalmers, editor, *Proceedings of the 18th Spring Conference on Computer Graphics (SCCG 2002)*, pages 13– 24, Budmerice, Slovakia, 2002.

[38] J. Oh, S. I. Wooley, T. N. Arvanitis, and J. N. Townend. A multistage perceptual quality assessment for compressed digital angiogram images. *IEEE Transactions on medical imaging*, 20(12), 2001.

[39] S. E. Palmer. *Vision science – photons to phenomenology*. The MIT Press, Cambridge, 3rd edition, 2002.

[40] J. Přikryl. *Radiosity Methods Driven by Human Perception*. PhD thesis, Technische Universitat Wien, 2001.

[41] M. Ramasubramanian, S. N.Pattanaik, and D. P. Greenberg. A perceptually based physical error metric for realistic image synthesis. In A. Rockwood, editor, *Siggraph 1999, Computer Graphics Proceedings*, pages 73– 82, Los Angeles, 1999. Addison Wesley Longman.

[42] M. Reddy. Perceptually optimized 3D graphics. *IEEE Computer Graphics and Applications*, 21(5):68– 75, Sep/Oct 2001.

[43] H. Rushmeier, G. Ward, C. Piatko, and P. Sanders. Comparing real and synthetic images: Some ideas about metrics. In *Eurographics Rendering Workshop 1995*, 1995.

[44] A. Schnayderman, A. Gusev, and A. M. Eskicioglu. A multidimensional image quality measure using singular value decomposition. *IS&T/SPIE Symposium on Electronic Imaging 2004*, 2004.

[45] J. Schumann, T. Strothotte, A. Raab, and S. Laser. Assessing the effect of nonphotorealistic rendered images in CAD. In *ACM Human Factors in Computing Systems, SIGCHI '96*, pages 35– 41, 1996.

[46] T. Strothotte and S. Schlectweg. *Non-Photorealistic Computer Graphics: Modeling, Rendering, and Animation*. Morgan-Kaufman, San Francisco, USA, 2002.

[47] D. S. Taubman and M. W. Marcellin. *JPEG2000: Image Compression Fundamentals, Standards and Practice*. Kluwer Academic Publishers, 2002.

[48] C. Taylor, Z. Pizlo, J. Allebach, and C. Bouman. Image quality assessment with a gabor pyramid model of the human visual system. *Proceedings of the 1997 IS&T/SPIE International Symposium on Electronic Imaging Science and Technology*, 3016:58– 69, 1997.

[49] P. C. Teo and D. J. Heeger. Perceptual image distortion. In *Human Vision, Visual Processing and Digital Display*, volume 2179, 1994.

[50] X. Tong, D. Heeger, and C. J. van den Branden Lambrecht. Video quality evaluation using ST-CIELAB. In *Proceedings of SPIE Human Vision and Electronic Imaging*, volume 3644, pages 185– 196, San Jose, CA, 1999.

[51] C. J. van den Branden Lambrecht. *Perceptual Models and Architectures for Video Coding Applications*. PhD thesis, École Polytechnique Fédérale de Lausanne, Switzerland, 1996.

[52] V. Volevich, K. Myszkowski, A.Khodulev, and E.A.Kopylov. Using the visual differences predictor to improve performance of progressive global illumination computations. *ACM Transactions on Graphics*, 19(2):122– 161, 2000.

[53] Z. Wang and A. C. Bovik. A universal image quality index. *IEEE Signal Processing Letters*, 9(3):81– 84, 2002.

[54] Z. Wang, A. C. Bovik, and L. Lu. Why is image quality assessment so difficult. In *Proceedings of IEEE International Conference Acoustics, Speech, and Signal Processing*, volume 4, pages 3313– 3316, 2002.

[55] Z. Wang, H. R. Sheikh, and A. C. Bovik. Objective video quality assessment. In B. Furht and O. Marqure, editors, *The Handbook of video databases: design and applications*, chapter 41, pages 1041– 1078. CRC Press, 2003.

[56] A. B. Watson. The cortex transform: Rapid computation of simulated neural images. *Computer Vision Graphics and Image Processing*, 39(3):311– 327, 1987.

[57] S. Winkler. *Vision Models and Quality Metrics for Image Processing Applications*. PhD thesis, EPFL, Switzerland, 2000.

[58] M. Yuen and H. R. Wu. A survey of hybrid MC/DPCM/DCT video coding distortions. *Signal Processing*, 70:247– 278, 1998.

[59] X. Zhang and B. A. Wandell. A spatial extension of CIELAB to predict the discriminability of colored patterns. In *Society for Information Display Symposium Digest*, volume 27, pages 731– 735, 1996.

