Perceptual modelling for 2D and 3D

Vincent Ricordel, Junle Wang, Josselin Gautier, Olivier Le Meur, Emilie Bosc

Projet PERSEE
SCHÉMAS PERCEPTUELS ET CODAGE VIDÉO 2D ET 3D
n° ANR-09-BLAN-0170

Livrable **D1.1** *17/11/2010*

Perceptual modelling for 2D and 3D

Vincent	RICORDEL	IRCCyN
Junle	WANG	IRCCyN
Josselin	GAUTIER	IRISA
Olivier	LE MEUR	IRISA
Emilie	BOSC	INSA

Table des matières

Introduction

Three-dimensional television (3DTV) is meant to enhance conventional 2D television by the added feeling of depth. The introduction of 3DTV will be successful if the perceived image quality and the viewing comfort are at least comparable to conventional television. Therefore, it is of relevant importance to understand human vision mechanisms, in order to provide the best image quality and a great 3D experience.

Stereoscopic vision is based on stereopsis : depth perception relies on the fusion of two slightly different viewpoints of the same scene and also on monocular cues. In 3DTV, stereoscopic video pairs can be provided by a multi-view video acquisition, and enhanced by depth estimations of the scene. Original stereoscopic pair or virtual rendered ones are then displayed on autostereoscopic display systems. In order to achieve a sufficient 3D experience, many fields should be studied.

This document is divided in three main parts. First, human depth perception will be addressed ; then an overview of 2D and 3D perceptual modelling is presented. The third part will address the applications.

Chapitre 1

Introductory elements of psychophysics

Achieving good image quality requires extensive research through the whole imaging process chain, i.e. content generation, coding algorithms, transmission and display technology. Preferences of customers drive the improvements to be done and thus, it is important to understand the mechanisms of vision, and the typical needs in 3D video. So, psycho visual aspects have to be considered when elaborating quality metrics that can drive the imaging process.

This chapter will first address the human visual system and human depth perception. Then elements on eye tracking will be presented.

1.1 Anatomy

The human vision is a very complex process that is still not fully understood. The first organ responsible for vision is identified as the eye, which consists of many parts (see figure 1.1).

Visual information is received through the eye and transmitted to the brain which processes it and allows us to interpret the environment. Through the hole in the middle of the iris, namely the pupil, the light enters. The light then refracts when entering the lens whose curvature can be changed by the attached muscles. At the back of the eye, the retina receives the images, but because of the optical characteristics, the projection is reversed. After a few pre-process through the retina's cells, the image is brought in to the brain which processes it. Those cells can be divided in two categories : the parvocellular cells that response to fine image details and chromatic information ; and the magnocellular cells which are sensitive to form, motion, depth. Then the photoreceptors (cones and rods) detect colours and bright light.

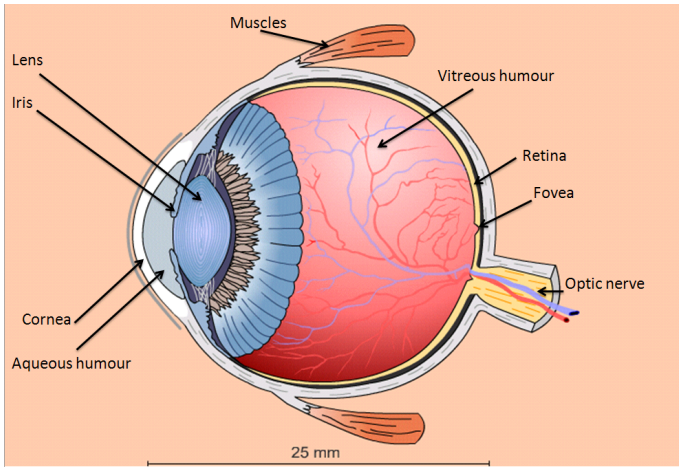

FIGURE 1.1 – Human eye[1]

Through visual pathways, the light arrives to the visual cortex, at the back of the brain. Later, the signals are brought to others areas such as V2 area and dorsomedial area (V6), and V5, all of which make a primary pathway. This pathway, also known as dorsal stream is associated with motion, representation of object locations and control of the eyes. The second pathway consists of V2 area, V4 area and the inferior visual cortex. It is known as the ventral stream and is associated with form recognition and object representation.

For more details on human anatomy, please refer to [2] and [3].

1.2 Human perception of depth

Both human eyes receive on their two retinæ the explicit 2D-images from their environment. But how the third dimension -the distance of the surface from the observer- which is lost in the optical projection from surfaces in a 3D world to the 2D retinæ, is regained ?

In the literature [4], the perception of spatial arrangement of surface to the observer is divided in two areas. First by the depth determination : distance of the surface to the observer in a 3D environment. Secondly by the surface orientation perception : the slant and tilt of the surface with respect to the viewer's line of sight. Depth and surface are however intrinsically interdependent, the orientation of surfaces give a distance information of its various parts from the observer. This concept of surface perception was originally proposed by [5] (in 1950), and was later modelized by [6] (1978) by a surface-based representation : the 2.5D sketch. A representation of oriented surfaces in depth has been proven to be necessary to vision. The next sections will help to understand higher level perceptual phenomenas.

It is now widely accepted that the human vision system (HVS) integrates different "cues" with more or less accuracy. The different depth cues, or sources of information giving the depth are presented in table 1.1. A classification of five depth characteristics are used : ocular versus optical, binocular versus monocular, static versus dynamic, relative versus absolute, and qualitative versus quantitative.

The depth determination vs surface orientation could be seen as another division, but they are in fact intrinsically interdependent, and one helps to determine the other. Next chapter present how some depth cues relied on pictorial information to determine surface orientation and then the depth of the various surface parts.

1.2.1 Ocular information

The ocular information are related to state of the eyes and their components. The focus of lens, is called **accommodation** and the angle between the two lines of sight of each eyes is called**convergence** or sometimes also **vergence**.

Accommodation

For each eye, the optical focus of the lens is controlled by the ciliary muscles around it. By applying different tensions on the lens, the shape of the lens vary temporarily : thin to focus on light to faraway objects, and thick

INFORMATION SOURCE	Ocular/ Optical	Binocular/ Monocular	Static/ Dynamic	Relative/ Absolute	Qualitative/ Quantitative
Accommodation	**ocular**	monocular	static	**absolute**	quantitative
Convergence	**ocular**	**binocular**	static	**absolute**	quantitative
Binocular disparity	optical	**binocular**	static	relative	quantitative
Motion Parallax	optical	monocular	**dynamic**	relative	quantitative
Texture accretion/del.	optical	monocular	**dynamic**	relative	qualitative
Convergence of parall.	optical	monocular	static	relative	quantitative
Position/ Horizon	optical	monocular	static	relative	quantitative
Relative size	optical	monocular	static	relative	quantitative
Familiar size	optical	monocular	static	**absolute**	quantitative
Texture gradient	optical	monocular	static	relative	quantitative
Edge interpretation	optical	monocular	static	relative	**qualitative**
Shading and shadows	optical	monocular	static	relative	**qualitative**
Aerial perspective	optical	monocular	static	relative	**qualitative**

TABLE 1.1 – Sources of information about Depth from [4] "This chart specifies five important characteristics of depth information : ocular versus optical, binocular versus monocular, static versus dynamic, relative versus absolute, and qualitative versus quantitative"

for nearby ones. So, if the HVS has information about the tension to apply on the muscles that control the shapes, then it has **absolute** information about the distance of the object to focus on (considering the visual system is properly "calibrated").

Different studies have shown accommodation is a weak but useful source of depth information at close distance, not especially to make direct judgment about distance but also to evaluate the size of objects [7]. Beyond 2 meters, accommodation provides hardly no depth information, as ciliary muscles are already in their most relaxed state.

But how does the HVS guess the accommodation to adopt ? The visual system guess the proper focus to apply on the retina by blurriness/sharpness analysis of edges : indeed the best indication of proper focus is the amount of "sharpness" of edges. Psychophysically speaking, as sharp edges contains more energy than blurry ones, it is likely that ciliary muscle tension is adjusted so it maximize the output of high spatial frequency channels.

Vergence

The second ocular depth information is the eye vergence : the extent between two eyes turned inward, i.e the angle in degree between the two line of sight. The objective of this function is to make the eyes fixate the same point in space so that emitted light from that point falls in the center of both foveae simultaneously. Then, the angle of vergence is correlated and varies directly with the distance to the fixated object. A close object will

be fixated with a large convergence angle, a far one with a small one. It is then a **binocular** source of depth information, and provides an **absolute** information about the distance : HVS can specify the actual distance to the fixated object.

The convergence can be expressed through right triangles trigonometry as

$$d = \frac{c}{2tan(a/2)}$$

with d the distance in meters and a the convergence angle in degree. From this asymptote behavior we can see that the angle of convergence decrease rapidly up to a meter or two, but very little after, where it tends to the asymptote. That also means that vergence control information is a reliable and accurate information up to two meters, as for the accommodation case.

Accommodation and vergence are interdependent and covary : change in the distance to an object will imply related change in accommodation and vergence. Studies with covered eyes also shown that the convergence is driven by the monocular accommodation.

Accommodation and vergence are then dependent contributions to depth perception and are among the few cues that give absolute distance to fixation point, but at close distance.

1.2.2 Stereoscopic information

The distance between the two human eyes enables to perceive the world from two slightly translated viewpoints. Regarding the last vergence section, we have seen that this viewpoints are also rotated to fixate the same object. The visual field then overlap in the central region of vision, so that the same point projected to left and right retinæ are displaced according to the distance of this point from the fixation point (and according to the distance from distant fixated point). This relative displacement is called **binocular disparity**.

Binocular disparity

The *direction* of disparity shows which points are closer and farther than the fixated point, the *magnitude* provides the quantitative information on how much closer and farther they are. As binocular disparity happened when a given point in the external world is not projected to the same position on the left and right retinæ, a closer-in-depth point than the fixated point will fall in outward direction on both fovea, this is the **crossed disparity**. At

the opposite, a farther point will project in the inward direction, this is **uncrossed disparity**, as illustrated in figure 1.2.

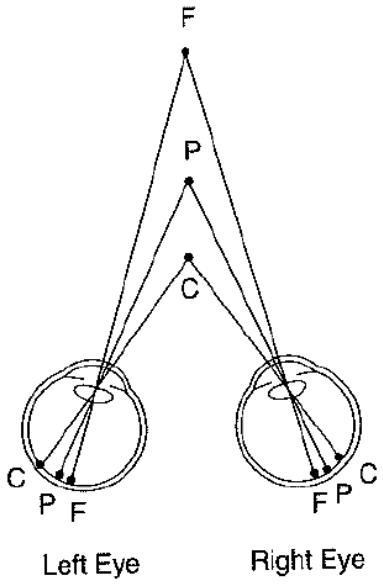

FIGURE 1.2 – Crossed versus uncrossed binocular disparity. When a point P is fixated, closer points (C)are displaced outwardly in crossed disparity.Farther points (F) are displaced inwardly in uncrossed disparity

The stereopsis, the process of perceiving distance to object or "depth" based on their lateral displacement, has to deal with the correspondence problem : how the visual system process to determine which feature in one retinal image correspond to another one in the second retinal one ? First theorists assumed that a process of shape analysis realized this task before stereopsis, but stereogram tests tends to say the opposite : stereopsis come first without any monocular shape information.

Vertical disparity

As seen before, binocular disparity consists of viewing the same object from two viewpoints. But now if this object is moving not only along a depth axis, but along an horizontal line from the eyes viewpoint, it will introduced a vertical binocular disparity. Let's consider the image in figure 1.3, the object seen by the right eye is bigger in right retinea than in left one. In fact the object is bigger along the horizontal axis in both direction, but also along the vertical axis. There is a vertical disparity between corresponding points in space.

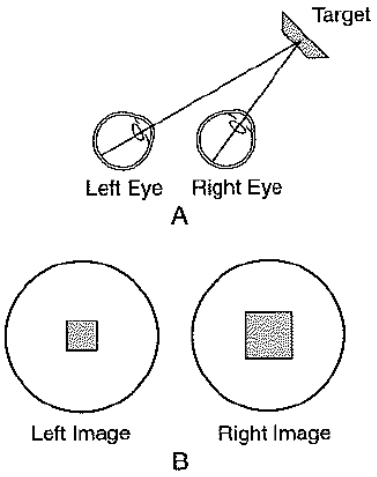

FIGURE 1.3 – Illustration of the vertical disparity. Here a surface of an object is closer to one eye than the other, differences in size on each retinae (B) leads to horizontal and vertical disparity

Da Vinci stereopsis

[8] underline the fact that, in binocular condition, an another perception of depth is allowed through the occluded areas. For different object depth, a part of object surface will be seen by just one eye. Considering to surface as figure 1.4, only the right eye can see the right part of the farther surface **occluded** by the closer surface on the left eye. This monocular perception

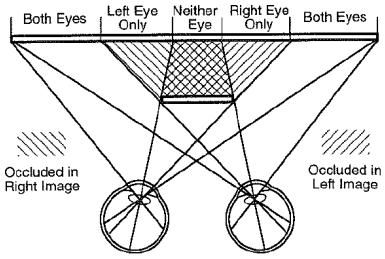

FIGURE 1.4 – Illustration of the Da Vinci stereopsis. Occluded part on one retinal image can be disoccluded on the other retinal one.

of regions provides an important binocular information about the relative position and distance of surfaces. Indeed, the depth information arise from the single monocularly viewed region, which belong necessarily to the farther surface. As with the vertical disparity, this relative depth cue can be seen as a subpart of the binocular disparity cue.

1.2.3 Dynamic cues

Displacement of object on the retinae over the time provides dynamic visual information to the HVS. It is then due to retinal image motion, or "optic flow". Whatever if an observer moves, the fixated object moves in an environment, or the whole environment is moving around an observer, direction and rate at which objects are retinally displaced depend on the **motion**, but also on the **distance** and the **position** from the object to the observer. The next sections describe how the visual system manage to get the "Depth from motion" information.

Motion parallax

The perception that image of points at different distances from the observer moves at different retinal velocities as the observer viewpoint is changing is an illustration of the motion parallax. In other terms, the differential motion of pairs of points due to their different depths relative to the fixation point provides the motion and then the depth information.

An analogy can be done with the binocular disparity (see section 2.1) Binocular disparity is related to the difference between a pair of displaced, but taken in the same time, retinal images, whereas motion parallax involves the difference between a pair of displaced through time retinal images. As for binocular disparity, the nature of retinal motion parallax is related to the distance to objects in the environment but also to the observer's fixation point.

In addition, the perception we have of motion parallax i.e the real movement of an object on the retinal image is a perception of depth rather than movement. We naturally experienced from motion parallax objects at different depth rather than objects at different motion speeds.

Consequently, motion parallax gives only a relative information about the distance to an object like binocular disparity, it tells how much closer or farther on object is from the fixated one.

It is also worth to note that motion parallax has been described to be sufficient for depth perception, in case on blindness of other depth cues. Experiments shows that motion parallax is sufficient only when the spatial information is rich enough.(Texture presence instead of uniformed colored surface).

Optic flow : case of a moving observer

The optic flow is more related to an high-level visual process to extract depth from motion, and linked to the motion parallax, than a simple and

strong depth cue. [9] advanced that when an observer is moving, the image motion is structured and depends on the structure of the 3D environment, its oriented surface, and the observer's motion. He introduced the concept of motion gradients to describe the motion of regions, their quantity, (speed) and direction. An observer moving leftward while fixating a point in the middle of a line along depth (by example a straight road directed to the horizon) will lead to a relative motion leftward for the closer points on the line (and the closer they are, the faster they will move), rightward for the farthest one.

Optic flow : case of moving objects

In the visual field, an object moving with respect to the observer enables also to perceive the depth. Relative movement of points depending of their positions on the object give to the HVS the distance information : which points of a surface is closer, which points is farther. [10] described the phenomenon through the kinetic depth effect (KDE). They backprojected the shadow of a 3D bent-wire figure. When the figure is static, the wire is stationary, no depth is perceived. Then, when the wire figure is rotated, it pops into a 3D shape. Recovering depth information from object rotation is nevertheless ambiguous, the 2D retinal motion could be perceived as a figure which is deforming over the time. But people perceived instead a rigid object consistent with the moving image : an object in rotation.

Texture accretion/deletion

Still in the moving context, a further source of depth information can be find in the appearance-disappearance of texture behind a moving edge [11] : the accretion-deletion of texture. As the edge always belongs to the closer surface, and the occluded-disoccluded texture to the farther surface, depth from this motion can be perceived. As a parallel can be drawn between motion parallax and binocular disparity, accretion/deletion of texture revealed over the time what is revealed across the views of left and right eyes in da Vinci stereopsis : occlusion information.

1.2.4 Pictorial information

The remaining depth cues are called pictorial cues because there are available in static monocularly viewed pictures, i.e they are signals obtained independently of any stereopsis or motion. However, these cues are important, sufficient to extract the depth from 2D pictures, and can even overcome the

stereo depth information in tricky experiments where stereo is reversed by optical devices.

Perspective projection

The perspective projection is a generic term that regroup all pictorial sources of information coming from the projection of a 3D scene on a 2D surface, either a picture or a retina.

Convergence of parallel lines / Vanishing point

The parallel lines in a 3D world rarely projects as parallel lines but as line converging to a vanishing point on the horizon line. (the distance between parallel lines becomes increasingly small with the increasing distance to the observer). Importantly, there are infinitely many vanishing points on the "horizon" line of other planes in 3D space (as shown on figure 1.5).

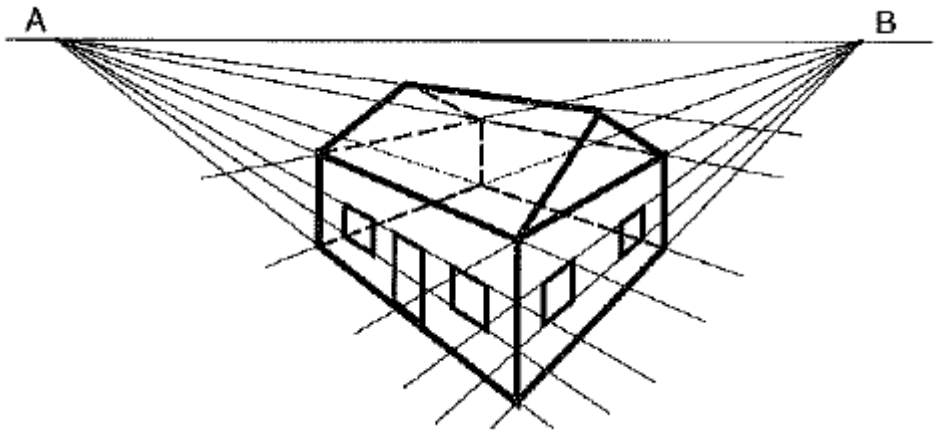

FIGURE 1.5 – A draftsman drawing in two-point perspective projection, where the house parallel edges converge to two vanishing points A and B on the horizon line (from [4])

Position relative to the horizon of a surface

The height of objects on the level plane relative to the horizon gives an extra information on their possible position in space. This cue is said to be quantitative [12], as geometrically speaking, the distance d from the observer to any point P on the surface can be determined from the horizon angle A (the angle between the line of sight to the horizon and the line of sight to

the point on plane), and the perpendicular distance to the surface h (figure 1.6).

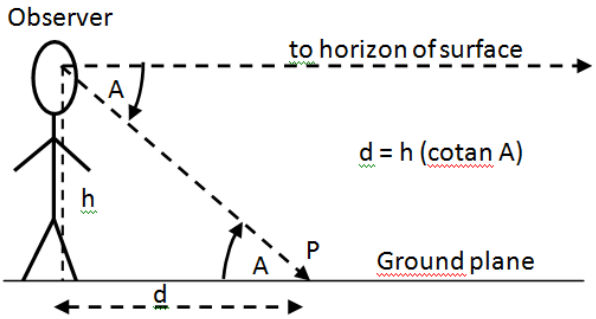

FIGURE 1.6 – Distance as a function of the horizon angle to a point on a surface (from [4])

As observers have information of the height of their eyes above the ground plane, and as long as they know the horizon lines, the angular distance to the horizon gives an efficient quantitative information about distance to objects along it.

