**Science Research**
2014; 2(5): 135-149
Published online November 20, 2014 (http://www.sciencepublishinggroup.com/j/sr)
doi: 10.11648/j.sr.20140205.20
ISSN: 2329-0935 (Print); ISSN: 2329-0927 (Online)

# Differences between the actual scene model and the image model for computation of visual depth information of early vision

**Zhao Songnian[1, *], Yu Yunxin[2], Zhao Yuping[3], Jin Xi[4], Cheng Wenjun[1]**

[1]LAPC, Institute of Atmospheric Physics, Chinese Academy of Sciences, Beijing 100029, China
[2]Institute of Geophysics of SSB, Beijing 100029, China
[3]Beijing Jetsen Century Technology Co., Ltd., Beijing 100191, China
[4]China rescue and salvage of ministry of the People's Republic of China, Beijing 100736, China

## Email address:
zsnzhao@126.com (Zhao S.)

## To cite this article:
Zhao Songnian, Yu Yunxin, Zhao Yuping, Jin Xi, Cheng Wenjun. Differences between the Actual Scene Model and the Image Model for Computation of Visual Depth Information of Early Vision. *Science Research.* Vol. 2, No. 5, 2014, pp. 135-149.
doi: 10.11648/j.sr.20140205.20

**Abstract:** In this paper, we introduced two viewing modes, "Scene Mode" and "Picture Mode", for early visual depth perception depending on the dimensions of the object being viewed. The essential difference between these two modes of visual depth perception is still unclear. We discuss the basic methods of introducing a three-dimensional Cartesian system into a plane to express the depth information of an image, estimate the loss of depth information caused by this approach, and provide an analysis of the important role of providing depth information based on size constancy and vanishing point in the two viewing modes. We studied the problem of how the retina and visual cortex separate the plenoptic (all-optical) function, which is the input representation of vision, by neural computing in scene mode. We also studied the problem of how to extract information about the position and angle of light beams in the light field, and then determined the output representation of the visual depth perception. In the absence of any stereoscopic cues, such as texture, gradient, shade, shadow, color, occlusion, and binocular disparity, we compare the main differences of visual depth perception between scene mode and picture mode using a cube being viewed and its line drawing, which respectively represent the two modes.

**Keywords:** Vision, Perceptual Mode, Vanishing Point, Dimensions, Size Constancy

## 1. Introduction

The main process of vision is first to "look" at objects and the environment; the viewer then "sees" the outside world, finds something of interest, and then extracts and processes related information from this visual input [1-4]. In the past, vision and visual perception research has focused on the neural response to specific stimuli; owing to the limit of current experimental equipment and methods (e.g., fMRI, ERPs), visual stimulus images rarely involve actual scenes. Comparative studies of the visual perception mechanism for actual scenes and stereo images on a 2D plane are also rare [5-9]. In fact, complex visual objects can in essence be divided into two categories according to the physical space dimension, namely three-dimensional actual scenes and various images displayed on a 2D plane, such as television, video, photographs, pictures, and paintings. In particular, the difference between the stereoscopic perception of viewing an actual scene and the stereoscopic perception obtained from viewing a photographic image of the same scene is worthy of in-depth study as a basic theory problem. In this paper, the term "three-dimensional perception" means the depth perception obtained from an actual scene, and the term "stereoscopic perception" means the depth perception obtained from pictures, to show that they are different in physical space dimensions. This problem is closely related to neural computing, visual information, cognitive psychology, computer vision, and image processing. It should be particularly mentioned that the main content of this paper and the references cited are only limited to human early vision and

visual depth perception, and do not involve computer vision, or photography.

To study the above issues, a useful way to begin is to ask: why are stereo visual perceptions formed from images and paintings displayed on a flat surface? For example, a transparent cube drawn in the plane (e.g., a Necker cube [10-11]), even without a texture gradient, light and shade, shadows, color and occlusion cues and binocular parallax, is very naturally perceived as a cube. Similarly, when a meandering line is drawn in the plane that extends into the distance, a stereoscopic perception of a river is easily generated. How is such a visual perception effect formed? There is at present no relatively clear understanding of this phenomenon, let alone a theoretical neuroscientific explanation. It is quite commonplace, and we are accustomed to it; but even with careful consideration, it is not easy to explain. Marr, as one of the founders of visual computational theory and computer vision, expressed his confusion on this issue: when proposing the principle of graceful degradation, he pointed out that if the visual system calculates a rough two-dimensional description from an image, it will be able to calculate a rough three-dimensional description represented by this image. In other words, human vision can perceive a real three-dimensional description from a stereoscopic image on a two-dimensional plane. Marr asked: "Contours of the images are two-dimensional, but we are often to understand these contours from a three-dimensional perspective. Therefore, the key question is how do we make a three-dimensional interpretation of the two-dimensional contour? Why can we make such an interpretation?" [2].