8 Dissertation thesis

Title: Perceptually Based Image Quality Assessment

Abstract

The aim of dissertation is to develop a general model for perceptual image quality assessment to overcome the limitations of contemporary approaches that are insufficient when processing other than photorealistic inputs. The model will exploit the knowledge of human visual system, and the information contents of the input image. Intended approach has many applications that are not only limited to the image comparison and assessment of image quality. It could be used to compare the perceptual impact of different rendering algorithms, or to analyse the effect of various acceleration techniques. The second objective of the dissertation work is the utilization of the perceptual assessment model in the rendering algorithms. The output of a model will be used as a feedback to a rendering algorithm for iterative steering the quality of the rendered image. Whereas the accuracy of a common error metric can be verified by instruments, the only way to measure accuracy of a perceptually based metric is to see how well observers perform in visual tasks. The third aim of the dissertation work is therefore the execution of subjective verification tests.

Keywords

computer graphics, human perception, human visual system, image quality, vision models, image comparison, acceleration of rendering, image processing

9 Publications of the author

[1] M. Čadík and P. Slavík. Comparing Image-Processing Operators by Means of the Visible Differences Predictor. To appear (accepted) In *Proceedings of WSCG 2004.* Plzen, University of West Bohemia Press, 2004.

[2] M. Čadík. Automatic Comparison of NPR Techniques. To appear (accepted) In *Proceedings of Workshop 2004.* Prague, Czech Technical University in Prague, 2004.

[3] M. Čadík. Visualization of the Light Load Using Java 3D. In *Poster 2003.* Prague, Czech Technical University in Prague, p. IC4, 2003.

[4] M. Čadík, P. Slavík and J. Přikryl. Experimental System for Visualisation of the Light Load. In *Proceedings of WSCG 2003.* Plzen, University of West Bohemia Press, vol. p, p. 37– 40, 2003.

A Image Pyramids

When we need to work simultaneously with various image resolutions, it is useful to utilize image pyramids. Image pyramid is a powerful, but conceptually simple structure for representing images at more than one resolution [4]. It is a collection of decreasing resolution images arranged in the shape of a pyramid. The base of the pyramid contains a high-resolution representation of image being processed, the apex contains a low-resolution approximation. As we move up the pyramid, both size and resolution decrease, see Figure 30.

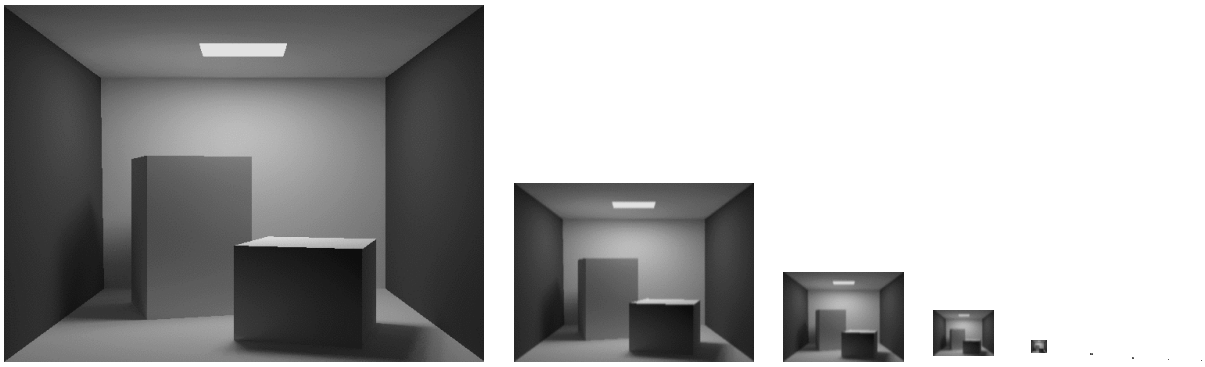

Figure 30: Nine levels of the Gaussian pyramid for the "Cornell box" image. The original image is the leftmost, each higher level array is roughly half the dimensions of its predecessor.

When the base level J is size $2^J \times 2^J$, where $J = log_2 N$, intermediate level j is size $2^j \times 2^j$, where $0 \leq j \leq J$. Fully populated pyramids are composed of $J + 1$ resolution levels up to $2^0 \times 2^0$, but most pyramids are truncated to $P + 1$ levels, where $1 \leq P \leq J$. The total number of elements in a $P + 1$ level pyramid for $P > 0$ is

$$N^2\left(1 + \frac{1}{(4)^1} + \frac{1}{(4)^2} + \cdots + \frac{1}{(4)^P}\right) \leq \frac{4}{3}N^2.$$

Both the original image, which is at the base of the pyramid, and its P reduced resolution approximations can be accessed and manipulated directly. The creation of the next level of a pyramid from an input image is composed of three sequential steps [17]:

1. Compute a reduced-resolution approximation of the input image. This is done by filtering the input and downsampling the filtered result by factor of 2. A variety of filtering operations can be used, including neighborhood averaging, which produces a *mean pyramid*, lowpass Gaussian filtering, which produces a *Gaussian pyramid*, or no approximation, which results in a *subsampling pyramid*.