Relative size

Depth in perspective projection also involves size comparison of identical -or supposed to be- objects : distant objects will then project smaller images onto the retina. Considering h the height of the object, a the visual angle it subtends, the right triangle has a height h and and an angle a. The distance to the object d can be expressed with the tangent of the angle a :

$$d = \frac{h}{tana} \tag{1.1}$$

The problem is that in order to know the distance, the size of object h must be known. This information of object size is ambiguous because we can't tell from a given image size whether it's a smaller object nearby or a larger one farther away. The HVS uses a heuristic to overcome this indetermination. It assumes that two identical objects have the same actual size so that their relative distances can be determined from relative image sizes projected on retina. Painting in 1.7 illustrates this idea where the different sizes of men recreate the depth effect.[13] recently conclude from ground/ceiling experiments that the perceived layout of objects in a scene depends both on the positions of the objects relative to a background surface and relative to the horizon.

FIGURE 1.7 – "Jour de pluie à Paris", painting by Gustave Caillebote, 1877

Familiar size

Familiar size cue could seem similar to relative size but involves an "a priori" of the size of the objects. By familiar, we mean that most of objects have a characteristic size or range of size with which people are familiar. For example, the majority of adult men vary between 1.6 and 1.9 meters, table are about 80 cm above the ground etc. If the size of a unique object is known to the observer, then the previous size-distance equation can be solved for its actual distance to the observer. However, [4] underlined the fact that this "automatic knowledge" process is unconscious.

Texture gradients

Gibson [5] also mentioned texture gradient as an important cue among perspective projection. Systematic changes in size and shape of texture elements, i.e stationarity and regularity of environmental surfaces provide a good information about surface orientation. As the size of texture elements decrease with distance, it can be used to estimate the relative distance to the different parts of the involved surface and also its orientation. Once again, an heuristic assumption is made. The distance to texture elements based on their image-size will be accurate only if these elements (texels) are objectively similar in size. Projected shape of texture elements can also carry

information about the orientation of the surface (illustration in figure 1.8).

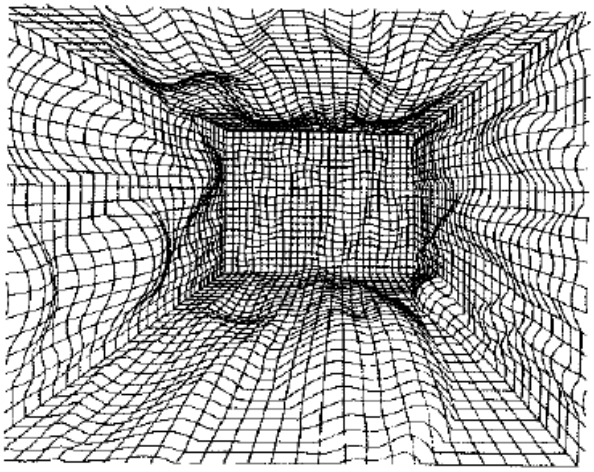

FIGURE 1.8 – Artificial texture gradients. Artificial surfaces of arbitrarily complex shapes can be rendered bu using identical texture elements (from [14])

Different algorithms were developed for estimating surface orientation from textural shape analysis, based on different heuristic assumptions.[15] devised an algorithm for slant and tilt recovering of small patches based only on a weaker assumption : texture elements are approximately invariant over small translations along the surface. Concretely, they find the best-fitting parameters of surface orientation and curvature accounting for shape and size between nearby elements.

Edge interpretation

The edge interpretation gives another pictorial information about depth : the occlusion from an object, by an opaque object -delimited by edges- and appearing nearer the viewer gives an ordinal depth relation : which object is farther than any other and how much it is farther. This edge interpretation is a relative rather than absolute, qualitative rather than quantitative cue. Nevertheless, these relations are available from any distant viewpoints, since occluding object is opaque. Different computational models of edge interpretation have been given in the past years, but no link has been made with the actual processes involved in human vision.

Shading information

Like previous aspects of depth perception, the visual analysis of shading often rests on heuristic assumptions to solve an underconstrained inverse problem. One of them is that we implicitly assume that illumination come from above, which is almost always the case in our environment. Figure 1.9 shows surfaces with 2 rows of indentations. The top ones appear to be

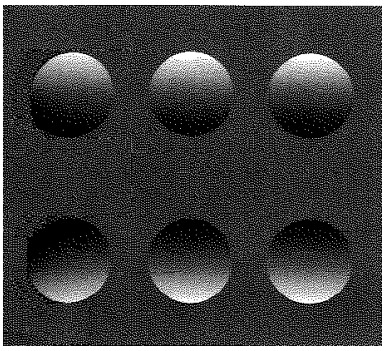

FIGURE 1.9 – Direction of illumination and perceived convexity (from [4])

convex bumps, and the bottom ones to be concave. This perception is based on heuristic and is veridical only if the illumination comes from above in this scene.(if you reverse the figures, this will reverse the perceived relief).

This assumption of illumination resolves surface orientation determination problem, and give good indications about depth.

Aerial perspective

Aerial perspective (sometimes called atmospheric perspective) gives an additional indication about the depth in particular conditions. Large objects viewed from far away, like building, mountains, sea, appear "washed out" i.e with a lower contrast because of additional atmosphere. The farther they are, the more the atmosphere contains particles or pollutants that scatter light. It can also take a blue tint as the particles in atmosphere scatter longer wavelengths of light than shorter wavelength (in the range we perceive the blue).

1.2.5 Integration of these different information sources

Perceiving depth in a scene involves numerous widely different sources. But how the HVS accomplish the integration of all these sources, since they all result in the same interpretation of surfaces oriented in depth ? By putting

different cues into conflict, scientists try to understand how the visual system **integrates** the different information : is there any dominance, interaction or compromise between two conflicting sources ?

Some studies show that pictorial source of information almost always wins out in case of conflict, but more recently, [16] [17] demonstrated that the weak fusion model is also a good candidate to the compromise solution. It consists in numerous different estimations of depth may computed independently and in parallel, before an integration by a mathematical combination (averaging, additive, multiplicative) at different locations in the depth map. The result is then a compromise between the different cues.

This description run counter to the fact that only absolute depth maps can be combined to have good depth results. As we have seen, **quantitative** sources of **relative** depth such as binocular disparity or motion parallax need to interact with other cues to produce **absolute** depth information. For example, binocular disparity specifies only ratios of distances to surfaces. On the contrary, convergence cue specify absolute depth, only for the fixated object. Together, they determine the absolute distance to object in the field of view. A complete depth map can then be computed if these two sources interact, and these can be generalize to any other absolute source on information. The knowledge of one absolute distance determines the distances to every object.

The model of modified weak fusion by [17] takes into account some limited interactions between depth sources we mentioned. Interestingly, he describes a representation where different sources of information are upgraded to the level of one absolute depth for the whole visual field : the "depth map". With analogy to saliency models, the "promotion" is the process that upgrade information from a depth source to this metric depth map. Convergence, accommodation can then be used to promote binocular disparity via scaling, before a combination of promoted depth maps by numerical integration is finally realized.

To conclude, the combination of the different depth cues into a unique representation of the 3D layout of surface remains a complex and an open issue. Even if it sounds plausible, there is no systematic physiological evidence of these interaction integration-based scenarios. The research on these topics are just beginning.

1.3 Eye tracking

Eye tracking is a technique which records the eye-movements so that the researchers can know both where a person is looking at any given time and the sequence in which their eyes are shifting from one location to another[18]. Eye tracking plays a substantial role in the research of psychology, biology and also computer vision. Especially in our future research of visual attention, it is necessary to use the eye-movement data from eye tracking experiment as the ground truth to evaluate the performance of computational models.

In this section, several eye tracking techniques will be first introduced. After that, the measurements of different eye-movements will be described. And we will introduce the setup of eye tracking experiment and how to generate the human saliency map as the output of the experiments.

1.3.1 Introduction of eye tracking technology

The technology of eye tracking appeared firstly more than 100 years ago for the research of reading [19]. Different techniques have been applied in eye tracking. For instance, the "electro-oculographic techniques" need to put electrodes on the skin around the eye so that eye movements can be detected by measuring the differences in electric potential. Some other methods relied on the wearing of large contact lenses covered the cornea (the transparent front part of the eye) and sclera (the white part of the eye). A metal coil was embedded around the lens so it moved along with the eye. The eye-movement could be measured by fluctuations in an electromagnetic field when the eye was moving[20]. These methods affect observers' eye-movement and are inconvenient to implement. Nowadays, modern eye trackers use video-based technologies to determine where a person is looking (i.e., their so-called "point-of-regard")[18]. Corneal-reflection/pupil-centre method is used by most commercial eye trackers to measure the point-of-regard. The corneal reflection (shown in figure 1.10 and figure 1.11) is also known as (first) Purkinje image. During the eye tracking, a camera focuses on one or both eyes to get images. Contrast is then used to get the location of the pupil, and infrared light is used to create a corneal reflection. By measuring the movements of corneal reflection relative to the pupil, it is then possible to know the head movement, eye rotation, the direction of gaze and consequently the point-of-regard.

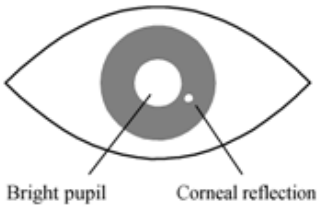

FIGURE 1.10 – Corneal reflection and pupil as seen in the infrared camera image.[18]

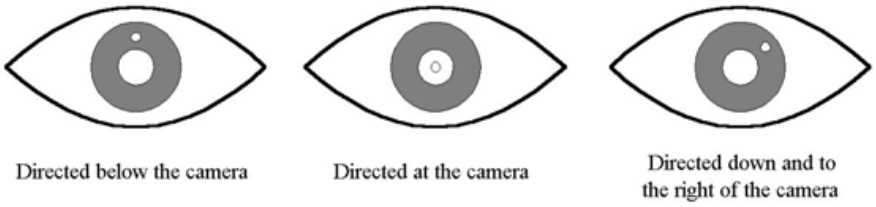

FIGURE 1.11 – Corneal reflection position changing according to point of regard.[18]

1.3.2 Introduction of eye-movement

It is assumed that what a person is looking at indicates the thought "on top of the stack" of cognitive processes [21]. This "eye-mind" hypothesis means that the recordings of eye-movement can provide a trace about where a person's attention is being directed. Different kinds of eye-movements can be recorded thanks to the development of eye tracking techniques. Two main measurements are "fixations" and "saccades". Some other measurements, including "gaze" and "scanpath", stem from these two basic measures. Pupil size and blink rate are also two measurements usually studied.

Fixation

Fixation is the location where the eyes are relatively stationary. Fixation last for 218 milliseconds on average, with a range of 66 to 416 milliseconds[18]. Several metrics derived from fixation are described as follow :

- Number of fixations overall. Goldberg et al.[22] suggest that more fixations overall indicates a less efficient visual search of the scene .
- Fixations per area. Experiments show that more fixations on a particu-

lar area indicate a greater interest or importance of a target[23]. And it may also mean that the target is complex in some way and difficult to encode [21]. But Jacob et al. [24] suggest that, in a search task, a higher number of fixations often mean a greater uncertainty in recognizing a target item.

- Fixations duration. A longer fixation can be interpreted in two ways : it's difficult to extract information, or the object is more engaging in some way [21].
- Gaze. Gaze is defined as the total duration of all fixations within a prescribed area. It is also referred to as "dwell", "fixation cluster" or "fixation cycle"[18]. It could be used to compare attention distributed between targets, or to measure the anticipation in situation awareness if longer gaze fall on an area of interest before a possible event occurring [25].
- Fixation spatial density. Cowen et al.[26] suggest that highly concentrated fixations in a small area means focused and efficient searching, and evenly spread fixations indicate widespread and inefficient search . It is also found that if an object contains an area which is with highly concentrated fixations, the object is tended to be considered as with high importance [23].
- Repeat fixations. It is also called "post-target fixations". A higher number of fixations off-target after the target has been fixated (i.e., a lower number of repeat fixations) means that the target lacks meaningfulness or visibility [22].
- Time to first fixation on-target. A shorter time to first-fixation on an object or area indicates that the object or area has better attention-getting properties [27].
- On-target fixations. It means the number of fixations on-target divided by the total number of fixations. A lower ratio means lower search efficiency [22].

Saccades

Saccades are those quick, simultaneous movements of both eyes in the same direction [28]. They are also the fast movement of eyes occurring between fixations. It is generally believed that no encoding takes place in human visual system during saccades, so vision is suppressed and it is difficult for us to get any clues about the complexity or salience of an object from the happening saccades. However, we could still get some information about visual perception from several saccade metrics as follow :

- Number of saccades. A larger number of saccades indicate that more

searching take place during the observation[22].
- Saccade amplitude. Saccade amplitude is computed by measuring the distance between one saccade's start point (a fixation) and its end point (another fixation). Larger amplitude indicates the existence of more meaningful cues, since the attention is drawn from the distance[29].

Scanpaths

Scanpath is a metric derived from the measurement of both fixations and saccades. It means a complete saccade-fixate-saccade sequence. The area covered by scanpath indicates the area observed. A longer scanpath means less efficient searching[29]. Besides, we can compare the time spent for searching (saccades) to the time spent for processing (fixation) in a scanpath. A higher saccade/fixation ratio means more searching or less processing.

Blink rate and pupil size

The blinking of eyes and the changing of pupil size are two eye-movements that could also be recorded by eye tracking experiments. They can be considered as a cue of cognitive workload. A lower blink rate is assumed to indicate a higher cognitive workload [30], and a higher blink rate may indicate visual fatigue [31]. The changing of pupil size also indicate some kinds of cognitive effort [32]. However, the blink rate and the pupil size can be easily affected by many factors during the observation, such as the luminance of environment. Because of this, blink rate and pupil size are not widely used in eye tracking research.

1.3.3 Introduction of eye tracking experiment procedure

The setup of the experiment

Generally, all subjects of an eye tracking experiment should be adults who have either normal or corrected-to-normal visual acuity. For some experimets which are to get the bottom-up information from the subjects, the subjects should be naive to the purpose of the eye tracking experiment, and have never seen the images or videos which will be presented in the eye tracking experiment.

The eye tracking experiment needs to be carried out in a dark room. Visual distractions (e.g., colorful or moving objects around the screen or in the testing environment) should be eliminated. The experiment room should

be also silent, since evidences shows that some kinds of sound affect eye movements [33].

To each subject, a calibration process is obligatory before the start of the eye tracking experiment. This calibration works by displaying some dots one by one on the screen. For each dot, if the eyes fixate on it for longer than a certain threshold time and within a certain area, the system records the relationship between eye position and the specific x-y coordinate on the screen. This procedure is repeated over a 9 to 13 point grid-pattern to gain an accurate calibration over the whole screen[18]. If the duration of experiment is long, the calibration procedure should then be repeated at regular intervals to maintain an accurate point-of-regard measurement.

During the experiment, stimuli (i.e. images or videos) are presented in a random order. For each image or video, there should be enough time for the subjects to look. Evidences suggest that the behavior of visual perception varies according to time[23, 34]. The interval of each stimulus should be also long enough to eliminate the effect of the previous stimulus. Sometimes the first couples of stimuli are considered as the training group, the eye tracking data of them are discarded from the final result[34].

The measurement of different eye-movements

As mentioned previously, there are two kinds of basic eye-movements : fixations and saccades. Hence, the process of fixation identification, which separates and label fixations and saccades coming from raw eye-tracking data, is an essential part of eye-movement analysis.

The fixation identification algorithms need to identify not only the fixations and the saccades taking place between one fixation and another, they need also to figure out those smaller eye movements that occur during fixations, such as tremors, drifts, and flicks[35]. The fixation identification a critical aspect of eye-movement data analysis, and it can have significant effects on later analysis. Evidences show that different identification algorithms can produce great different interpretations even when analyzing the same eye-tracking data

Salvucci et al. [35] suggest that the classification of fixation identification algorithms can be with respect to spatial or temporal characteristics. For spatial characteristics, three criteria have been used to distinguish three primary types of algorithms :

- Velocity-based. These algorithms take advantage of the fact that fixations points have much lower velocities compared with the saccade points. Generally, the sampling rate of an eye-tracker is constant, so we can ignore the temporal component implicit in velocities.

- Dispersion-based. This kind of algorithms emphasize the spread distance (i.e., dispersion) of fixation points. It assumes that fixation points generally occur near one another, but saccades points are far away from others.
- Area-based. This kind of algorithms identify the points locating within given areas of interest (AOIs) which represent relevant visual targets.

For temporal characteristics, two criteria are included :

- Duration information. This criterion is based on the fact that fixations are rarely less than 100 ms and usually in the range of 200-400ms[35].
- Local adaptivity. This criterion means that the interpretation of a given point is influenced by the interpretation of temporally adjacent points.

Velocity-based Algorithms

Among all the velocity-based algorithms, Velocity-Threshold Identification (I-VT) is one of the simplest algorithms to realize. The velocity profiles of eye movements show two distribution of velocities : low velocities for fixations, and high velocities for saccades. These velocity-based discriminations are straight forward and robust.

I-VT calculates firstly point-to-point velocities for each point. Each velocity is computed as the distance between the current point and the next (or previous) point. Each point is then classified as a saccade point or fixation point based on a velocity threshold : if the velocity is higher than the threshold, it becomes a saccade, otherwise it becomes a fixation point. Finally, I-VT translate each fixation group into a <x, y, t, d> representation. <x, y> represent the centroid of the points, t and d means the time of the first point and the duration of the points respectively.

Dispersion-based Algorithms

Dispersion-based Algorithms utilizes the fact that fixation points tends to cluster closely together because of their low velocity. The Dispersion-Threshold Identification (I-DT) is one of the Dispersion-based algorithms. It identifies fixations as groups of consecutive points within a particular dispersion. To help alleviate equipment variability, it incorporates a minimum duration threshold of 100-200ms [36] because of the fact that fixations usually have a duration of at least 100ms.

The I-DT algorithm uses a moving window to cover consecutive data points. The moving window begins at the start of the protocol. It contains initially a minimum number of points which is determined by the given duration threshold. The I-DT then compute the dispersion of the points in the

window by summing the differences between the points' maximum and minimum x and y : $D = [max(x) - min(x)] + [max(y) - min(y)]$. If the dispersion is above a dispersion threshold, the window moves to the following point. If the dispersion is below the threshold, the window represents a fixation and will be expended until the window's dispersion is above the threshold. The final window is marked as a fixation which centers at the centroid of the points and be with the given onset time and duration. Two parameters are required in I-DT, the dispersion threshold and the duration threshold.

Area-based Algorithms

This algorithm identifies only fixations that occur within specified target areas. It also utilizes a duration threshold to help distinguish fixations in target areas from passing saccades in the areas. Area-of-Interst Identification (I-AOI) is one of this kind of algorithms. At first, I-AOI labels points within a specified area as fixation points for that target, labels points outside the area as saccades. rerererIt then collapses those consecutive fixation points for the same target into fixation groups, removing saccade points. After that, fixation groups that do not span the minimum duration threshold are removed. Finally, I-AOI maps each fixation group to a fixation at the centroid of its points.

The output of the experiment : Human Saliency Map

Saliency map is a topographically arranged map that represents visual saliency of a corresponding visual scene[37]. Because it is not easy to use the raw data which is from an eye tracker to quantitatively compare or analysis the eye movements, the generation of saliency map is crucial. This kind of saliency map is usually called "Human Saliency Map" since it comes from the actual movements of the human eye (compared with the "Predicted Saliency Maps" coming from computational models).