When viewing a real 3D scene, although a two-dimensional visual image is formed on the retina, the viewer is able to generate three-dimensional visual perception and can perceive the three-dimensional structure of the actual scene. However, when viewing various images on a two-dimensional plane, the stereoscopic structure of the image can be perceived. In this article, to compare these with each other, we denote viewing an actual scene as "Scene Mode", and denote viewing various images on a two-dimensional plane as "Picture Mode". Obviously, Marr raised issues related to the basic characteristics of scene mode and picture mode, and also related the similarities and differences of their neural mechanisms. Thus, this is a basic problem in vision, cognitive psychology, computer graphics, computer vision, image rendering, 3D display and computational photography. However, this article's scope is mainly limited to the aspect of human early vision.

Scientific and technical personnel in these areas have been exploring this basic question of human depth perception, particularly various cues of three-dimensional perception [12-20]. They have also focused on how to better express the three-dimensional structure of the scene through image rendering methods [21-23], and how to better improve the quality of 3D displays. Currently, many of these results have already been achieved. However, for the study of human early visual depth perception, a more appropriate research approach involves excluding a variety of cues that can cause three-dimensional visual perception (such as texture, gradient, shading, shadow, color, occlusion, any other three dimensional cues and binocular disparity). We also use a simple line drawing of a cube as a basic stimulus pattern (and therefore do not have to consider the complex structure of the image [24]). Finally, we take the question, "Why can we express a three-dimensional structure in a two-dimensional plane?" as a key research topic. In our research, we discuss a basic method of expressing the stereoscopic structure of an image on a two-dimensional plane, and provide an analysis and an estimation of the loss of depth information caused by this approach. Furthermore, we outline the important roles of size constancy and vanishing point in the stereo perception of images expressed on the plane, as well as conjugate mapping between vanishing point and visual image [25,26]. We also outline the mechanism of neural computations of the retina and the visual cortex using a dual-parameterized method separating full-optical (plenoptic) functions to extract visual information; we determine the plenoptic functions as the input representation of visual depth perception in scene mode, compared with picture mode, in which pictures are input as a representation of human vision.

# 2. 2D and 3D Methods and Geometry Introduction

### *2.1. Basic Representation Method and the Information Loss of Three-Dimensional Visual Perception*

In three-dimensional space, the three axes of the Cartesian rectangular coordinate system are orthogonal to each other (strictly speaking, the three planes (xoy; yoz; zox) are perpendicular to each other). While it is possible to draw up a Cartesian coordinate system in the plane, the three coordinate axes are not orthogonal to each other; in fact, we also cannot draw the orthogonal coordinate system on the plane. In three-dimensional space, axes (x, y, z) are orthogonal to each other, so to distinguish these cases when the Cartesian coordinate system is represented in a two-dimensional plane, we denote an axis as the $\hat{z}$-axis instead of the z-axis; that is, in the two-dimensional plane, the three axes of the Cartesian coordinate system are denoted by the symbols x, y, and $\hat{z}$, as shown in Figure 1.

In fact, the $\hat{z}$-axis and the xoy plane are in the same actual plane (M-plane), and therefore, it is impossible to actually draw up a $\hat{z}$-axis that is perpendicular to the xoy plane. That is, we cannot really draw the $\hat{z}$-axis perpendicular to the ox axis. However, in Figure 1, note that the included angle between the $\hat{z}$-axis and the ox-axis is $\beta$. When $\beta = 90^{o}$, in the two-dimensional plane, the $\hat{z}$-axis coincides only with the oy-axis. When $\beta = 180^{o}$, the $\hat{z}$-axis coincides only with the ox-axis.

Only when $\beta \neq 0^{o}$ or $\beta \neq 180^{o}$ will we perceive the oy-axis as perpendicular to the $\hat{z}$ox plane, and also perceive the yo$\hat{z}$ plane perpendicular to the $\hat{z}$ox plane. That is to say, the o$\hat{z}$-axis divides the plane M into three distinct parts,

namely xoy, yo $\hat{z}$ and $\hat{z}$ ox. In this case, although the xoy plane and the yo $\hat{z}$ planes are not orthogonal (the included angle between both of them is $\beta$), the xoy plane and the yo $\hat{z}$ plane are both orthogonal to the $\hat{z}$ ox plane. In this case, the human visual system can strongly perceive that the xy $\hat{z}$ coordinates system forms a stereoscopic space. In this article, we will call this "stereoscopic perceptual space" in the sense of cognitive psychology, to distinguish it from three-dimensional physical space, in which planes xoy, yoz and zox are orthogonal to each other. This terminology is only for the sake of brevity.