2. Upsample the output of the previous step – again by a factor of 2 and filter the result. This creates a *prediction* image with the same resolution as the input. By interpolating intensities between the pixels of the *Step 1* output, the *interpolation filter* determines how accurately the *prediction* approximates the input to *Step 1*.

3. Compute the difference between the *prediction* of *Step 2* and the input to *Step 1*. This difference, labeled the *level j prediction residual*, can be later used to reconstruct progressively the original image.

Executing this procedure P times produces two intimately related $P+1$ level approximation and prediction residual pyramids. If a prediction residual pyramid is not needed, *Steps 2* and *3* can be omitted.

B Multiscale Transforms

This section presents a brief introduction to a series of multiscale transforms that are widely applied to signal and image processing [7]. In general, signal analysis by using a multiscale transform is characterized by the following elements. A signal $u(t)$ presents a set of features and structures occuring at different spatial scales. This signal is to be analyzed by a multiscale transform $U(b,a)$ involving two parameters: b, associated with the *time* variable t of $u(t)$, and a, associated with the *analyzing scale.* The scale parameter a is usually related to the inverse of the *frequency* f, i.e., $\frac{1}{a} \cong f$, leading to a dual interpretation of these transformations and suggesting the terms *time-scale* and *time-frequency.*

Many different multiscale-based signal analysis frameworks can be summarized by the Algorithm 1.

Algorithm 1: Typical Multiscale Analysis
(1) Obtain the signal $u(t)$ to be analyzed
(2) Calculate the multiscale transform $U(b,a)$ of $u(t)$
(3) Extract the important scale characteristics of $u(t)$ from $U(b,a)$

The *scale-space* approach to multiscale signal analysis is one of the most popular multiscale methods. The scale-space of a signal allows tracking its singularities (or of one of its derivatives) through the scale dimension. This fits to conjecture by Marr that the perceptually important dominant signal points correspond to singularities remaining along longer scale intervals. The scale space is defined as follows.

Definition 1 *Let $u(t)$ be the signal to be analyzed, $g_a^{(1)}$ be the first derivative of the Gaussian function $g_a(t)$ and $U^{(1)}(t,a) = u(t) * g_a^{(1)}(t)$. Let $\left\{U^{(1)}(t,a_0)\right\}_{zc}$ denote the zero-crossings set of $U^{(1)}(t,a_0)$. The* ***scale-space*** *of $u(t)$ is defined as the set of zero-crossings of $U^{(1)}(t,a)$, i.e.*

$$\left\{(b_0,a_0)|a_0,b_0 \in R, a_0 > 0, and\ b_0 \in \left\{U^{(1)}(t,a_0)\right\}_{zc}\right\}$$

The term scale-space is sometimes also used for $U(t,a)$, obtained by the convolution of the signal with a series of Gaussians. In general, the extrema of $U^{(n)}(t,a)$ can be defined from the zero-crossings of the scale-space generated by $g_a^{n+1}(t)$.

The *time-frequency* transforms have originated as an alternative to Fourier analysis capable of *signal local analysis.* Fourier transform has engaging properties, but it operates on the whole signal. The short-time Fourier transform has been defined as an attempt to circumvent this problem by introducing an *observation window* aimed at selecting a signal portion during the transform. The following transform was defined from the Fourier equation:

$$U(b,f) = \int_{-\infty}^{\infty} g^*(t-b)u(t)\exp(-j2\pi ft)dt,$$

where $g(t)$ is the window function that *slides* over the signal $u(t)$. Time-frequency analysis is based on the above considerations, and one of its most important tools is the *Gabor transform,* i.e., the short-time Fourier transform where $g(t)$ is a Gaussian window.

The *time-scale* transform (or modernly the *wavelet* transform) faces the following problem: high-frequency events frequently occure along short time intervals while low-frequency components remain longer during the signal evolution. The application of the short-time Fourier transform has the drawback that the analyzing window size is the same for all frequencies. Therefore, in the wavelet transform the kernel size varies with the frequency, allowing high-frequency events to be localized with better time resolution, while low-frequency components are analyzed with better frequency resolution. This property is known as *relative bandwidth* or *constant-Q*. The continuous wavelet transform is defined as:

$$U[\psi, u](b, a) = U_\psi(b, a) = \frac{1}{\sqrt{a}} \int_{-\infty}^{\infty} \psi^*\Big(\frac{t-b}{a}\Big) u(t) dt.$$

In the case of the Morlet wavelet, the Gabor transform and the so-called Gabor wavelets [25], the transform presents a strong and interesting biological inspiration (i.e. the receptive fields of neural cells involved in visual processing).