The generation of human saliency map is under the assumption that it is an integral of Gaussian point spread functions locating at the positions of a set of successive fixations :

$$S_{human} = H(x) = \frac{1}{K} \sum_{k=1}^{K} h(x_k)$$

where $H(x)$ represents the human saliency map and $h(x_k)$ represents the point spread function (PSF). It is assumed that each fixation x_k gives rise to a Gaussian distributed activity. And the width of the Gaussian kernel

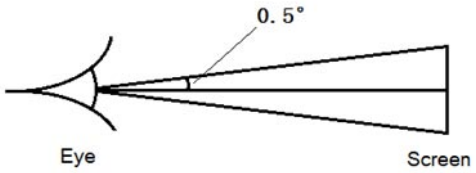

FIGURE 1.12 – Example of the calculation of PSF's radius

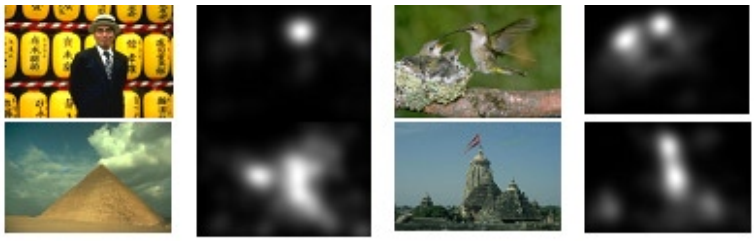

FIGURE 1.13 – Some examples of human saliency map. [23]

depends on the size of fovea and the distance between the observer and the screen. Figure 1.12 is an example of the calculation of PSF's radius.

Generally, the area covered by 0.5-degree visual angle is considered as the area covered by the fovea. Therefore, we can use trigonometric function and the distance between eye and screen to computer the size of Gaussian kernel.Figure 1.13 shows some examples of saliency maps generated by this method.

1.4 Conclusion

This section presented the fundamental elements enabling human perception of depth and the eye tracking principles. Human vision and human perception of depth are not fully understood because the organs involved in the vision process (eyes, brain) are still under investigation. However, as far as it is known, human perception of depth requires both monocular cues and binocular cues. Eye tracking technique allows researchers to know where a person is looking. This is an helpful tool in order to design perceptual models.

Chapitre 2

Perceptual modelling

The experience of the user of a communication system determines its quality. For this reason, engineering metrics, or models are meant to predict the performance of this experience. Those techniques are based on models of human perception. This chapter will then address 2D perceptual modelling, then visual attention models will be presented, and the last section introduces 3D perceptual modelling.

2.1 2D perceptual modelling

Human Visual System (HVS) modelling is meant to perceptually optimized image processing. The HVS models can be classified into two types : neurobiological models and models based on psychophysical properties of the vision. The models based on neurobiology estimate the actual low-level process in human visual system including the eye and optical nerve. However, this type of models are not widely used, because of their overwhelming complexity [38].

The psychophysical models, which are built upon psychophysical experiments, are used to predict aspects of the human vision. They are typically implemented in a sequential process shown as follow :

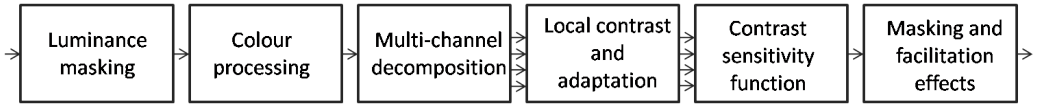

FIGURE 2.1 – Psychophysical HVS model [38]

2.1.1 Luminance masking and Color processing

The first stage in the processing chain of HVS modelling is the so-called luminance masking or lightness non-linearity. This stage take care of the non-linear perception of luminance by the HVS. The color processing stage is modelled by the transformation of the input signal into an adequate color space, usually based on opponent colors [39]. As parts of the same process, luminance masking and color processing take place simultaneously in HVS.

2.1.2 Multi-channel decomposition

Thanks to the measurements of receptive fields in the primary visual cortex, it is known that the HVS bases its perception on multiple channels that are tuned to different range of spatial frequencies and orientations. These channels exhibit approximately a dyadic structure[40]. This behavior can be modelled by a multi-resolution filter bank or a wavelet decomposition[38]. The example of the multi-resolution filter is the cortex transform which is a flexible multi-resolution pyramid. Besides, it is believed that there are channels processing different object velocities or temporal frequencies [39]. These include on temporal low-pass and one or two band-pass mechanisms in the HVS. They are respectively referred to as sustained channels and transient channels.

2.1.3 Local contrast and adaptation

After the input signal is decomposed into channels, the local contrast and adaptation stage takes place. It is widely accepted that the response of the human visual system depends much more on local luminance variations to the surrounding than the absolute luminance. This property is known as Weber-Fechner law [41]. Contrast is widely used in vision models to measure this relative variation. For the simple patterns, a contrast measure is simple to define. However, it is much more difficult to define the contrast measure in complex images since it depends on the image content. And the contrast measure is also influenced by the adaptation to a specific luminance level or color.

2.1.4 Contrast sensitivity function

The human visual system has different sensitivity to different spatial frequencies. The HVS usually has a decreasing sensitivity for higher spatial

frequency. This phenomenon is modelled by the Contrast Sensitivity Function (CSF).

It is not easy to correctly model the CSF for color image, thus the sensitivity to color and the sensitivity to pattern is assumed independent for simplicity. The CSF for each channel of the color space are modeled independently. The CSFs for achromatic channel are summarized in [42], and the CSFs for color channels are described in [43, 44, 45]. Besides, [46] describes the details of an efficient CSF-modelling method in combination with the wavelet decomposition. Figure 2.2 shows a 2D CSF function. Figure 2.3 shows the CSF functin for different luminance levels.

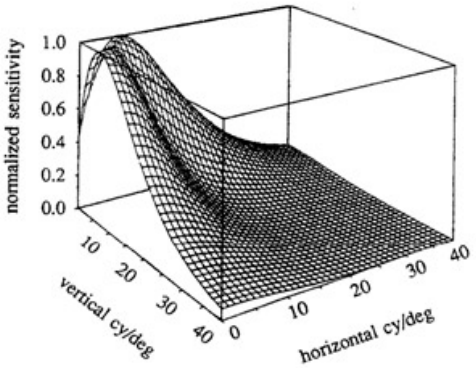

FIGURE 2.2 – The normalized two-dimensional CSF model [47]

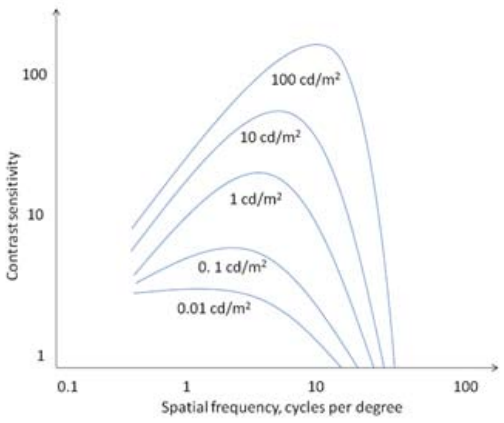

FIGURE 2.3 – Contrast sensitivity function for different luminance levels, adapted from [38]

The human visual system has also different sensitivity to different temporal frequencies. The simplest way to measure temporal sensitivity is to present an observer with a small flickering light, and find the smallest amount of flicker that the observer can detect. When temporal sensitivity is measure at a range of flicker rates ,the results can be plotted in a temporal contrast sensitivity function. Similar to the shape of spatial CSF, the shape of temporal is low-pass or slightly band-pass.

Furthermore, the spatio-temporal contrast sensitivity functions are utilized to describe the interaction between spatial and temporal frequencies. These sensitivity functions are commonly used in vision models for video [48]. The implementation of spatio-temporal contrast sensitivity function is described in [49] and [50].

2.1.5 Masking and facilitation effects

Masking effect means that a visual stimulus which is visible by itself cannot be detected due to the presence of another visual stimulus. Facilitation effect can be considered the opposite effect of masking : a visual stimulus which is not visible by itself becomes visible due to the presence of another stimulus. Masking effect explains why some distortions disturbs in some region while they are hardly noticeable in another region.

Several different spatial masking effects are described in[51, 52], but this distinction is not clear-cut [39]. The terms contrast masking, edge masking, texture masking are usually mentioned to describe the masking effects caused by strong local contrast, edges, and local activity, respectively. Temporal masking is a brief elevation of visibility threshold caused by temporal discontinuities in intensity, e.g. at scene cuts. It occur not only after a discontinuity but also before [53].

2.2 Visual attention modelling in 2D

2.2.1 Introduction of Visual Attention

This is William James's suggestion of attention which dates back to 1890 : "Everyone knows what attention is. It is the taking possession of the mind, in clear and vivid form, of one out of what seem several simultaneously possible objects or trains of thought. Focalization, concentration of consciousness are of its essence. ... "

Visual attention is one of the most important features of the human visual system. It concerns which part of an image/video attracts the gaze of an observer. Figure 2.4 gives an example of visual attention :

FIGURE 2.4 – A example of visual attention

When people look at the figure, most people will immediately notice the sheep eating grass at the centre of the image. Only a small number of people will pay attention to the fence, or wonder the species of the trees which are far behind that sheep. This is visual attention. To J.K. Tsotsos, visual attention is a mechanism which has at least the following basic components[54] :

- The selection of a region of interest in the visual field ;
- The selection of feature dimensions and values of interest ;
- The control of information flow through the network of neurons that constitutes the visual system
- The shifting from one selected region to the next in time.

In people's everyday life, the environment presents far more perceptual information than human can effectively process. In order to keep the essential visual information, humans have developed a particular strategy, first outlined by James. This strategy involves two mechanisms. The first one refers to the sensory attention driven by environmental events, commonly called bottom-up or stimulus-driven. The second one is the attention to both external and internal stimuli, usually called top-down or goal-driven[55].

Why do most people first notice the sheep ? It is because the sheep is visually salient due to its color and shape which are very different from the ones of the grass. This kind of visual differentia is generated by the physical input, which can be analyzed by the computers thanks to the technology of image processing. This is called bottom-up or stimulus-driven. It is also the reason why the flickering advertising banners on the webpage can attract people's attention.

Besides bottom-up, there is something else. The sheep is at the centre of the image. As an individual unit, it is the biggest one. It is the main character of this scene. Due to this, it is reasonable to make a conjecture that the sheep may be the object which the photographer wants to highlight. So people pay more attention to the sheep. This is called top-down or concept-driven or goal-driven. It is also the reason why students focus their attention on the professor during the class.

2.2.2 Computational model of visual attention

The classification of computational model

Most of the researches of visual attention are to generate a saliency map indicating where the most visually interesting regions are located. In the past few years, several models which used different mathematical tools, have been proposed to compute the saliency maps. According to Le Meur et al.[56], these computational models can be grouped into three different categories : Hierarchical models ; Statistical models ; Bayesian models.

- Hierarchical model. The computational architectures of the hierarchical models are similar. This kind of models is characterized by using hierarchical decomposition which might involve a Gaussian, a Fourier-based or wavelet decomposition. Then a difference of Gaussian is applied on the subbands. After the salience decomposition level has been estimated, different methods are used to aggregate the information across all the

levels to generate the final saliency map.

- Statistical model. This kind of models utilizes probabilistic methods to compute the saliency value of each location. It measures different features of a current location and the features from the regions surrounding the current location. All the information used in these models is from the current image. The selection of features is critical.
- Bayesian model. The most different part of this kind of models from the others is the combination of prior knowledge and bottom-up saliency. The prior knowledge might concern, for example, the statistic of visual features in the natural scene, the distribution of various features, and so on. Itti and Baldi's 'Surprise theory'[57] can be considered as belonging to this group. They proposed the definition of Surprise which measures the distance between posterior and prior beliefs of the observers.

The features utilized in the computational models

The performances of all the three kinds of models mentioned previously are close. On the other hand, one thing that can greatly affect the performance of the computational models is the selection of visual features. Some features have been used for a long time and have been shown to be efficient in the computation of saliency map. For instance, color, intensity, orientation, contrast, edge strength, locations are the features used widely in many models.

Besides these features, new features are recently found to be efficient to affect the visual attention. Kootstra et al. [58] suggested that symmetry could be considered as a visual feature. They proposed an operator that computes the symmetry value of each location and generated the saliency map according to the symmetry value. In the 3D case, Jukka et al.[59] show that the complex stereoscopic structures and the structures nearer than the actor captured the gaze of observers. This might suggest that the complexity of structure and the depth information could also be used as feature in the prediction of saliency map. Gal and Cohen-Or[60] suggested that the saliency of a region can be computed according to its size relative to the whole object, its curvature, the variance of curvature, and the number of changes of curvature.

Some principle visual attention computational models

The model of Itti et al. At 1998, Itti et al.[61] proposed this computational model which is one of the most earliest and representative models. It has been also widely utilized as a reference to compare or evaluate the performance of other models. This model builds on a biologically plausible architecture proposed by Koch and Ullman [62]. It explains human visual search strategies by relating to a so-called "feature integration theory" [63]. This model's framework is also based on the concept of saliency map which is a two-dimensional topographic representation of conspicuity for each pixels of the image.

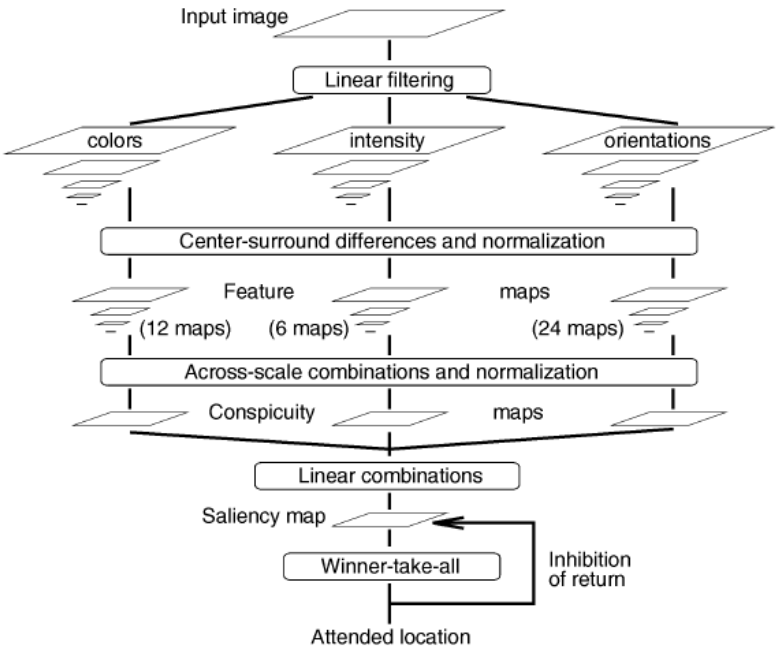

FIGURE 2.5 – The diagram of Itti et al's model.

Visual input provided in the form of static color images is first computed by a set of linear "center-surround" operations akin to visual receptive fields. It is decomposed into three channels : colors, intensity, orientations. This process yields a set of topographic feature maps :

- The first set of feature maps is constructed for intensity contrast. There are two kinds of intensity contrasts : one is detected by neurons sensitive to dark centers on bright surrounds and the other one is detected by

neurons sensitive to bright centers on dark surrounds. Here, both of them are simultaneously computed in a set of six maps $I(c,s)$, with $c \in \{2,3,4\}$ and $s = c + \delta, \delta \in \{3,4\}$

$$I(c,s) = |I(c) \ominus I(s)| \tag{2.1}$$

where c represets "center" and s represents "surround"

- The second set of maps is constructed for color channels. A so-called "color double-opponent" system exists in cortex. It refers to 2 pairs of colors : red/green and blue/yellow. Due to this, maps $RG(c,s)$ and maps $BY(c,s)$ are created :

$$RG(c,s) = |(R(c) - G(c)) \ominus (G(s) - R(s))| \tag{2.2}$$
$$BY(c,s) = |(B(c) - Y(c)) \ominus (Y(s) - B(s))| \tag{2.3}$$

- The third set of maps is concerned with the local orientation contrast between the center and surround scales.

In total, 42 feature maps are computed : six for intensity, 12 for color, and 24 for orientation. For the three channels, features maps are normalized and combined across scales and orientations into a "conspicuity maps" for each channel. Finally, an overall saliency map is obtained by linearly combining the channels.

This method provides good results, however, it can still be improved in some respects[64] : Firstly, this model allows each location compete for conspicuity within each channels, but it separates each channel independently. The second one is that it has many parameters that need to be hand-selected.

The model of Le Meur et al. The work of Le Meur et al. is another very representative model. This is also a bottom-up model based on Feature Integration Theory (FIT) from Treisman and Gelade[63] and the biologically plausible architecture proposed by Koch and Ullman[62]. This model was first described in [55] and then modified in [65], which added the measurement of movement.

The model proposed in [55], which is for still images, builds on a coherent psychovisual space to obtain a saliency map. Being justified with psychophysical experiments, this space is used to combine visual features, such as intensity, color, orientation, spatial frequencies. These features are normalized to their individual visibility threshold. The visibility threshold associated

to each value of each component is calculated by the usage of accurate non-linear models which simulate visual cells behaviors.

This proposed computational bottom-up model contains several properties of human visual cells in mind. Three aspects of the vision process are tackled, the visibility, the perception, and the perceptual grouping.

The first vision process utilized is the visibility, which simulates the sensitivity which is limited in human visual system. A coherent normalization is first used for scaling all the visual data. After that, all the visual data is then grouped into a psychovisual space which bases on the following four mechanisms :

- Transformation of the RGB Luminance into the Krauskopf's Color Space. The relation of color spaces is given as follows :

$$\begin{pmatrix} A \\ Cr1 \\ Cr2 \end{pmatrix} = \begin{pmatrix} 1 & 1 & 0 \\ 1 & -1 & 0 \\ -0.5 & -0.5 & 1 \end{pmatrix} \begin{pmatrix} L \\ M \\ S \end{pmatrix}$$

- Early visual features extraction.
- Contrast sensitivity functions. CSF has been widely used to measure the visibility of natural images components. This model apply the 2D anisotropic CSF designed by Daly on the achromatic component[66]. On the other hand, the CSFs of the two color visual components Cr1 and Cr2 are modeled using sinusoidal color gratings.
- Visual masking. It refers to the modification of the differential visibility threshold of a stimulus because of the influences of the context. Due to the subband decomposition, three types of masking are defined here : intrachannel masking, interchannel masking and intercomponent masking. The intrachannel masking is the most important masking effect. This model apply the function designed by Daly [66] to model the intramasking effect for the achromatic component. On the other hand, the function designed by Le Callet[67] was used to model the intramasking effect for the chromatic components.

The second process is perception. The goal of this process is to determine the achromatic components which are necessary for the calculation of the saliency map. This process contains two mechanisms that can effectively detect the achromatic reinforcement by chromatic context and the center/surround suppressive interaction. The achromatic reinforcement takes advantage the color dimension which can efficiently guide the attention to the most salient

areas of our visual field. The center/surround suppressive interaction simulates the mechanism of the visual systems, which is used to select relevant areas and reduce the redundant incoming visual information.

The third process is perceptual grouping. It refers to the human visual ability that group and bind visual features to organize a meaningful higher-level structure. Perceptual grouping comprises numerous mechanisms. Facilitative interactions mechanism, the most common one, is concerned in this model. This facilitative interaction is usually named as contour enhancement or contour grouping. In this paper, two butterfly filters described by [68] are used for simulating the contour grouping. The two filters, $B^0_{\theta_{i,j,A}}$ and $B^1_{\theta_{i,j,A}}$, are obtained by a directional term $D_{i,j}(x,y)$ and a proximity term circle Cr blurred by a Gaussian filter G(x,y) :

$$B_{\theta_{i,j,A}} = D_{i,j}(x,y) \cdot C_r(x,y) * G(x,y) \tag{2.4}$$

with

$$D_{i,j}(x,y) = \begin{cases} cos(\frac{\pi/2}{\alpha}\varphi) & if -\alpha < \varphi < \alpha \\ 0 & otherwise \end{cases} \tag{2.5}$$

Finally, this computational model sums directly the output of the different achromatic channels to obtain a two-dimensional spatial saliency map. Furthermore, Le Meur et al. proposed combining a temporal saliency map with the spatial saliency map to measure the visual attention in video. The computation of temporal saliency map is based on the assumption that the contrast of movement is the most important attractor of attention. By the fusion of spatial saliency map and the temporal saliency map, the final saliency map is obtained (shown in figure 2.6).

Compared to other computational model, the advancement of this model is the usage of the coherent normalization of visual features. However, this model can still be improved. The combinations of more early visual features might effectively improve the performance. And it is not difficult to combine more features thanks to the coherent normalization method proposed in this paper.