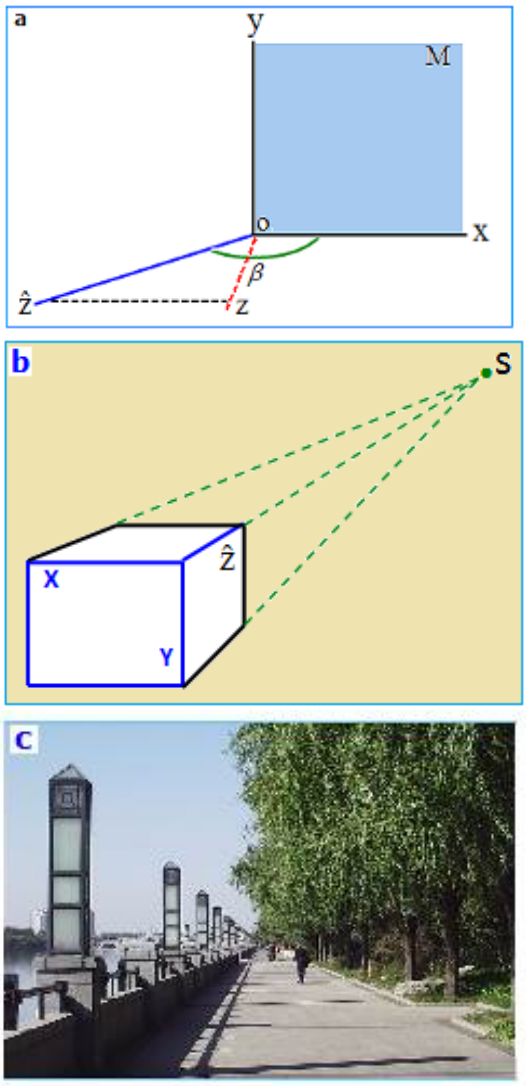


***Figure 1.*** *Stereoscopic perceptual coordinate system on a plane. Human vision can perceive the spatial stereoscopic effects formed by the three coordinates axes x, y and* $\hat{z}$ *. (a) The basic expression of three-dimensional Cartesian coordinate system on a plane; (b) the relationship between the Cartesian coordinate system and the affine coordinate system is established by the vanishing point on the plane; (c) the image of a real scene corresponding to one vanishing point of part (b) (Figure 1(c) with permission from Li Shuzhong and Sung Guangyu).*

To compare the three-dimensional physical space and the psychological stereoscopic space, one can "imagine" the following mathematical treatment: let the xy $\hat{z}$ coordinate system coincide with the xyz coordinate system. The specific steps are: first, make the xoy plane in three-dimensional physical space and the xoy plane in the three-dimensional perception of space overlap. Second, let the $\hat{z}$ ox plane coincide with the zox plane. It is important to note that the yo $\hat{z}$ plane is not coincident with the yoz plane; the included angle between them is $(\beta - 90^o)$ (without considering the sign of the angle). We know that, according to the practice of computer graphics or computer vision, when the xoy plane is set as the imaging plane, the z-axis and the $\hat{z}$ -axis will just carry the depth information of the xyz coordinate system and the xy $\hat{z}$ coordinate system, respectively (of course, this is true for visual perception). We naturally want to know what the difference is between the following two kinds of depth information: one that is obtained in the z-direction while looking at an actual scene (corresponding to the xyz coordinate system), and the other that is perceived in the $\hat{z}$-direction while looking at a picture with the same scene (corresponding to the xy $\hat{z}$ coordinate system).