The model of Zhang et al. The model of Zhang et al. [64] starts from an assumption that an important goal of the visual system is to find potential targets and build up a Bayesian probabilistic framework. From this framework, different kinds of saliency emerge in different ways. Bottom-up saliency emerge naturally as the self-information of visual features, and overall saliency emerges as the pointwise mutual information between the features

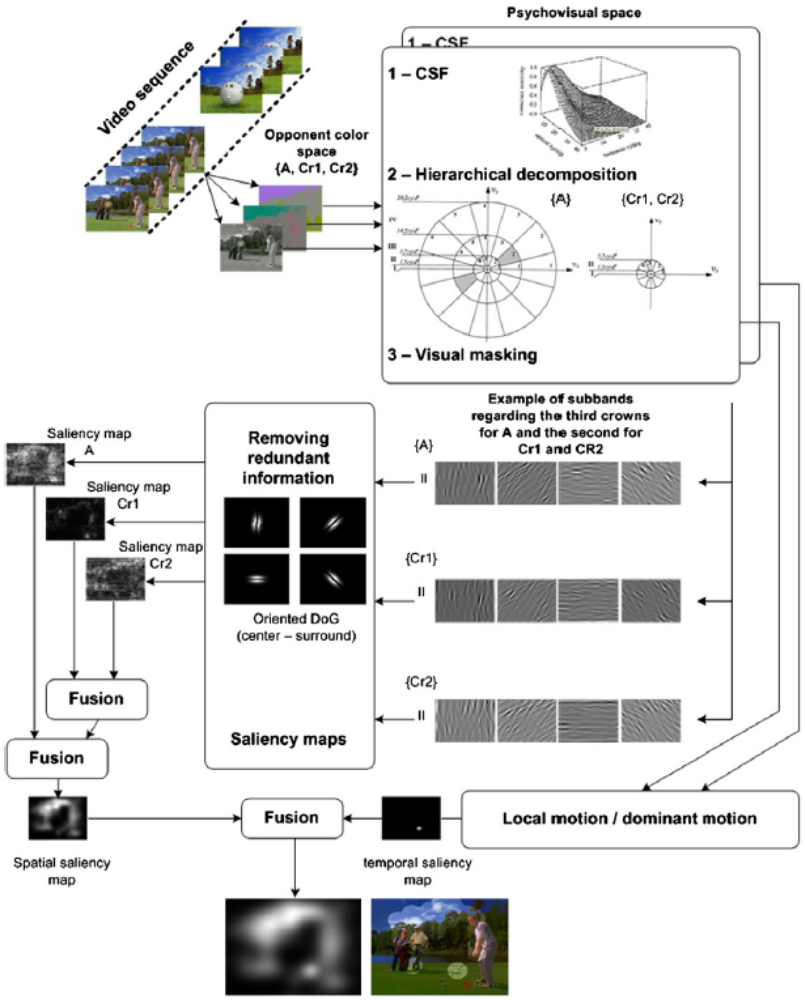

FIGURE 2.6 – The diagram of Le Meur et al's model [65]

and the target when searching for a target.

Compared to other bottom-up saliency measures, which are defined solely in terms of the image currently being viewed, this model is defined based on natural statistics collected from a set of images of natural scenes. And this is the reason why it is named SUN. Besides, compared to the others, it involves only local computation on images, without calculation of global image statistics or saliency normalization or winner-take-all competition.

Within the Bayesian framework proposed, z denotes a point, (in this

model, it is a pixel of the image), C denotes whether or not a point belongs to a target class, L denotes the location, F denote the visual features of a point. Saliency of z can be defined as $p(C = 1|F = f_z, L = l_z)$. Here, f_z represents the features values observed at z, l_z represents the location of z. Bayes' rule can be used here to calculate this probability :

$$S_z = p(C = 1|F = f_z, L = l_z) = \frac{p(F = f_z, L = l_z|C = 1)}{p(F = f_z, L = l_z)} \tag{2.6}$$

Due to the assumptions that features and location are independent and conditionally independent given C = 1 and the distribution of a feature does not change with location, the formulation is given by :

$$\log S_z = -\log p(F = f_z) + \log p(F = f_z, C = 1) \tag{2.7}$$

The first term on the right side of the equation, $-\log p(F = f_z)$, is the self-information. The rarer the features are, the more informative they are. The second term,$\log p(F = f_z, C = 1)$, is a log-likelihood term which favors feature values consistent with our knowledge of the target. It corresponds to the top-down effect when searching for a known target. The third term in the equation, $\log p(F = f_z, L = l_z)$, is independent of visual features and represents any prior knowledge of where the target is likely to appear. By omitting the third part of the equation, the location prior, the resulting express is obtained :

$$\log S_z = \log \frac{p(F = f_z, C = 1)}{p(F = f_z)p(C = 1)} \tag{2.8}$$

This equation can be called the pointwise mutual information between the visual feature and the presence of a target. It expresses the overall saliency. If it is the free-viewing condition, the log-likelihood term is unknown, so the overall saliency reduces to just the self-information term :$\log S_z = -\log p(F = f_z)$. The following part of the paper focuses on this term.

This model takes color images as input and calculates their saliency maps. It calculates the features in two different ways, DoG (difference of Gaussians) and ICA (independent component analysis), either of which is used in most saliency algorithms. By comparing the result and the data obtained from experiment, it is found that this algorithm under the proposed Bayesian framework performs as well as or better than existing algorithms in predicting people's eye fixations in free viewing.

The "Suprise" model of Itti et Baldi Itti et Baldi [57] proposed a concept of surprise which was central to sensory processing, adaptation, learning, and attention by describing a formal Bayesian definition of surprise.

Surprise quantifies how data affects observers by measuring the difference between prior and posterior beliefs of the observer. By using this framework, we can measure the extent to which people direct their gaze towards surprising items while watching a video.

Bayesian Definition of Surprise : In this paper, the surprise proposed is a general concept, which can be derived from first principles and formalized across spatio-temporal scales, sensory modalities, and, more generally, data types and data sources. But for a principled definition of surprise, it should always contain two elements :

- First, surprise can exist only in the presence of uncertainty, which can arise from intrinsic stochasticity, missing information, or limited computing resources.
- Second, surprise can only be defined in a relative, subjective, manner and is related to the expectations of the observer. The same data may carry different amount for different observers, and even for a same observer taken at different times.

Besides, due to the probability and decision theory, a consistent definition of surprise must also involve :

- Probabilistic concepts to cope with uncertainty
- Prior and posterior distributions to capture subjective expectations.

Measure of Surprise : It is necessary to capture the background information of an observer by the prior probability distribution $\{P(M)\}_{M\in\mathcal{M}}$ over the hypotheses or models M in a model space $\mathcal{M}$. D is a new data observation on the observer. Given the prior distribution, the fundamental effect of this new data observation is to change the prior distribution $\{P(M)\}_{M\in\mathcal{M}}$ into the posterior distribution $\{P(M|D)\}_{M\in\mathcal{M}}$ via Bayes theorem :

$$\forall M \in \mathcal{M}, P(M|D) = \frac{P(D|M)}{P(D)} P(M) \tag{2.9}$$

In this framework, surprise elicited by data is formally measured as some distance measure between the prior and posterior distributions. It can be realized by using the relative entropy or Kulback-Leibler (KL) divergence. Thus, surprise is defined by the average of the log-odd ratio taken with respect to the posterior distribution over the model class $\mathcal{M}$

$$S(D, M) = KL(P(M|D), P(M)) = \int_{\mathcal{M}} P(M|D) \log \frac{P(M|D)}{P(M)} dM \tag{2.10}$$

"wow" is consider as the unit of surprise. For a single model M, it may be defined as the amount of surprise corresponding to a two-fold variation

between $P(M|D)$ and $P(M)$, i.e. $\log P(M|D)/P(M)$.

According to the experimental result given in paper [57], it is found that the metric using surprise to predict the attractor outperforms all other computational models which include entropy metric, contrast metric, saliency metric, flicker metric, motion metric.

The model of Gao et al. Gao et al. [69] proposed a model which evaluated the plausibility of a generic principle for visual saliency : all saliency decisions are optimal in a decision-theoretic sense. The discriminant saliency hypothesis and a classical assumption, that bottom-up saliency is a center-surround process, are combined to derive a (decision-theoretic) optimal saliency architecture. Under this architecture, the saliency of each image location can be obtained by computing the discriminant power of a set of features with respect to the classification problem that opposes stimuli at center and surround.

The discriminant saliency hypothesis is that all saliency decisions about the state of the surrounding environment are optimal in a decision-theoretic sense, e.g., that have minimum probability of error. Discriminant saliency is defined with respect to two classes of stimuli : The first one is the class of stimuli of interest and the second one is a null hypothesis. With respect to these two classes and lowest expected probability of errors, the locations of visual fields that can be classified as containing stimuli of interest are denoted as salient. This can be accomplished by applying two mathematical processes :

- Defining a binary classification problem that opposes stimuli of interest to the null hypothesis ;
- Equate the saliency of each location in the visual field to the discriminant power (with respect to this problem) of the visual features extracted from that location.

In fact, the definition of saliency mentioned above is applicable to a large number of problems. For instance, both top-down and bottom-up saliency can be specialized by different specifications of stimuli of interest and null hypothesis. However, this model focuses on the problem of bottom-up saliency.

The "Discriminant center-surround saliency" proposed by this model can be considered as a result of combining the discriminant saliency hypothesis and the classical assumption that bottom-up saliency is a center-surround process. This classical assumption is formulated as a classification problem. At each image location l, it includes definitions :

- Stimuli of interest : observations within a neighborhood W_l^1 of l (referred to as the center)
- Null hypothesis : observations within a surrounding window W_l^0 (referred to as the surround)

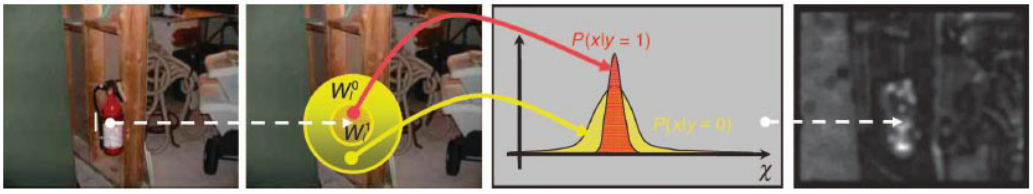

FIGURE 2.7 – Illustration of discriminant center-surround saliency [69]

The saliency of location l, S(l), is equal to the discriminant power of X for the classification of the observed feature vector $x(j), \forall j \in W_l = W_l^0 \cup W_l^1$, into center and surround. $X(l) = (X_l(l), ..., X_d(l))$ is a random process of dimension d, drawn conditionally on the state of a hidden variable Y(l). S(l) is quantified by mutual information between features, X, and class label, Y :

$$S(l) = I_l(X;Y) = \sum_c \int P_{X(l),Y(l)}(x,c) \log \frac{P_{X(l),Y(l)}(x,c)}{P_{X(l)}(x)P_{X(l)}(c)} dx \qquad (2.11)$$

By using the equation above to estimate the mutual information at each image location after the stage of extracting visual features, we can achieve our purpose : the discriminant saliency detection in static imagery. The performance of this model is good.

The model of Bruce and Tsotsos At 2009, Bruce et al. [70] put forth a model of saliency computation within the visual cortex based on the premise that localized saliency computation serves to maximize information sampled from one's environment. The framework of the proposed computational model is depicted in the figure 2.8 :

In this model, the first operation for the input image is the independent features extraction. For each location i,j in the image, the response of various learned filters with properties reminiscent of V1 cortical cells are computed. This operation may be thought of as measuring the response of various cortical cells coding for content at each individual spatial location. This yields a group of coefficients for each local neighborhood of the scene, $C_(i,j)$.

The second stage is density estimation. In this stage, the content of each local neighborhood $C_(i,j,k)$ of the image is characterized by several coefficients a_k. These coefficients are corresponding to the various basis filters

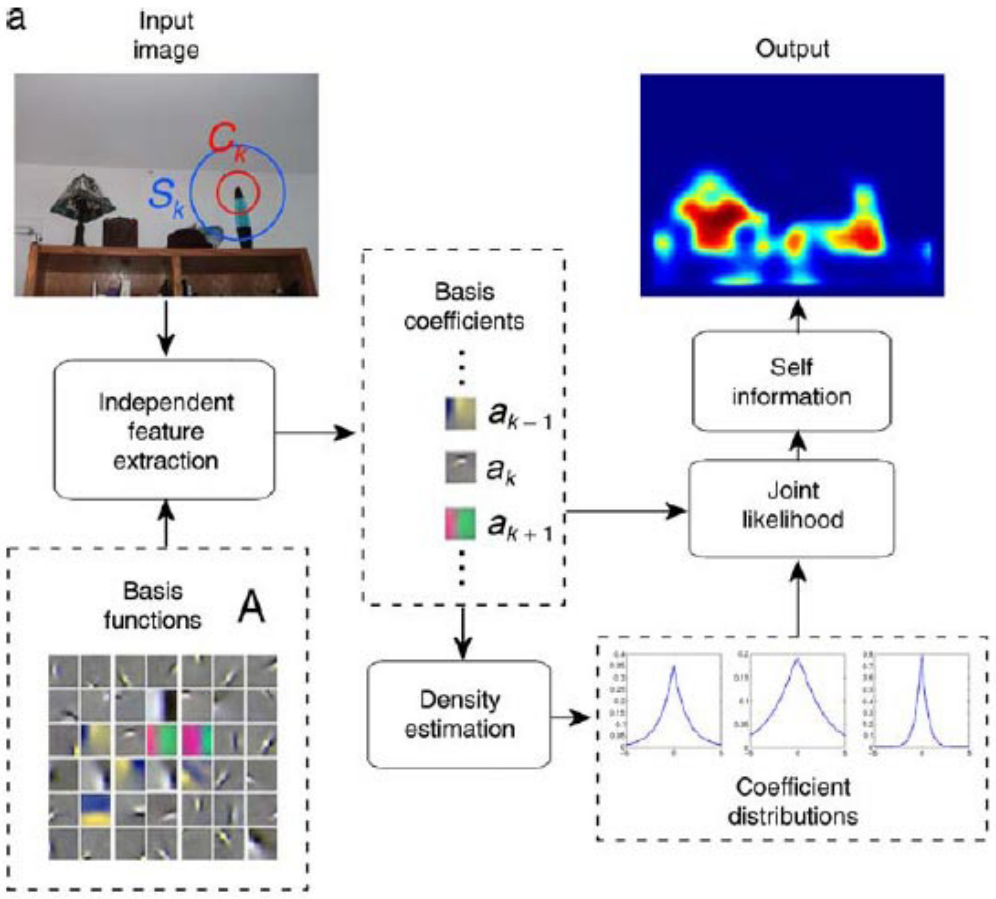

FIGURE 2.8 – The diagram of Bruce and Tsotsos's model [70]

coding for that location, and i,j are corresponding to the position of that local neighborhood. At one spatial location and in the surrounding regions of that location, there are a set of coefficients for a same filter type. Based on a non-parametric or histogram density estimate, the coefficients in the surround form a distribution that may be used to predict the likelihood of the coefficients of $C_(i, j)$. For computational parsimony, this paper supposes that each pixel in the image contributes equally to the density estimate and it is performed based on a 1000 bin histogram density estimate.

The following stages are Joint likelihood and Self-information. Thanks to the operations before, a likelihood estimate corresponding to a single filter type can be afforded by a density estimate for any single coefficient based on coefficients corresponding to the same filter type from the surround. Based

on the independence assumption, the product of the likelihoods associated with each individual filter yields an overall likelihood for all coefficients corresponding to a single location. The Shannon Self-Information of this overall likelihood p(x) is given by $-\log(p(x))$ and the resulting information map depicts the saliency.

This model for visual saliency computation builds on a first principles information theoretic formulation noted as Attention based on Information Maximization (AIM). Although this model is built entirely on computational constraints, the result exhibits considerable agreement with the organization of the human visual system.

The model of Oliva and Torralba Oliva and Torralba propose a model of attention guidance based on global scene configuration[71]. It is based on the assumption that, by directing their attention to relevant regions in the image, human observers will use visual context information to facilitate the search while looking for a specific object in a complex scene. The goal of this paper is to try to locate probable locations of people in scenes, and then obtain a saliency map from the result.

The saliency map is computed using a hardwire scheme (shown in figure 2.9) : Processing the local image features by center-surround inhibition and then using a winner take all strategy to select the most salient regions. The most commonly used image features are the outputs of multiscale oriented band-pass filters. In this paper, each color subband is decomposed by using a steerable pyramid with 4 scales and 4 orientations. Each location has a features vector with 48 dimensions. On the other hand, the saliency is defined in terms of the likelihood of finding a set of local features in the image. To define the saliency, this paper uses a probabilistic definition : the saliency of a location is large when it is more unexpected to find the image features at that location. This probability is approximated by fitting a Gaussian to the distribution of local features in the image.

There are three important processes in this model :

- Contextual modulation of saliency. In the process of contextual modulation of saliency, it formulates the object detection as the evaluation of a probability function. It is the probability of the presence of the object given a set of local measurements. After that, the probability is decomposed into three factors by using Bayes rule : the object likelihood, the local saliency and the contextual priors. The terms that do not require

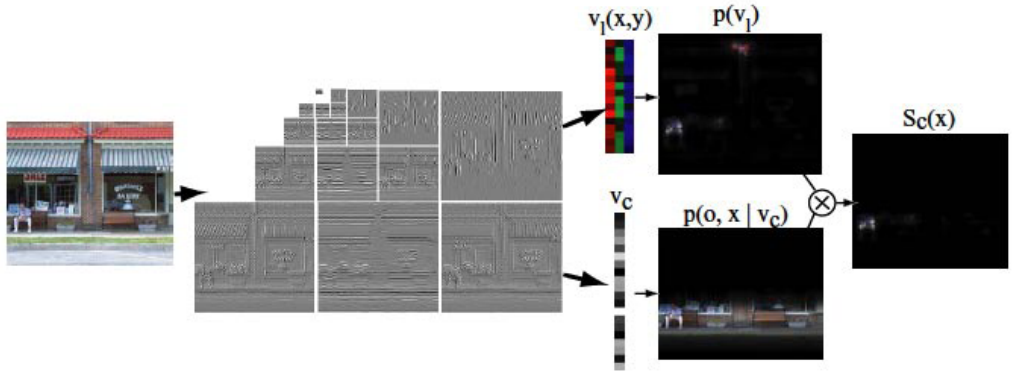

FIGURE 2.9 – Attentional system for object detection integrating local saliency and contextual priors about target location. [71]

knowledge of the appearance of the target are attached importance to.

- Computing contextual image features. The structure of the whole image is describe d by contextual features. Here, the model reduces the dimensionality of the local features to represent the context.
- Learning the location of people. A database of images for which the location of the people is known is used for training the PDF. In this paper, the learning is performed using the EM algorithm and the PDF is then modeled by using a mixture of Gaussians.

In summary, this computational model is concerned with the visual context in which objects embedded. In predicting the locations at which peoples will fix, this computational model does not differ with the performance of other more complex models.

2.3 3D perceptual modelling

In this section, a state-of-the-art of existing models of visual attention integrating depth cue is presented. Below is a resuming table of models.

EM	Input type	Depth dimension involved	Operations	Biological/ empirical model
Ouerhani et al.	Depth map	Stereoscopy (binocular and vertical disparity, da Vinci stereopsis)	Depth as identical input to L.Itti's model	biological for color/luminance processing, not for the depth
Zhang et al.	Depth map and motion (t+k ;-w<k<w)	Stereoscopy and dynamic (motion)	Itti's model Op. + block based optical flow for temporal attention model + depth as disparity	empirical for depth processing and integration
Park et al.	Stereo pair and motion(t-1)	Stereoscopy and dynamic	simple SAD to have disparity, Double Opponent receptive field for Motion	Biologically inspired for motion (DORF in MT Clark2004) and texture(Itti's Model) but arbitrary integration with weights on each cue
Maki et al.	Depth map and motion(t-1)	Stereoscopy and dynamic	stereo disparity calculation(stereo algorithm), motion pursuing algo, depth-based cue integration by a pursuit/saccade mode	biological for the integration only
Fernandez et al.	Stereo pair + shape and motion parameters	Stereoscopy and dynamic(velocity and acceleration)	Dynamic stereo selective attention : feature extraction/integration. Charge disparity analysis and depth-based segmentation	biological stereo attention focus and 2D charge map, inspired by biological visual hemifields

TABLE 2.1 – Existing models of visual attention integrating depth

A relative low amount of article exists in the literature, considering the depth as a cue that can provide essential information to model the behavior of the Human Visual system in attentional tasks. Existing methods differ in the location, and the way the depth is processed, if it's based on biological assumption or not, and if they also implement motion as an additional cue.

2.3.1 Depth and Motion extension to Saliency models

Itti based saliency models

Ouerhani et al [72] proposed to integrate the depth in an extension of the Itti's saliency model, as a conspicuity map. To the n conspicuity map in competition in the integration process, m additional maps coming from the depth are proposed, but only the depth as binocular disparity information is kept : n+1 maps are combined. Under empirical geometrical assumption, *mean curvature* and *depth gradient* are considered but finally not used. Mean curvature is a surface feature providing information about the geometry of

scene. It is underlined that preprocessing operations like smoothing could overcome the inconvenient of the mean curvature noise sensitivity. Adding such preprocessing operations to a computational depth cue could lead to high inaccuracy and and remains highly content-dependent. (parameters, size of smoothing operators). Depth gradient is presented as a feature vector efficient to underline depth changes in angles, but also in depth discontinuity. Finally, experiments with color + depth as conspicuity maps shows "subjective" improvements in the resulting saliency maps.

$$S = \sum_{i=1}^{n+m} w_i C_i$$

with m=1, the depth map. $w = (M - \overline{m})^2$, where M is the maximum activity of the conspicuity map and m is the average of all its local maxima. w measures how the most active locations differ from the average. Thus, depth locations which stand out from their surrounding are promote, through a peak response in depth feature which compete with color and is finally propagate to final saliency maps. These experimental results show the effectiveness of channel competition that might effectively takes place in the different depth processing areas of the HVS.

Zhang et al. The stereoscopic visual attention model of [73] relies on the same input characteristic for depth information but not on the same integration. Indeed the input disparity is taken as in [72] but is converted into "perceive depth" map D to give more importance to close objects, as an hypothesis. Still in difference with previous model, the depth is integrated with other feature maps through a weighted coefficient k_D , that seems to be manually setted and not obtained by square difference between maxima and all local maxima.

Another additional feature is integrated into the model : the motion, as a major stimuli of attention. A block based optical flow algorithm is used to estimate motion between consecutive frames. To deal with camera motion, motion maps of different resolution are decomposed with Gaussian pyramids of 9 levels, and center-surround difference is used to separate the "attentive object" motion from background motion. This results in the motion saliency map S_m.

The correlation between u and v channels is also empirically used and weighted by minimum of either static and motion, static and depth, or motion and depth feature maps. Finally, the Zhang stereoscopic visual attention

model is defined as :

$$S_{SVA} = \Psi \left[D \cdot \left(k_s S_s + k_m S_m + k_D D - \sum_{uv \in \{sm, sD, mD\}} e_{uv} C_{uv} \right) \right]$$

The features maps after integration are multiplied again by depth map to give more importance to the closer pixel value. The resulting images show pertinent potential attention areas, but suffers from an empirical integration of features and especially of the depth (converted in "perceived depth" under empirical assumption).

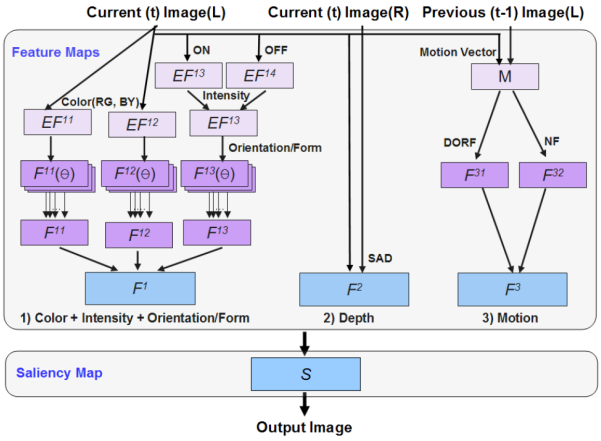

FIGURE 2.10 – Schematic diagram of the Park's model (from [74])

Park et al. Similarly, [74] presented an "automatic focusing attention model "for a stereo pair of image sequences", integrating depth and motion. The model estimates the depth from two left and right input views with a standard SAD (Sum of Absolute Difference) stereo matching algorithm.

Based on psychological studies showing MT (middle temporal cortex) that cells respond best to movement in its selective direction within its receptive field, biological functions of "Double Opponent Receptive Field"(figure 2.11) and "Noise filtration"(figure 2.12) are implemented to calculate the motion map. The response of a pixel is computed by :

The double opponent receptive fields responds to a visual stimulus when

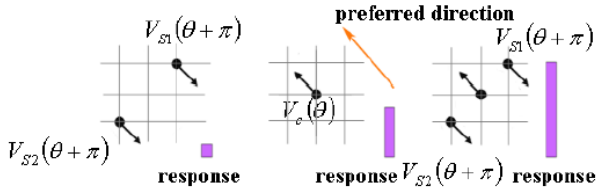

FIGURE 2.11 – Park's model based on "Double Opponent Receptive Field"(from [74])

a pair of surround motion directions is opposed, it is computed by

$$If\,|V_C(\theta)| \neq 0,$$
$$|R(\theta)| = \left|-\frac{1}{2}V_{S1}(\theta+\pi) + V_C(\theta) - \frac{1}{2}V_{S2}(\theta+\pi)\right|$$
$$else\,|R(\theta)| = \left|-\frac{1}{4}\{V_{S1}(\theta+\pi) + V_{S2}(\theta+\pi)\}\right|$$

where $V_C(\theta)$ presents the motion vector of an observation point, $V_{S1}(\theta+\pi)$ and $V_{S2}(\theta+\pi)$ the surrounding motion vector.

The noise filtration function considered the case of a pair of cells selectively tuned to opposite directions. The MT cell's response is then very weak in one of both direction.

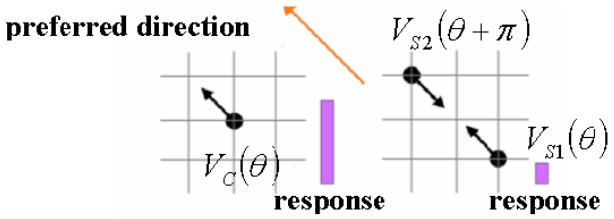

FIGURE 2.12 – Park's model based on "Noise filtration" (from [74])

$$If\,|V_C(\theta)| \neq 0, and\,|V_{S1}(\theta) - V_{S2}(\theta+\pi)| \geq 0$$
$$|R(\theta)| = |V_{S1}(\theta) - V_{S2}(\theta+\pi)|$$
$$else\,|R(\theta)| = 0$$

These two oriented filters are biological-inspired functions that might be processed in MT brain area to obtain a kind of motion saliency map. Unfortunately concerning the integration step, arbitrary weights are given to each of these resulting features (5 for motion, 3 for color/illumination/orientation

and 1 for depth), citing general studies showing that motion is more important than color, which is itself more important than orientation, form, or depth. So the final saliency map is computed in a heuristic way as :

$$S = 3\ S_S + 1\ D + 5\ S_M$$

Experiments show an improvement in the detection of supposed salient regions, with comparison to their previous spatiotemporal model without depth. A comparison to other depth spatiotemporal models is not given.

Saccade/pursuit model

Maki et al.[75] proposed a computational model of preattentive cues processing and latter depth-based integration, where the idea is to maintain the attention to the closest moving object under biologically inspired mechanisms. Three parallel stages of stereo disparity, image flow and motion detection calculation are realized first. The cue integration guides the attention to the salient part, by combination of two independent pursuit and saccade modes. As shown in both figures below, the processes of pursuit (figure 2.13) and saccade (figure 2.14) is based on iterative disparity/flow histogram computation, logic operation and generation of iterative target masks. The

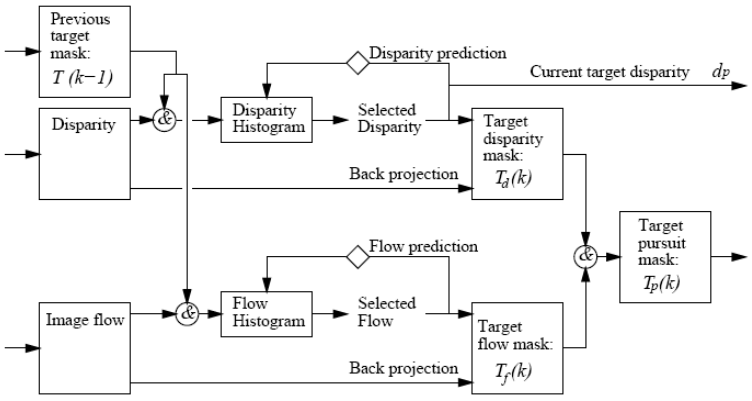

FIGURE 2.13 – Schematic flow diagram of the attentional pursuit. Diamonds indicate a one frame delay in the feedback, circles a logic AND operation (from [75])

principle of this last mode is that an "interesting part" or distractor calculated only from motion relative to the background trigger an attentional shift. This newly detected moving target is carried in disparity histogram and finally in the target saccade mask. Under arbitrary depth-based criterion (the closer

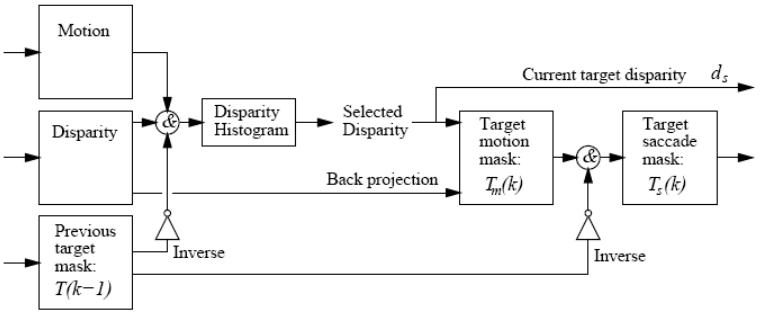

FIGURE 2.14 – Schematic flow diagram of the attentional saccade mode (from [75])

a target is, the higher priority it has), either pursuit mask or saccade mask are selected to produce the final target mask. Results show approximative target mask, especially on the border of close objects to segment, but model benefits of simple iterative algorithms.

Charge map based model

The "dynamic stereoscopic selective visual attention" (DSSVA) model from Fernandez et al.[76] is a stereoscopic extension of their previous model. It is based on two biologically inspired methods : accumulative computation and lateral inhibition.

The first**Dynamic Selective Visual Attention** stages (DSVA), illustrated on figure 2.15 consists of different subtasks :

- an "Attention Building" is modelized with gray level images segmented into a lower number of levels.
- then "feature extraction" and their "integration" output an Interest map, and stores for each image pixel the result of the comparison with three-discrepancy classes : active/inhibited/neutral, as a result of evaluating motion detection between two consecutive time instants and the observers's guidelines.
- the "Working Memory" is obtained for each gray-level band in subtask "Attention Building". In the same logic each pixel [x, y] can be active, inhibited or neutral.
- The value of each pixel in WM is the max value of WMi at each gray-level band.
- Lastly, monocular attention focus is obtained on a reinforcement basis : V_{active} pixels in WM reinforce attention in the MAF, $V_{inhibited}$ ones decrement the attention value.(eq 5) where D(MAF) and C(MAF) are

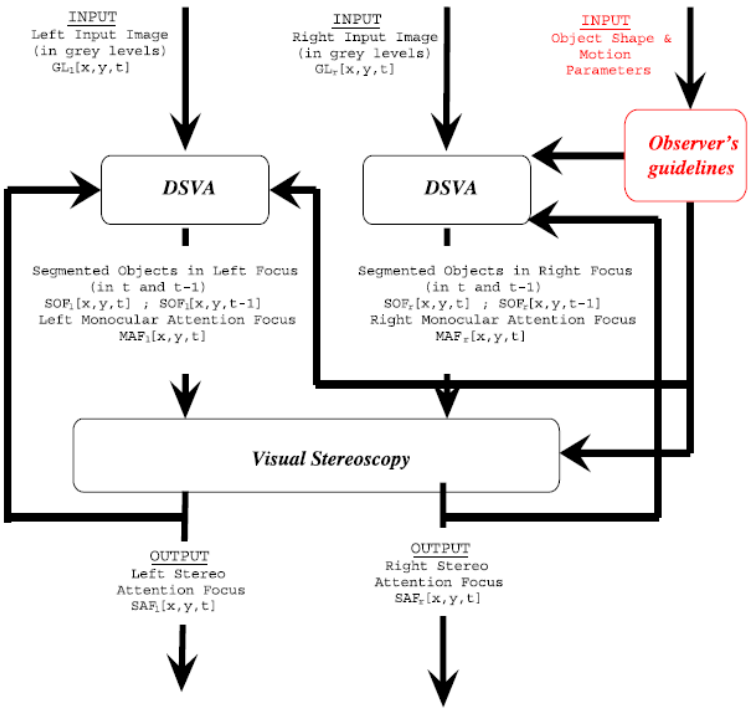

FIGURE 2.15 – Schematic flow diagram of the attentional saccade mode.

the discharge and charge constants.

– Segmented objects in focus are obtained through simple filtering :

$$SOF_{l/r}[x,y] = \begin{cases} GLB_{l/r}[x,y,t] & \text{if } MAF_{l/r}[x,y,t] > Ch_{min} \\ 0 & \text{else,} \end{cases}$$

Secondly, the Visual Stereoscopy is obtained through 4 subtasks :

– Obtention of the 2D charge map inspired by the biological visual hemifields. Left and right motion charge memories are obtained by means of AC PSM to provide info about movement.
– Charge disparity analysis : a 3D depth map or disparity map, shows the depth of points in the scene where there has been movements. The decision stands on the disparity with the greatest reliability, based on epipolar, ordering and disparity restrictions.
– Obtention of Depth : left and right stereo working memories are calculated as in DSVA (Attention Building)
– 3D Attention Reinforcement : the couple of stereo attention foci is obtained as for monocular attention foci. (Vactive vs Vinhibited)

The tests are done on synthetic stereo images sequences in low resolution (320*240) with 2 configurations : one based on size and shape only, another

on depth only. Even if the condition of experimentation are quite restricted, it gives a global idea of the capacity of the method. A clear segmentation is realized on object, but fails with turning motion of objects.

Conclusion

The different models presented here introduce the use of depth and/or motion as important, contributing and possible improving cues to classical 2D models. They introduced some considerations of motion (with DORF [74]) and depth or stereoscopy (charge disparity analysis in[76]), but fails to attribute quantitative motion and depth based on objective psychophysical facts : arbitrary weighting are used instead.

2.4 Conclusion

This section presented the previous studies regarding perceptual modelling in 2D and 3D and the visual attention modelling in 2D. These models are conceived in order to optimize communication systems, for the assessment of image quality, for coding applications, etc.

Chapitre 3

Applications

Perceptual models can be incorporated in communication systems in order to improve their performances and consequently the experience of the user. This chapter focuses on the possible applications of perceptual models. The first section introduces perception-based quality metrics, then the second section addresses the coding field. Finally section 4 addresses super-resolution and section 5 introduces motion sharpening.

3.1 Quality metrics

During acquisition, compression, transmission, and storage, digital images suffer a wide variety of distortions that may cause visual degradations of the image quality. So image quality needs to be assessed. There are two types of quality assessments : subjective quality assessment, which refers to viewers opinions about quality of an image over a display system, and objective quality assessment, which is comes from objective quality metrics.

A visual quality metric is a tool that can optimize the performance of digital imaging systems by taking into account the human perception of visual information and predicting automatically subjective ratings. Including a visual quality metric within a compression scheme can reduce the visibility of artefacts and increase the subjective quality, i.e. the quality perceived by the viewer. Perceived video quality depends on many factors (viewing distance, display size, resolution, contrast, brightness, sharpness, naturalness, etc.)

Feature-based quality metrics try to assess presence of visual artefacts by evaluating features in images or video streams. Generally they are built according to the same scenario : when the typical artefacts are identified, a database containing those artefacts is created and submitted to subjective

quality experiments. The subjective assessments for each artefact are combined and balanced to create the prediction function that evaluates the impact of the artefacts. The created prediction function should then be close to subjective ratings. Consequently, the validation of an objective quality metric depends on its closeness to subjective ratings.

Although feature-based quality metrics are efficient, especially when a limited number of artefacts are expected, they require a consistent database, and a sufficient number of observers for the subjective quality tests that allow their validation.

We can distinguish three kinds of metrics : the full-reference (FR) metrics, which need the entire reference data to process the comparison ; the no-reference (NR) metrics, which do not need any reference and are generally based on blockiness estimation ; the reduced-reference (RR) metrics, which extracts only some features of the reference for the comparison. In the following subsection, metrics used for image and video sequences assessments will be presented.

3.1.1 Overview of visual quality metrics

FR metrics

A widely used metrics is Peak Signal to Noise Ratio (**PSNR**), because of its simplicity and mathematically easiness to deal with for optimization purposes. Based on the logarithm of the inverse of Mean Square Error (MSE), it is often considered as not suitable for perceptual assessments. This is due to the fact that it does not take into account the temporal and spatial dependencies between samples, it is a pixel-based metric. It is known that two different images can have the same PSNR while being widely perceptually different [77]. The MSE between two pictures I and $\tilde{I}$ is defined as follows :

$$MSE = \frac{1}{XY} \sum_l \sum_c [I(l,c) - \tilde{I}(l,c)]^2$$

where XxY is the size of one image, $I(l,c)$ is the value of one pixel in I. The PSNR in decibels is defined as :

$$PSNR = 10log_{10}(\frac{m^2}{MSE})$$

where m is the maximum value that a pixel can take (255 for 8-bit images).

In order to overcome these limitations, the Structural SIMilarity (**SSIM**) was introduced. Based on the fact that human vision extracts and identifies objects from a scene, SSIM metrics evaluates luminance, contrast and structural comparisons. The SSIM between two signals x and y is defined as follows :

$$SSIM(x,y) = \frac{(2\mu_x\mu_y+C_1)(2\sigma_{xy}+C_2)}{(\mu_x^2+\mu_y^2+C_1)(\sigma_x^2+\sigma_y^2+C_2)}$$

where, if the two signals x and y contain N samples, the statistical features are :

- $\mu_x = \bar{x} = \frac{1}{N}\sum_{i=1}^{N} x_i$
- $\mu_y = \bar{y} = \frac{1}{N}\sum_{i=1}^{N} y_i$
- $\sigma_x^2 = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})^2$
- $\sigma_y^2 = \frac{1}{N-1}\sum_{i=1}^{N}(y_i - \bar{y})^2$
- $\sigma_{xy} = \frac{1}{N-1}\sum_{i=1}^{N}(x_i - \bar{x})(y_i - \bar{y})$
- and the constants : $C_1 = (K_1L)^2$, $C_2 = (K_2L)^2$. L is the dynamic range of the pixel values (for 8-bit images, L=255), $K_1 = 0.01$, and $K_2 = 0.03$

SSIM system can be seen on figure 3.1.

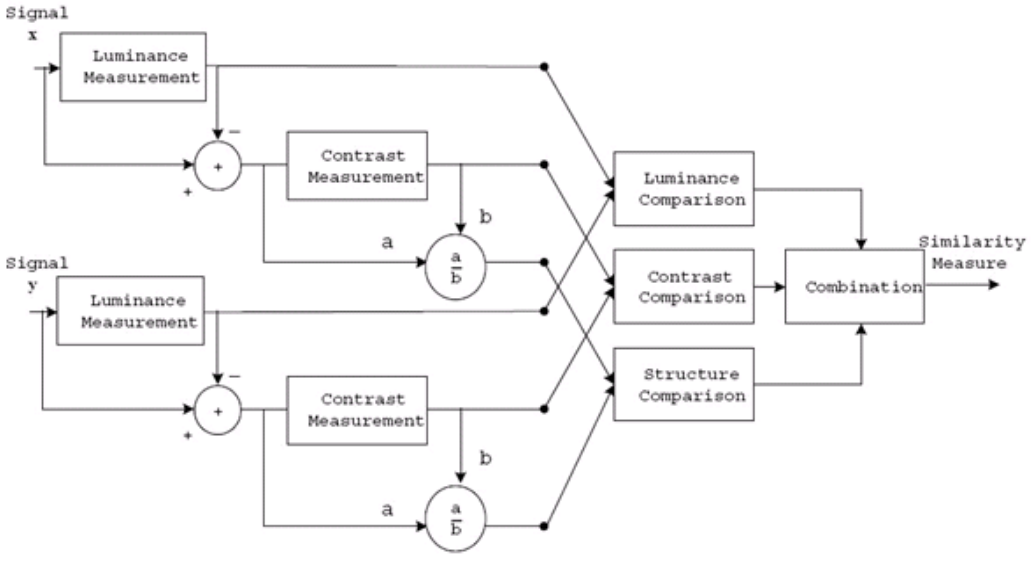

FIGURE 3.1 – SSIM system

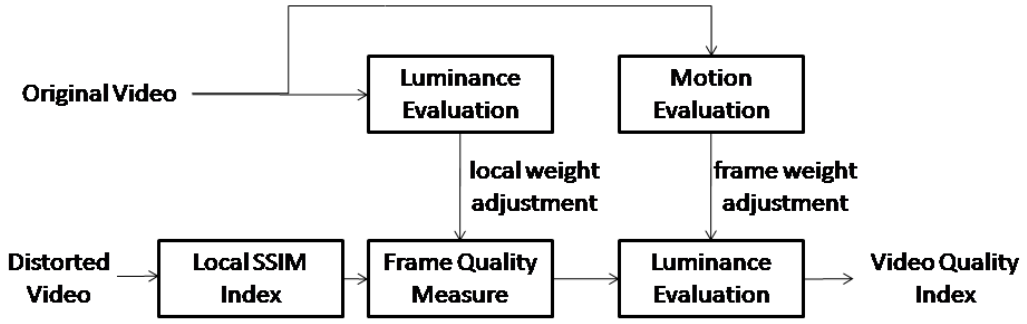

FIGURE 3.2 – VSSIM system [78]

As an extension for video data, **VSSIM** was proposed [78]. It is not an average of SSIM of both frames, but rather made of adjustments for weighted SSIM (figure 3.2 : for example dark regions that do not attract fixations are assigned smaller weighting values). This metric is applied by moving a window on the imageand calculating an SSIM map.

VSSIM measures the distorted video in three levels. The first level is the local region level, where the measurement is applied on 8×8 windows of the Y, Cr and Cb color planes. Let note $SSIM_{ij}^{Y}$, $SSIM_{ij}^{Cb}$ and $SSIM_{ij}^{Cr}$ the $SSIM$ index values of the Y, Cb and Cr components of the j-th sampling window in the i-th video frame, the local quality index is defined as :

$$SSIM_{ij} = W_Y SSIM_{ij}^{Y} + W_{Cb} SSIM_{ij}^{Cb} + W_{Cr} SSIM_{ij}^{Cr}$$

where W_Y, W_{Cb} and W_{Cr} are the weights.

The second level is the frame level, where the local quality values are combined into a frame-level quality index using :

$$Q_i = \frac{\sum_{j=1}^{R_s} w_{ij} SSIM_{ij}}{\sum_{j=1}^{R_s} w_{ij}}$$

where Q_i is the quality index measure of the i-th frame in the video sequence, and w_{ij} is the weighting value given to the j-th sampling window in the i-th frame. The third level is the sequence level that averages all the frames level values to produce the overall quality of the video sequence by :

$$Q = \frac{\sum_{i=1}^{F} W_i Q_i}{\sum_{i=1}^{F} W_i}$$

where F is the number of frames in the video sequence and W_i is the weighting value assigned to the i-th frame. Weighting values are assigned according to the local luminance, as explained in [78] A comparison [79] of objective metrics for video states that VSSIM is a computationally simple and consistent method with human ratings, but does not address distortions such as vertical distortions being more noticeable than horizontal distortions.

An interesting approach consists of determining whether the distortions are below a threshold of visual detection. Visual Signal to Noise ratio (**VSNR**) is based on [80]. If the distortions are suprathreshold, multiscale wavelet decomposition is used and a Euclidean distance is performed in the distortion-contrast space. Let $E = \tilde{I} - I$ be the distortions contained in the distorted image $\tilde{I}$, I be the original image, the VSNR is defined as :

$$VSNR = 10log_{10}(\frac{C^2(I)}{VD^2})$$

where $C(I)$ denotes the Root Mean Square (RMS) of the original image I, and VD is the visual distortion defined as :

$$VD = \alpha d_{pc} + (1 - \alpha)\frac{d_{gp}}{\sqrt{2}}$$

the parameter $\alpha \in [0, 1]$, d_{gp} is a measure of the extent to which global precedence has been disrupted, d_{pc} denotes a measure of the perceived contrast of the distortions.

Results show that it is generally competitive with current metrics of visual fidelity and has low computational complexity and low memory requirements.

Also, the **Sarnoff JND vision model** should be mentioned [81]. It predicts the perceptual ratings that human subjects would assign to a degraded image. It is based on differences between degraded image and original image. Those differences are quantified in units of the modelled human just-noticeable difference (JND). Figure 3.3 illustrates the architecture : the images of the pyramids are filtered with a Gaussian filter, then they normalized. After calculating three contrast measures, a contrast energy mask is processed. The algorithm ends with a pooling stage.

The DCT-based metric, called Digital Video Quality (**DVQ**) metric, has performances similar to Sarnoff JND model. The metric is described in [82] :

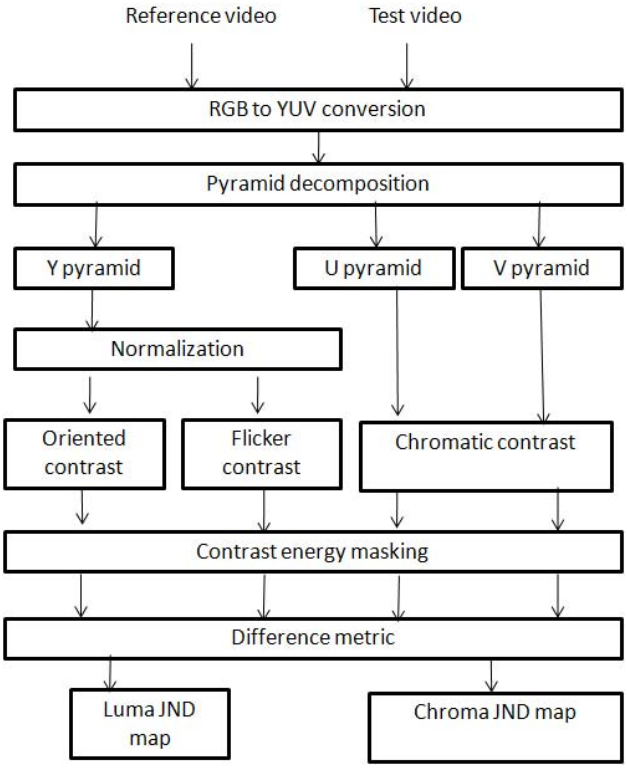

FIGURE 3.3 – JND Model system

it contains spatial and temporal filtering stages, a contrast masking stage and a probability summation stage.

In [83], the proposed metric, Visual Information Fidelity (**VIF**), seems to outperform Sarnoff's JND and SSIM, for the run tests. This metrics is derived from a statistical model for natural scenes, a model for image distortions, and a human visual system model.

RR metrics

Video Quality Metric (**VQM**)[84] is a standardized method of objectively measuring video quality that closely predicts the subjective quality ratings. It is based on the measure of the amount and orientation of activity in spatio-temporal blocks from the sequence.

C4 [86] is a metric based on the comparison between structural informa-

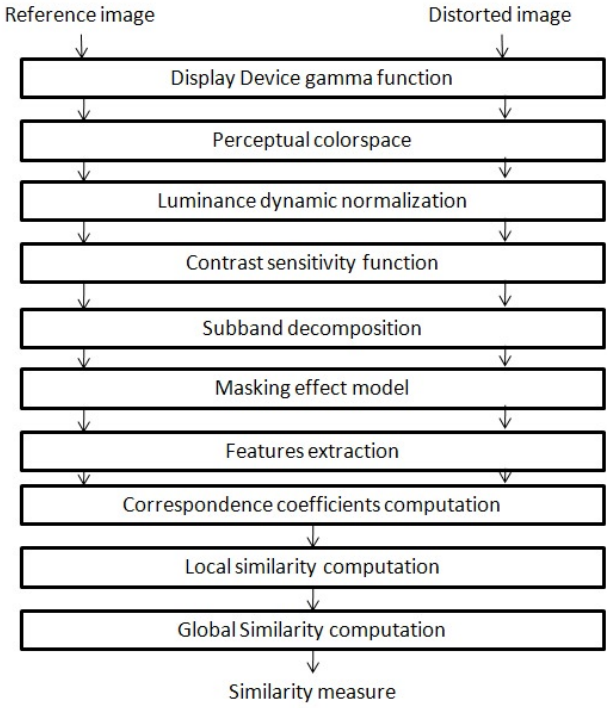

FIGURE 3.4 – C4 architecture [85]

tion extracted from the distorted and the original images (see figure 3.4. It relies on a model of the human visual system. In two steps, the perceptual representation is first built for the original and the distorted image ; then representations are compared to compute the quality score. However it is a time-consuming method because it integrates a global contrast sensitivity function.

3.1.2 Towards more 3D-adequate quality metrics

Perceived 3D image quality

Before evaluating image quality, it is essential to understand and know the typical 3D-imaging distortions. Many distortions have been described in the literature [87].

The ***keystone effect*** makes the image look like a trapezoid. It is due to a converging camera configuration where the left and right cameras are po-

sitioned at an angle toward each other. It is more noticeable with increasing camera base distance decreasing convergence distance, and decreasing lens focal length and causes minimal eye strain.

The ***ghosting effect*** is also known as ***crosstalk***. It is perceived as ghost, shadow, or double contours due to imperfect image separation. But it can be due to the display, which makes difficult the task to elaborate a specific algorithm. Crosstalk is suspected to be the main contributor to visual comfort and image quality.

When depth is perceived as unnatural, that is to say the scene appears to be divided into discrete depth planes, the distortion is called ***cardboard effect***. This is due to either image acquisition parameters, or compression parameters resulting in a coarse quantization of disparity or depth values. Along the same sequence, it can occur that an object or part of an object is assigned to different depth layers in time, which results in a ***flickering depth percept***. Studies [88] have shown that in this case, a perceptual gain can be obtained by reducing the depth resolution.

On multiview autostereoscopic display, the ***picket-fence effect*** is the distortion that makes appear vertical banding in an image due to the black mask between columns of pixels in the liquid crystal display (LCD).

In 3D video quality, we face the problem of ***binocular suppression*** [89]. This phenomenon is due to artefacts that cause contradictory depth cues to be sent to each eye. Similarly to asymmetric video encoding (for instance in stereoscopic encoding, one view can be encoded with higher quality than the other) which results in the masking of the artefacts of the worse view, the risk is to suppress the stereopsis because there is no combination of both values.

Even though, it has been shown that image quality is important for visual comfort, it is not the only factor of great 3D visual experience. New concepts have to be considered such as presence (the feeling of *being there and react*, investigated in [90], [91] and [92]), naturalness (i.e. perceptual realism, or what observers perceive as a truthful representation of reality), and viewing experience. Those concepts are widely studied in [93].

Objective assessments

Because very complex concepts are involved in 3D visual experience, the simple idea to apply 2D quality metrics to 3D video is hardly conceivable. In fact, whereas 3D video aim is to provide a depth feeling, most used metrics in the literature do not incorporate perceptual factors related to reproduced depth, 3D image impairments, and visual comfort. In [93], the study investigates whether 2D image quality models are sufficiently adequate to measure 3D quality, and which criterion is consistent for 3D video quality assessment. Seven experiments show that image quality evaluation does not reflect the added value of depth (which has been neglected so far). A model of the added value of 3D is proposed and incorporates image quality and perceived depth : with a linear regression analysis, viewing experience and naturalness were predicted and convincing fits were found.

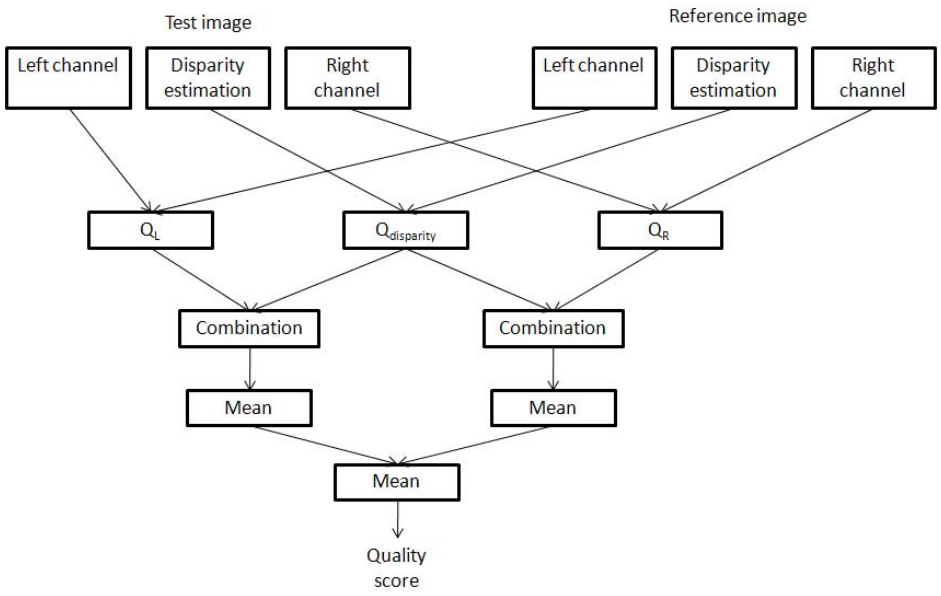

FIGURE 3.5 – Proposed metric in [94]

In [94], two 3D quality metrics are proposed and rely on both the use of 2D metrics and depth information. SSIM or C4 are combined to a disparity distortion measure. Figure 3.5 shows one of the proposed systems : Q designates the quality measure processed with SSIM. The quality measures from 2D quality assessment performed to left and right images of a stereopair (each pixel of the images has a SSIM measures, stored is an "M_{map}") is fused

with the disparity distortion (measured locally). The final score is the mean score of left and right distortions measures. The combination is calculated using :

$$Ddl(p) = M_{map}(p)(1 - \frac{\sqrt{Disp.Or(p)^2 - Disp.Dg(p)^2}}{255})$$

where p is the index of the current pixel, M_{map} is the SSIM map obtained from the reference and the distorted images (left or right images), $Disp.Or$ is the disparity of the reference stereopair and $Disp.Dg$ is the disparity of the degraded stereopair. Results show that combining the disparity distortion measure with SSIM metric enhances performances, and the results are close to perceptual objective metrics such as C4.

In [95], the proposed FR metric is composed of two components : stereoscopic quality (Q_s) and monoscopic quality (Q_m), as illustrated in figure3.6. This FR metric compares the initial stereo-frames, designated as L_ref and R_ref, with the distorted stereo-frames, denoted as L_test and R_test. The proposed metric should assess the amount of binocular cues preserved by comparing the perceptual disparity maps. Perceptual disparity maps are generated by applying a "block matching" based on a perceptual measure (SSIM in the experiments). A stereo-similarity map is also generated from the SSIM measures of the optimum disparity vectors found during the "block matching" process. Then, they multiply the two maps element-wise, and obtain the stereoscopic quality. This is done for various scales of the stereopairs and averaged to obtain the final Q_s. Cyclopean images are made from the reference pair and the distorted pair, as the human eye would see it as a monoscopic image. Then, the two images are perceptually compared (using SSIM). This process is done at different scales, and the final Q_m results in the mean value. The relative importance of Q_m and Q_s for the overall stereoscopic quality is to be studied as future work, since it is different from observer to observer, depending on how much the observerï£¡s visual system relies on binocular cues for depth perception. The metric is tested on distorted stereo images and stereo video sequences. The aim is to build a perceptually-aware feedback for a H.264 based stereo video encoder. The results show that Q_s and Q_m follows the subjective opinion (MOS) better than SNR values.

In [96], the proposed metric can evaluate depth image based rendering for video plus depth representation. It is based on Color and Sharpness of Edge Distortion (CSED). Color distortion evaluates the luminance loss of

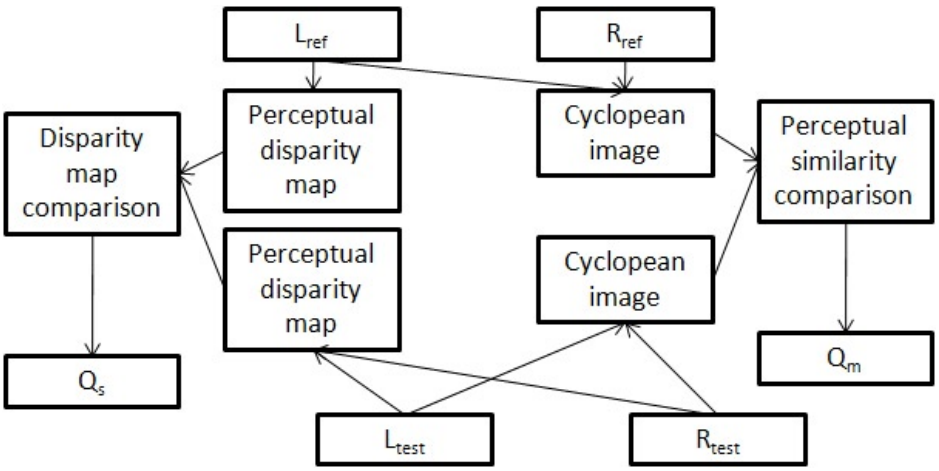

FIGURE 3.6 – Proposed perceptual metric in [95]

the rendered image and the latter measures a depth-weighted proportion of remaining edge to the original edge. So the originality of this metric is that it also gives information about synthesis error. The metric follow the subjective ratings. Color distorsion is measured according to :

$$S_1(x,y) = \frac{2\mu_x\mu_y+C}{\mu_x^2+\mu_y^2+C}$$

where C is a constant to take effect only when $\mu_x^2+\mu_y^2$ is small, and μ_x and μ_y are the mean luminance of a 5×5 window centered by x and y, respectively the disoccluded pixel in rendering image and of the corresponding pixel in reference.

Sharpness is measured using :

$$S_2(x,y) = \frac{\sum_{i\in\delta_x \cap \delta_y} D_i}{\sum_{j\in\delta_y} D_j}$$

where δ_x, δ_y are the edge pixels along the boundary position of the hole region in x and y respectively, and D is the intensity of corresponding pixel in depth image.

In [38], an imaging process channel for feature-based 3D quality estimation is proposed. It takes into account the binocular mechanisms. It is a FR metric that first transform each channel into a perceptual color space such

as S-CIELAB ; then the next stage estimates the visibilities of depth cues ; finally, the processed visual information of the test and reference video sequences are compared by a *structural similarity* metric (for CSF perceptual salience estimation).

In [97], a perceptual metric for stereoscopic video, based on the human visual system mechanisms, is proposed. It is based on the most relevant properties of the human visual system, for stereo video, namely : contrast sensitivity function, multi-channel and masking effect, and depth perception. The metric consists of different steps (figure 3.7) :

- perceptual decomposition : after temporal decomposition, the output is subject to spatial decomposition where the left and right views are decomposed into several subsets of high frequencies and disparity information. This spatial decomposition is performed with a 3D wavelet decomposition technology based on disparity compensation view filtering (DCVF) ;
- contrast conversion and masking : the perceptual response to contrast $r_{k,\theta}(x,y)$and the perceptual response to depth r_z are computed. They are expressed as :

$$r_{k,\theta}(x,y) = K_s csf_k \frac{c^2_{k,\theta}(x,y)}{\Delta + \sum_\theta c_{k,\theta}}$$

  where K_s is a gain factor, csf_k is the contrast sensitivity of scale k, $c_{k,\theta}(x,y)$ is the contrast of the subset.
  And the perceptual response to depth is as follows :

$$r_z(x,y) = K_z dsf c_z(x,y) + (1 - K_z) dsf |H^l_t(x,y) + H_v(x,y) - H^r_t(x,y)|$$

  where $c_z(x,y)$ is the depth contrast, dsf is the depth sensitivity function, $H^l_t(x,y)$, $H^r_t(x,y)$ are the high frequency band after temporal filtering, K_z is the weight coefficient and H_v is the high frequency band after filtering.
- pooling and quality mapping : the two kinds of responses are computed for left and right views and the overall distortion e between the reference video and the distorted video sequences, for frame i is :

$$\begin{aligned} e_i = &\textstyle\sum_k A_k (\sum_{\theta,x,y} |r^r_{k,\theta}(x,y) - r^d_{k,\theta}(x,y)|^4)^{\frac{1}{4}} \\ &+ B_z (\textstyle\sum_{x,y} |r^r_z(x,y) - r^d_z(x,y)|^4)^{\frac{1}{4}} \end{aligned}$$

  where A_k and B_z are weight coefficients determined experimentally,

$r^r_{k,\theta}(x,y)$, $r^r_z(x,y)$ and $r^d_{k,\theta}(x,y)$, $r^d_z(x,y)$ are the responses of the reference sequence and the distorted sequence respectively.

The multi-channel vision model proposed is based on 3D wavelet decomposition. Results show that the proposed method is more perceptually consistent with human ratings than currently used pixel-to-pixel based method such as PSNR and MSE.

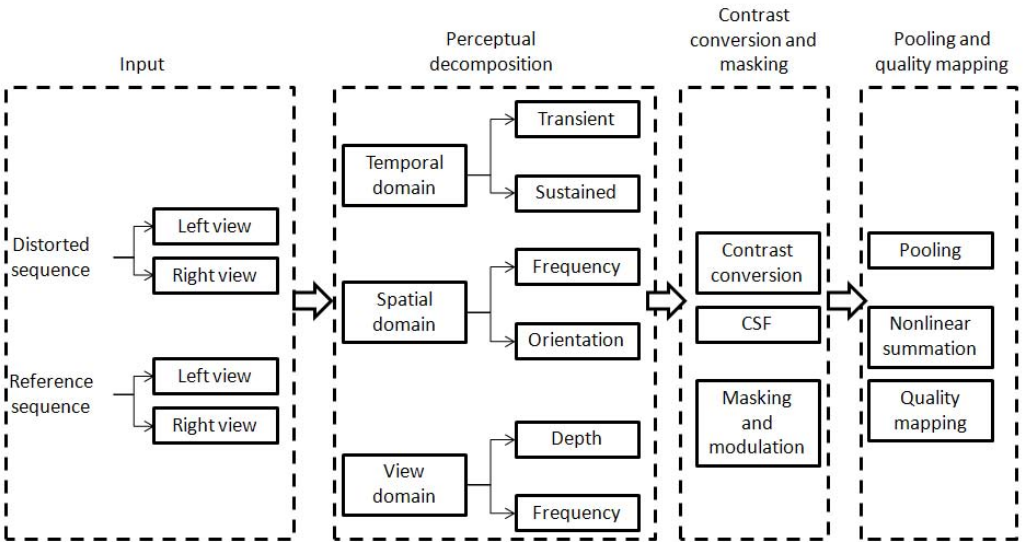

FIGURE 3.7 – Proposed perceptual metric in [97]

The complete mechanisms involved in human visual system are not entirely understood but are dedicated to various studies currently. The new 3D quality metrics that will assess 3D content (images and video sequences) seem to require the knowledge, a modelling tool or the incorporation of such mechanisms. The previously presented studies seem to outperform pixel-based metrics, as they integrate more appropriately perceptual concepts. It is a fairly new research area but the studies already show that 2D inspired metrics are no longer sufficient in the 3D domain because not every dimensions are considered in those priors metrics.

Most of the presented metrics are applied on stereoscopic images or video sequences (only texture sequences are processed) but we may want to apply them on multiview video (MVV) data, or multiview video plus depth (MVD) data. A conventional monoscopic video quality metric (say VSSIM) can be used as a quality measure for video plus depth data by measuring the quality of virtual views that are rendered from the distorted color and depth sequences. The reference view is then the virtual view rendered from undistorted texture and depth maps. This could be extended to MVD. An extension to MVD data can be imagined for [96].

It is worth underlining the fact that since nor disparity neither depth maps are natural images, it is non-sense applying perceptual-based distortion metrics. Evaluating the perceptual quality of a 3D content seems to require considering the global rendering of the content. In a freeviewpoint context, direct intervention, or comparison between two depth or disparity maps is a controversial subject because we do not know exactly the relationship between depth and texture, i.e. we do not know exactly which artefact in the depth map will end up in which artefact in the rendered view. It is then difficult to quantify the quality of a depth map, in a perceptual context. So it may be helpful studying the interactions between depth, texture and rendered views, in the case of the use of multiview video plus depth data.

3.2 Coding

As digital images and video products are designed for human viewing, human visual system (HVS) models are meant to faithfuily reproduce perceptually important infomation and eliminate the infomation that the visual system cannot perceive. Perceptually lossless compression method aimed to achieve this. In the meantime, another problem is the assessment question : because of the lack of standardization in the field of objective quality assessment, and the lack of reliable comparisons of the performance of the different state-of-the-art metrics, it is often difficult to evaluate the codecs performances.

Concerning perceptually lossless methods, the challenge remains either to preserve the best the image quality at a fixed bit rate, or to reduce the rate required for a given quality. In [98], a perceptually based fractal image compression is proposed. It is based on the assumption that all real-world images are rich in affine redundancy. The proposed method uses a human visual (HVS) model that improves the encoding fidelity. The fractal compression is believed to be best suited for textures and natural images, relying on the basic assumption of redundancy. Also, fractal compression can achieve high coding efficiency and good perceptual quality levels. Typical artifacts are blocking artifacts and image blurring.

Wu and al., in [99], used an advanced human vision model to identify and to remove psycho-visual redundancy. The proposed coder is built on the JPEG2000 framework. Instead of embedding the vision model into a post-compression rate-distortion optimization stage, they proposed to embed it into a "visual pruning function". For each frequency level, at each orientation and location, a reference coefficient and a set of distorted coefficients are generated, by applying a progressive bit-plane truncation from the least significant bit. Then, comparing the reference coefficient and each distorted coefficient using their vision model generates perceptual distortion measures, and a set of percentage responses. Finally the set of distortion measure, and the set of percentage reponses are gathered. Coefficient are then truncated to a perceptually optimal bit-plane level according to predetermined thersholds. Results shown higher compression ratio gains comparing to lossless methods, without any visible distortion.

In [100], the author proposed a saliency model that is used to realize a saliency-based compression. The quantization varies according to the measure of salience. Fine quantizaztion is applied on high saliency regions while,

coarser quantization is applied on low saliency regions. The experiments the author presents are JPEG-based. The results show that the saliency-based compressed images are more pleasant to look at but, because of the losses, its SNR scores are worst than JPEG for the same compression ratio.

Video compression quality was improved by using a saliency map estimator in [101]. A ROI-based video compression setup and low bit-rate MPEG-4 video encoding were used. This saliency-based model for visual attention operates both top-down and bottom-up information. It creates a skin conspicuity map and orientation, intensity and color conspicuity maps within a wavelet subband analysis.

Based on the fact that human vision is sensitive to spatial frequencies and as well to moving velocities, a visual measure was proposed to be included in video compression methods, in [102]. Three visual factors are involved : motion attention model, unconstrained eye-movement incorporated spatiovelocity visual sensitivity model, and visual masking model. Quantization parameters at macroblock level for video coding are adjusted on the basis of the combined measure the three factors, i.e. of the measure of motion attention of the current macroblock, the estimated spatiovelocity and the spatiotemporal distortion masking measure.

Research efforts on foveation techniques led to more efficient image and video coding systems in [103]. It is based on the phenomenon of point of fixation : it removes high frequency information redundancy from the regions away from the fixation point. The fovea is the region of highest visual acuity (see Figure 1.1) because of its high concentration of cones. The point a human observer gazes at is called a fixation point or foveation point because it is projected on the fovea. The visual sensitivity decreases dramatically with distance from the fixation point. As a result, removing redundancy is the major motivation of this technique since the bandwidth can be reduced. Foveation techniques are used in many other application fields such as image quality assessment, image segmentation, stereo 3D scene perception and volume data visualization.

3.3 Super-resolution

Sadaka et al. proposed a Perceptually Attentive Super-Resolution (PASR) method [104]. In this method, a perceptual model based on just noticeable distortion thresholds is utilized to select the active pixels which need to be processed by the Super-resolution algorithm. These active pixels are iterated upon until the desired visual quality is reached. After that, a visual attention model is used to get the attended regions in which the active pixels are processed at a higher accuracy by the SR algorithm compared with the pixels lying in other regions. By using the visual attention model, the proposed method significantly reduce the computational complexity of the super-resolution algorithm without loss of the desired perceptual quality.

Super-resolution (SR) methods are widely utilized to get High-Resolution (HR) images, which are unaliased and sharp/deblurred, by combining information from multiple Low-Resolution (LR) frames of the same scene. However, the SR techniques always suffer from high computational complexity.

On the other hand, Perceptual Quality Metrics (PQM) is usually utilized to imitate the human visual system perception of distortions in order to improve the assessment of image/video quality. Due to the human visual attention, artifacts in salient regions are likely to be more annoying to the observer than artifacts in less salient regions. Thus, the visual attention based PQM can be effectively used to assess the visual quality of images.

In this Perceptually Attentive Super-Resolution method, Sadaka et al. first utilize the visual attention model purposed by Itti et al. to detect attended regions. This model is a bottom-up approach based on neural receptive field stimuli of low-level image features of contrast in intensity and orientation. Besides the visual attention model, the PASR method used a Just Noticeable Difference model proposed by Ferzli et al to measure the contrast sensitivity threshold.

The MAP-based SR framework proposed by Hardie et al is used to estimate the HR image. Being different from the original framework, only a subset of pixels, referred to as active pixels, are updated at each iteration in PASR. The selection of active pixels is based on comparisons with locally computed Just Noticeable Difference threshold. The existence of active pixels and iteration will not stop until the threshold is reached.

The PASR method also divides the image into a foreground and a back-

ground based on the result of visual attention computational model. The threshold to the foreground region has a lower threshold compared to the background region. Therefore, the subset of active pixels in the salient foreground region is treated with a higher accuracy. On the other hand, the algorithm continues iterating on a smaller subset of active pixels belonging to the visual salient areas, so the number of pixels that need to be processed is significantly reduced without a visually perceived loss in quality. The numbers of active pixels per iteration are shown in figure 3.9.

The performance of this Perceptually Attentive Super-Resolution, a Super-resolution algorithm combing the perceptual model and visual attention model, is shown in figure 3.8. Figure 3.10 shows the performance.

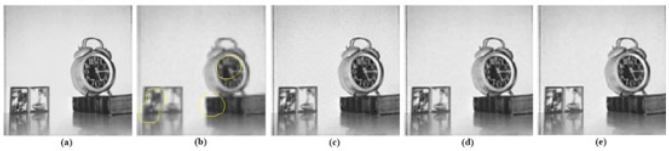

FIGURE 3.8 – Visual performance results for the 256x256 CLOCK image with a resolution enhancement factor 4, number LR images = 4, and additive Gaussian noise variance = 4, regularization operator = 150. (a) Original image. (b) Bilinearly interpolated LR image with 3 VA regions. (c) MAP SR method [4], $\varepsilon = 0.0001$. (d)SELP SR method [3], $\varepsilon = 0.0001$. (e) Proposed PASR, $\varepsilon = 0.0001$, >=15, VA regions = 3.

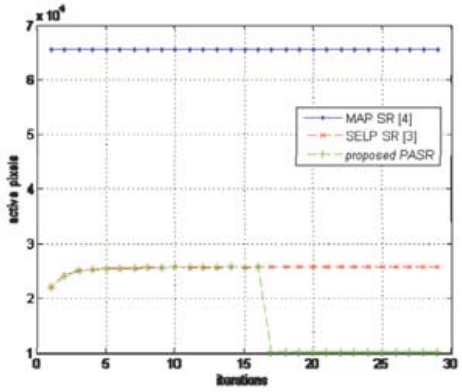

FIGURE 3.9 – Active Pixels per Iteration.

	The CLOCK image	
	PSNR(dB)	Pixel Saving
Bilinear	24.0975	0%
MAP SR	28.3438	0%
SELP SR	28.5747	62.32%
PASR	28.5324	**72.71%**

FIGURE 3.10 – PSNR and Percentage Pixel Savings Result

3.4 Adaptive 3D rendering

Another application of the perceptual modeling of human visual system and visual attention is an adaptive 3d Rendering method based on Region-of-Interest proposed by Chamaret et al [105].

This paper proposed a post-processing method to improve the Quality of Experience (QoE) of 3DTV by utilizing visual attention model. It is widely accepted that the mismatch of accommodation and vergence of 3D video causes the visual strain. This kind of mismatch is especially serious when objects pop out from the screen (i.e. be with a positive parallax value). This causes most visual strain. However, this kind of mismatch does not exist at the zero parallax plane (i.e. the screen).

Therefore, if people looks at the objects in front of the screen for a long time, or the distance between the object and the screen is too large, the visual strain increases and the QoE declines. Consequently, to improve the QoE, it is possible to limit the visual strain by moving the areas with a large parallax value to the zero parallax plane.

To solve the problem, it is necessary to : (1) find the ROI which attracts observers' attention ; (2)move this area to the zero parallax plane ; (3)keep observers' attention on the ROI. The diagram is shown in figure 3.11 :

3.5 Motion Sharpening

The studies plan to extend to the temporal domain the characterization of the HVS, exactly the idea is to extend the spatial CSF to the spatio-temporal domain. The simpler way consists in computing the spatio-temporal CSF, as the product of the spatial CSF by the temporal CSF. But this decoupled

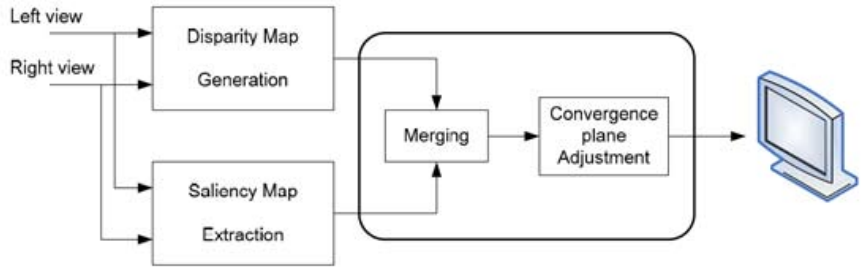

FIGURE 3.11 – Overview block diagram

model is not appropriate, in particular at low frequencies, because the spatial and temporal proprieties are very linked[106, 107].

Some works have also demonstrated the importance of the motion for the contrast perception (with the increase of the thresholds for the detection[108]). Other works[109] have showed that the motion is able to modify the perception of the blur. In this context we address the phenomenon known as the "motion sharpening" (MS).

Indeed the motion can improve the visual quality : some static and blurred images can be perceived sharper when they are animated (note : the "motion blurring" is the opposite effect, when the motion blurs the perception of objects that are sharp when they are static).

The MS have been studied in the literature using natural images[110, 111] or synthetic stimuli[112, 113, 114, 115, 116, 117, 118, 119]. Only few papers aims at modeling the MS. [117]have showed that the perception of the blur is inversely proportional to the motion of the moving objects. [120] have showed that the perceived sharpness decreases when the number of objects increases in the sequence. [111] have showed that if the sequence contains only moving objects with a low saliency, the perceived sharpness is low because these objects don't belong to a region of interest. [114, 115, 116] wanted at finding the parameters that characterized the MS, he showed that the ability of the HVS to discriminate "patterns" decreases with the motion. [118] have also demonstrated that the motion can decrease the discrimination of the blur. [113, 121] have showed the independence between the MS and the contrast.

Some works aims at exploiting the MS for video coding applications. As example we can cite[111] who proposed to use a (spatio-temporal) lowpass filter to smooth the sequence before its coding. They explained that 40% of the information can be reduced using this method. [122, 123] proposed to fil-

ter some images of the sequence (a sort of temporal sub-sampling) before the coding. The filtering takes into account of the characteristics that influence the MS. Using a H.264 coder, from 9.13% to 14,51% of the bit-rate have been saved without visible distortions.

3.6 Conclusion

This section presented many applications for perceptual models. Concerning the quality metrics, as much as it has been widely explored for 2D applications, 3D is still under investigation. For coding applications, even though HVS-based systems are thought to be more likely respectful towards human perception, it is difficult to evaluate the codecs. This section also introduced the super-resolution technique that allows to embed the accuracy in some specific regions ; the adaptive 3D rendering to improve the quality of experience of 3DTV ; and motion sharpening that can be exploited for coding applications by taking into account the motion in a sequence.

Bibliographie

[1] C. Perrin, "Oeil humain," http ://www.biologieenflash.net/animation.php ?ref=bio-0029-2, Feb. 2010. 8

[2] B. A. Wandell, *Foundations of vision*, Sinauer Associates, 1995. 8

[3] K. Ramamohan Rao, *Digital video image quality and perceptual coding*, CRC Press, 2006. 8

[4] S.E. Palmer, *Vision science : Photons to phenomenology*, MIT press Cambridge, MA., 1999. 9, 10, 16, 17, 18, 20

[5] J.J. Gibson and L. Carmichael, *The perception of the visual world*, Houghton Mifflin Boston, 1950. 9, 18

[6] D. Marr and H.K. Nishihara, "Representation and recognition of the spatial organization of three-dimensional shapes," *Proceedings of the Royal Society of London. Series B, Biological Sciences*, vol. 200, no. 1140, pp. 269–294, 1978. 9

[7] H. Wallach and L. Floor, "The use of size matching to demonstrate the effectiveness of accommodation and convergence as cues for distance," *Perception & Psychophysics*, 1971. 10

[8] K. Nakayama and S. Shimojo, "Da vinci stereopsis : depth and subjective occluding contours from unpaired image points," *Vision Research*, vol. 30, no. 11, pp. 1811–1825, 1990. 13

[9] J.J. Gibson, "The senses considered as perceptual systems," *Massachusetts : Houghton-Mifflin*, 1966. 15

[10] H. Wallach and DN O'Connell, "The kinetic depth effect," *Journal of Experimental Psychology*, vol. 45, pp. 205–217, 1953. 15

[11] J.J. Gibson, G.A. Kaplan, H.N. Reynolds Jr, and K. Wheeler, "The change from visible to invisible : A study of optical transitions.," *Perception & Psychophysics*, 1969. 15

[12] H.A. Sedgwick, "Space perception," *Handbook of perception and human performance.*, vol. 1, pp. 21–1, 1986. 16

[13] K. Ozkan and M.L. Braunstein, "Background surface and horizon effects in the perception of relative size," *Journal of*, vol. 6, no. 6, pp. 421, 2006. 17

[14] D. Marr, *Vision : A computational investigation into the human representation and processing of visual information*, Henry Holt and Co., Inc. New York, NY, USA, 1982. 19

[15] J. Malik and R. Rosenholtz, "A computational model for shape from texture," in *Ciba Foundation Symposium*, 1994, vol. 184, p. 272. 19

[16] J.J. Clark and A.L. Yuille, *Data fusion for sensory information processing systems*, Springer, 1990. 21

[17] M.S. Landy, L.T. Maloney, E.B. Johnston, and M. Young, "Measurement and modeling of depth cue combination : In defense of weak fusion," *Vision research*, vol. 35, no. 3, pp. 389–412, 1995. 21

[18] A. Poole and L. J Ball, "Eye tracking in human-computer interaction and usability research : current status and future prospects," *Encyclopedia of human computer interaction*, pp. 211–219, 2005. 22, 23, 24, 26

[19] K. Rayner and A. Pollatsek, *The psychology of reading*, Lawrence Erlbaum, 1989. 22

[20] A. T Duchowski, *Eye tracking methodology : Theory and practice*, Springer-Verlag New York Inc, 2007. 22

[21] M. A Just and P. A Carpenter, "Eye fixations and cognitive processes.," *Cognitive Psychology*, vol. 8, no. 4, pp. 441–480, 1976. 23, 24

[22] J. H Goldberg and X. P Kotval, "Computer interface evaluation using eye movements : Methods and constructs," *International Journal of Industrial Ergonomics*, vol. 24, no. 6, pp. 631–645, 1999. 23, 24, 25

[23] J. Wang, "Quantifying the relationship between visual salience and visual importance," San Jose, 2010. 24, 26, 29

[24] R. Jacob and K. Karn, "Eye tracking in human computer interaction and usability research : ready to deliver the promises," 2003, vol. 2, p. 4. 24

[25] C. Mello-Thoms, C. F Nodine, and H. L Kundel, "What attracts the eye to the location of missed and reported breast cancers ?," in *Proceedings of the 2002 symposium on Eye tracking research & applications*, 2002, p. 117. 24

[26] L. Cowen, L. J Ball, and J. Delin, "An eye movement analysis of webpage usability," in *People and Computers XVI : Memorable Yet Invisible, Proceedings of HCI*, 2002, pp. 317–335. 24

[27] M. D Byrne, J. R Anderson, S. Douglass, and M. Matessa, "Eye tracking the visual search of click-down menus," in *Proceedings of the SIGCHI conference on Human factors in computing systems : the CHI is the limit*, 1999, p. 409. 24

[28] B. Cassin, S. A.B Solomon, M. L Rubin, and M. Polasky, *Dictionary of eye terminology*, Triad Pub. Co., 2001. 24

[29] J. H Goldberg, M. J Stimson, M. Lewenstein, N. Scott, and A. M Wichansky, "Eye tracking in web search tasks : design implications," in *Proceedings of the 2002 symposium on Eye tracking research & applications*, 2002, pp. 51–58. 25

[30] D. Bruneau, M. A Sasse, and J. D. McCarthy, "The eyes never lie : The use of eye tracking data in HCI research," in *Proceedings of the CHI*, 2002, vol. 2. 25

[31] J. B Brookings, G. F Wilson, and C. R Swain, "Psychophysiological responses to changes in workload during simulated air traffic control," *Biological Psychology*, vol. 42, no. 3, pp. 361–377, 1996. 25

[32] S. P Marshall, *Method and apparatus for eye tracking and monitoring pupil dilation to evaluate cognitive activity*, Google Patents, 2000, US Patent 6,090,051. 25

[33] J. R Lackner, "Induction of illusory self-rotation and nystagmus by a rotating sound-field," *Aviation, Space, and Environmental Medicine*, vol. 48, no. 2, pp. 129–131, 1977. 26

[34] U. Engelke, H. J Zepernick, and A. Maeder, "Visual attention modeling : region-of-interest versus fixation patterns," in *Proceedings of the 27th conference on Picture Coding Symposium*, 2009, pp. 52–524. 26

[35] D. D Salvucci and J. H Goldberg, "Identifying fixations and saccades in eye-tracking protocols," in *Proceedings of the 2000 symposium on Eye tracking research & applications*, 2000, pp. 71–78. 26, 27

[36] H. Widdel, "Operational problems in analysing eye movements," *Advances in Psychology*, pp. 21–29, 1984. 27

[37] E. Niebur, "Saliency map," *Scholarpedia*, vol. 2, no. 8, pp. 2675, 2007. 28

[38] A. Boev, M. Poikela, A. Gotchev, and A. Aksay, "Modelling of the stereoscopic HVS," . 31, 32, 33, 69

[39] M. J Nadenau, S. Winkler, D. Alleysson, and M. Kunt, "Human vision models for perceptually optimized image processing-a review," *Proceedings of the IEEE*, vol. 2000, 2000. 32, 34

[40] M. A Garcia-Perez, "The perceived image : Efficient modelling of visual inhomogeneity," *Spatial vision*, vol. 6, no. 2, pp. 89–99, 1992. 32

[41] W. F Schreiber, *Fundamentals of electronic imaging systems : some aspects of image processing*, Springer-Verlag, 1993. 32

[42] P. G.J Barten, "Contrast sensitivity of the human eye and its effects on image quality," 1999. 33

[43] E. M. Granger and J. C. Heurtley, "Letters to the editor : Visual chromaticity-modulation transfer function.," *Journal of the Optical Society of America*, vol. 63, no. 9, pp. 1173, 1973. 33

[44] K. T. Mullen, "The contrast sensitivity of human colour vision to red-green and blue-yellow chromatic gratings.," *The Journal of Physiology*, vol. 359, no. 1, pp. 381, 1985. 33

[45] G. J. Van der Horst and M. A. Bouman, "Spatiotemporal chromaticity discrimination.," *Journal of the Optical Society of America*, vol. 59, no. 11, pp. 1482, 1969. 33

[46] M. J. Nadenau, J. Reichel, and M. Kunt, "Wavelet-based color image compression : Exploiting the contrast sensitivity function," *IEEE Transactions on Image Processing*, vol. 12, no. 1, pp. 58–70, 2003. 33

[47] S. Daly, "The visible differences predictor : an algorithm for the assessment of image fidelity," *Digital images and human vision*, vol. 11, 1993. 33

[48] S. Daly, "Engineering observations from spatiovelocity and spatiotemporal visual models," *Vision Models and Applications to Image and Video Processing*, pp. 179–200. 34

[49] D. H. Kelly, "Spatiotemporal variation of chromatic and achromatic contrast thresholds," *J. Opt. Soc. Am*, vol. 73, pp. 742–749, 1983. 34

[50] J Yang and W Makous, "Spatiotemporal separability in contrast sensitivity," *Vision research*, vol. 34, pp. 2569–2576, 1994, 19. 34

[51] S. A Klein, T. Carney, L. Barghout-Stein, and C. W Tyler, "Seven models of masking," in *Proceedings of SPIE*, 1997, vol. 3016, p. 13. 34

[52] A. B Watsona, R. Borthwickb, and M. Taylorb, "Image quality and entropy masking," . 34

[53] A. J Ahumada Jr, B. L Beard, and R. Eriksson, "Spatio-temporal discrimination model predicts temporal masking function," in *Proc. SPIE*, 1998, vol. 3299, p. 120. 34

[54] J. K Tsotsos, S. M Culhane, W. Y Kei Wai, Y. Lai, N. Davis, and F. Nuflo, "Modeling visual attention via selective tuning," *Artificial intelligence*, vol. 78, no. 1-2, pp. 507–545, 1995. 35

[55] O. Le Meur, P. Le Callet, D. Barba, and D. Thoreau, "A coherent computational approach to model bottom-up visual attention," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 28, pp. 802–817, 2006, 5. 36, 39

[56] O. Le Meur and P. Le Callet, "What we see is most likely to be what matters : Visual attention and applications," in *2009 16th IEEE International Conference on Image Processing*, 2009, pp. 3085–3088. 36

[57] L. Itti and P. Baldi, "Bayesian surprise attracts human attention," *Vision research*, vol. 49, pp. 1295–1306, 2009, 10. 37, 43, 45

[58] G. Kootstra, A. Nederveen, and B. de Boer, "Paying attention to symmetry," 2008, pp. 1115–1125. 37

[59] J. Woods Andrew, S. Holliman Nicolas, A. Dodgson Neil, J. Hakkinen, T. Kawai, J. Takatalo, R. Mitsuya, and G. Nyman, "What do people look at when they watch stereoscopic movies ?," 2010, vol. 7524, p. 75240E, SPIE. 37

[60] R. Gal and D. Cohen-Or, "Salient geometric features for partial shape matching and similarity," *ACM Transactions on Graphics (TOG)*, vol. 25, pp. 150, 2006, 1. 37

[61] L. Itti, C. Koch, and E. Niebur, "A model of saliency-based visual attention for rapid scene analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 20, pp. 1254–1259, 1998, 11. 38

[62] C. Koch and S. Ullman, "Shifts in selective visual attention : towards the underlying neural circuitry," *Hum Neurobiol*, vol. 4, pp. 219–27, 1985, 4. 38, 39

[63] A. Treisman and G. Gelade, "A feature-integration theory of attention," *Cognitive Psychology*, vol. 12, pp. 97–136, 1980, 1. 38, 39

[64] L. Zhang, M. Tong, T. Marks, H. Shan, and G. Cottrell, "SUN : a bayesian framework for saliency using natural statistics," *Journal of Vision*, vol. 8, pp. 32, 2008, 7. 39, 41

[65] O. Le Meur, P. Le Callet, and D. Barba, "Predicting visual fixations on video based on low-level visual features," *Vision research*, vol. 47, pp. 2483–2498, 2007, 19. 39, 42

[66] S. Daly, EK Co, and NY Rochester, "A visual model for optimizing the design of image processingalgorithms," 1994, vol. 2. 40

[67] P. Le Callet, A. Saadane, and D. Barba, "Interactions of chromatic components on the perceptual quantization of the achromatic component," 1999, vol. 3644. 40

[68] Z. Li, "A neural model of contour integration in the primary visual cortex," *Neural computation*, vol. 10, pp. 903–940, 1998, 4. 41

[69] D. Gao, V. Mahadevan, and N. Vasconcelos, "On the plausibility of the discriminant center-surround hypothesis for visual saliency," *Journal of Vision*, vol. 8, pp. 13, 2008, 7. 45, 46

[70] N. Bruce and J. Tsotsos, "Saliency, attention, and visual search : An information theoretic approach," *Journal of Vision*, vol. 9, pp. 5, 2009. 46, 47

[71] A. Oliva, A. Torralba, M. Castelhano, and J. Henderson, "Top-down control of visual attention in object detection," 2003, vol. 1. 48, 49

[72] N. Ouerhani and H. Hugli, "Computing visual attention from scene depth," in *International conference on Pattern Recognition*, 2000, vol. 15, pp. 375–378, saliency + Z. 50, 51

[73] Y. Zhang, G. Jiang, M. Yu, and K. Chen, "Stereoscopic visual attention model for 3d video," *Advances in Multimedia Modeling*, pp. 314–324, 2010, saliency + Z. 51

[74] M.C. Park and K.J. Cheoi, "Automatic focusing attention for a stereo pair of image sequence," in *Proceedings of the 2nd International Conference on Interaction Sciences : Information Technology, Culture and Human*. ACM, 2009, pp. 1185–1190, saliency + Z. 52, 53, 57

[75] A. Maki, P. Nordlund, and J.O. Eklundh, "A computational model of depth-based attention," in *International conference on Pattern Recognition*. Citeseer, 1996, vol. 13, pp. 734–739, saliency + Z. 54, 55

[76] A. Fernandez-Caballero, M.T. López, and S. Saiz-Valverde, "Dynamic stereoscopic selective visual attention (dssva) : Integrating motion and shape with depth in video segmentation," *Expert Systems with Applications*, vol. 34, no. 2, pp. 1394–1402, 2008, To read stereo saliency. 55, 57

[77] Z. Wang, A. C Bovik, H. R Sheikh, and E. P Simoncelli, "Image quality assessment : From error visibility to structural similarity," *IEEE transactions on image processing*, vol. 13, no. 4, pp. 600–612, 2004. 60

[78] Z. Wang, L. Lu, and A. Bovik, "Video quality assessment based on structural distorsion measurement," *Signal processing : Image communication*, vol. 19, no. 2, pp. 121–132, Feb. 2004. 62, 63

[79] M. H Loke, E. P Ong, W. Lin, Z. Lu, and S. Yao, "Comparison of video quality metrics on multimedia videos," in *2006 IEEE International Conference on Image Processing*, 2006, pp. 457–460. 63

[80] D. M Chandler and S. S Hemami, "VSNR : a wavelet-based visual signal-to-noise ratio for natural images," *IEEE Transactions on Image Processing*, vol. 16, no. 9, pp. 2284, 2007. 63

[81] J. Lubin and D. Fibush, "Sarnoff JND vision model," *T1A1*, vol. 5, pp. 97–612, 1997. 63

[82] A. B Watson, J. Hu, and J. F McGowan III, "Digital video quality metric based on human vision," *Journal of Electronic imaging*, vol. 10, pp. 20, 2001. 63

[83] H. R Sheikh and A. C Bovik, "Image information and visual quality," *IEEE Transactions on Image Processing*, vol. 15, no. 2, pp. 430–444, 2006. 64

[84] M. H Pinson and S. Wolf, "A new standardized method for objectively measuring video quality," *IEEE Transactions on broadcasting*, vol. 50, no. 3, pp. 312–322, 2004. 64

[85] M. Carnec, P. Le Callet, and D. Barba, "Objective quality assessment of color images based on a generic perceptual reduced reference," *Signal Processing : Image Communication*, 2008. 65

[86] M. Carnec, P. Le Callet, and D. Barba, "An image quality assessment method based on perception of structural information," in *2003 International Conference on Image Processing*, Spain, 2003, pp. 14–17. 64

[87] M. Meesters, W. Ijsselsteijn, and P. Seuntiens, "A survey of perceptual evaluations and requirements of three dimensional TV," *IEEE Transactions on Circuits And Systems for Video Technology*, vol. 14, no. 3, pp. 381–391, Mar. 2004. 65

[88] A. Schertz, "Source coding of stereoscopic television pictures," in *Image Processing and its Applications, 1992., International Conference on*, 1992, pp. 462–464. 66

[89] H. Asher, "Suppression theory of binocular vision," *The British Journal of Ophthalmology*, vol. 37, no. 1, pp. 37, 1953. 66

[90] W. A IJsselsteijn5a, H. de Ridderb, J. Freemanc, and S. E. Avonsd, "Presence : Concept, determinants and measurement," . 66

[91] W. IJsselsteijn, H. Ridder, J. Freeman, S. E. Avons, and D. Bouwhuis, "Effects of stereoscopic presentation, image motion, and screen size on subjective and objective corroborative measures of presence," *Presence : Teleoperators & Virtual Environments*, vol. 10, no. 3, pp. 298–311, 2001. 66

[92] W. IJsselsteijn, H. de Ridder, R. Hamberg, D. Bouwhuis, and J. Freeman, "Perceived depth and the feeling of presence in 3DTV," *Displays*, vol. 18, no. 4, pp. 207–214, 1998. 66

[93] P. Seuntiens, *Visual Experience of 3D TV*, Ph.D. thesis, 2006. 66, 67

[94] A. Benoit, P. Le Callet, P. Campisi, and R. Cousseau, "Quality assesment of stereoscopic images," 2008. 67

[95] A. Boev, A. Gotchev, K. Egiazarian, A. Aksay, and G. B Akar, "Towards compound stereo-video quality metric : a specific encoder-based framework," in *Proc. Southwest Symp. Image Analysis and Interpretation (SSIAI 2006)*, 2006, pp. 218–222. 68, 69

[96] H. Shao, X. Cao, and G. Er, "Objective quality assessment of depth image based rendering in 3DTV system," in *Proc. IEEE 3DTV Conference, Potsdam, Germany*, 2009. 68, 72

[97] Z. Zhu and Y. Wang, "Perceptual distortion metric for stereo video quality evaluation," *Wseas Transactions*, vol. 5, no. 7, 2009. 70, 71

[98] H. Lin and A. N. Venetsanopoulos, "Fast pyramidal search for perceptually based fractal imagecompression," in *Image Processing, 1996. Proceedings., International Conference on*, 1996, vol. 1. 73

[99] D. Wu, D. M. Tan, M. Baird, J. DeCampo, C. White, and H. R Wu, "Perceptually lossless medical image coding," *IEEE transactions on medical imaging*, vol. 25, no. 3, pp. 335–344, 2006. 73

[100] L. Itti, *Models of Bottom-Up and Top-Down Visual Attention*, Ph.D. thesis, Pasadena, California, Jan 2000. 73

[101] N. Tsapatsoulis, K. Rapantzikos, and C. Pattichis, "An embedded saliency map estimator scheme : Application to video encoding," . 74

[102] C. W Tang, "Spatiotemporal visual considerations for video coding," *Multimedia, IEEE Transactions on*, vol. 9, no. 2, pp. 231–238, 2007. 74

[103] Z. Wang and A. C Bovik, "Foveated image and video coding," *Digital Video, Image Quality and Perceptual Coding*, pp. 431–457. 74

[104] N. G Sadaka and L. J Karam, "Perceptual attentive superresolution," in *International Workshop VPQM*, 2009. 75

[105] C. Chamareta, S. Godeffroya, P. Lopeza, and O. Le Meura, "Adaptive 3D rendering based on Region-of-Interest," in *Proceedings of SPIE*, 2010, vol. 7524, p. 75240V. 77

[106] J. Robson, "Spatial and temporal contrast sensitivity functions of the visual system.," *Journal of the Optical Society of America.*, vol. 56, pp. 1141–1142, 1966. 78

[107] J. Koenderink and A. Doorn, "The international representation of solid shape with respect to vision.," *Biological Cybernetics, Springler.*, vol. 32, pp. 211–216, 1979. 78

[108] D. H. Kelly, "Motion and vision. II. stabilized spatio-temporal threshold surface.," *Journal of the Optical Society of America.*, vol. 69, pp. 1340–1349, 1979. 78

[109] V. S. Ramachandran, V. M. Rao, and T. R. Vidyasagar, "Sharpness constancy during movement perception (short note).," *Perception.*, vol. 3, no. 1, pp. 97–98, 1974. 78

[110] T. Takeuchi and K. K. De Valois, "Motion sharpening in moving natural images.," *Journal of Vision.*, vol. 2, no. 7, pp. 377, 2002. 78

[111] T. Takeuchi and K. K. De Valois, "Sharpening image motion based on the spatio-temporal characteristics of human vision.," in *SPIE Electronic Imaging.*, 2005 (San Jose, USA). 78

[112] D. C. Burr and M. J. Morgan, "Motion deblurring in human vision.," *Proceedings of the Royal Society B : Biological Sciences.*, vol. 264, pp. 431–436, 1997. 78

[113] M. A. Georgeson and S. T. Hammett, "Seeing blur : 'motion sharpening' without motion.," *Proceedings of the Royal Society B : Biological Sciences.*, vol. 269, no. 1429-1434, 2002. 78

[114] S. T. Hammett and P. J. Bex, "Motion sharpening : Evidence for the addition of high spatial frequencies to the effective neural image.," *Vision Research.*, vol. 36, pp. 2729–2733, 1996. 78

[115] S. T. Hammett, "Motion blur and motion sharpening in the human visual system.," *Vision Research.*, vol. 37, no. 18, pp. 2505–2510, 1997. 78

[116] S. T. Hammet, M. A. Georgeson, and A. Gorea, "Motion blur and motion sharpening : temporal smear and local contrast non-linearity.," *Vision Research.*, vol. 38, no. 14, pp. 2099–2108, 1998. 78

[117] P. J. Bex, G. K. Edgar, and A. T. Smith, "Sharpening of drifting, blurred images.," *Vision Research.*, vol. 35, no. 18, pp. 2539–2546, 1995. 78

[118] A. K. Pääkkönen and M. J. Morgan, "Linear mechanisms can produce motion sharpening.," *Vision Research.*, vol. 41, pp. 2771–2777, 2001. 78

[119] D. Burr and J. Ross, "Visual processing of motion.," *Trends in Neurosciences.*, vol. 9, pp. 304–307, 1986. 78

[120] S. Chen, H. E. Bedell, and H. Ogmen, "A target in real motion appears blurred in the absence of other proximal moving targets.," *Vision Research.*, vol. 35, pp. 2315–2328, 1995. 78

[121] S. T. Hammett, M. A. Georgeson, S. Bedingham, and G. S. Barbieri-Hesse, "Motion sharpening and contrast : Gain control precedes compressive non-linearity ?," *Vision Research.*, vol. 43, pp. 1187–1199, 2003. 78

[122] A. Fujibayashi and C. S. Boon, "A masking model for motion sharpening phenomenon in video sequences.," *IEICE TRANSACTIONS on Fundamentals of Electronics, Communications and Computer Sciences.*, vol. 91, no. 6, pp. 1408–1415, 2008. 78

[123] A. Fujibayashi and C. S. Boon, "Application of motion sharpening effect in video coding.," in *IEEE International Conference on Image Processing (ICIP).* San Diego, CA., October. 2008, pp. 2848–2851. 78