The easiest way is that the value on the $\hat{z}$ -axis is converted into the one on the z-axis, which is equivalent to the projection of the $\hat{z}$ -axis into the z-axis; its value is denoted by the symbol $z_\perp$, i.e.,

$$z_\perp = \hat{z}\cos(\beta - 90^o) \tag{1}$$

This means that by comparison with the value of corresponding depth information in the z-axis, the degree of narrowing $\Delta d$ of the depth information carried by the $\hat{z}$ -axis, may be calculated as

$$\Delta d = \hat{z} - z_\perp = \hat{z} - \hat{z}\cos(\beta - 90^o) = \hat{z}\,[1 - \cos(\beta - 90^o)] \tag{2}$$

It is not hard to imagine, if the yo $\hat{z}$ plane coincides with the yoz plane, or in psychological space coordinates, that the $\hat{z}$ -axis is really perpendicular to the ox-axis, which means that $\beta = 90^o$ and the loss of information $\Delta d = 0$. This is, of course, impossible, since the so-called "three-dimensional" coordinate system drawn in the plane, which is dependent on the psychological space formed by human visual perception, can be seen as "attached to" the two-dimensional physical plane. Thus, the constraint condition applied by the plane to the psychological space is $\beta \neq 0^o$ and $\beta \neq 90^o$. Perhaps humanity has accumulated such rich experience with visual perception and associated functions in three-dimensional space that even in the absence of stereo cues, such as perspective, texture gradient, shade, shadow, color and occlusion as well as binocular disparity, we are able to naturally perceive that the x-axis, y-axis and $\hat{z}$ -axis in Figure 1 are orthogonal to each other and form a space with stereoscopic properties.

It is important to note that we cannot draw the $\hat{z}$ -axis as truly perpendicular to the plane xoy; that is, we cannot draw the $\hat{z}$ -axis as perpendicular to the ox-axis, which is equivalent to saying that we cannot draw the $\hat{z}$ ox plane perpendicular to the plane xoy. In Figure 1 the plane $\hat{z}$ ox

linearly shrinks toward the ox-axis along the zo-axis, which contracts toward the horizontal vanishing line, as shown in Figure 2. Since the plane $\hat{z}$ ox is naturally tilted, it cannot be perpendicular to plane xoy. The inclined plane actually causes a kind of visual perception of the sky and the ground plane, and plane oxy is perpendicular to plane $\hat{z}$ ox or basic ground.

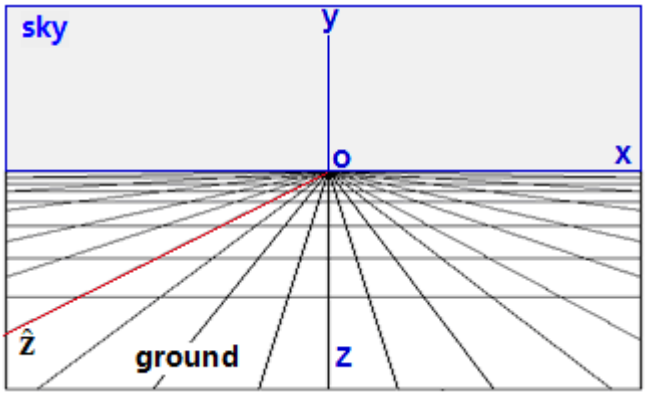


***Figure 2.*** *Linear shrinkage process of a plane toward the horizon (horizontal vanishing line). In this graph the axis* $\hat{z}$ *is also drawn, which is shown in Figure 1. Such a splitting of the plane can obtain the visual effects of three-dimensional perception. When shrinking the plane, the tilt angle* $\beta$ *with respect to the horizontal plane is determined by the angle between the* $\hat{z}$ *-axis and the ox- axis.*

This further shows the important role of the vanishing point and vanishing lines in forming stereoscopic visual perception from a picture on the plane.

In general, compared with the *z*-axis in three-dimensional space, the depth perception of the scene provided by the $\hat{z}$ -axis in the two-dimensional plane will be reduced. The amount of this reduction is given by $\cos(\beta - 90^{\circ})$, which indicates the information loss of stereoscopic perception obtained by viewing a three-dimensional image in the plane. When a three- dimensional image of the actual scene is expressed on a two-dimensional plane in the manner of Figure 1 (this is the only way to express it), the depth information provided in the two-dimensional image, with different angles $\beta$, means that the depth perception of scenes is also different.

It is particularly noteworthy that it is impossible to actually draw the $\hat{z}$ -axis perpendicular to the plane xoy; it is also impossible to really draw the $\hat{z}$ -axis perpendicular to the ox-axis. In Figure 2, it seems that the ground surface is contracting toward the ox -axis, which is also contracting toward the horizontal vanishing line, wherein the ground is naturally tilted so it cannot be perpendicular to the plane xoy. The inclined plane actually causes such a visual perception; that is, the ox-axis divides the entire plane into two parts: the sky and the ground. This means that, in scene mode, the imaging plane (or retina) of biological vision corresponds to the xoy plane here and is also perpendicular to the optical axis, and the optical axis is coincident with the depth axis (z-axis). However, in picture mode, the z-axis or $\hat{z}$ -axis cannot be perpendicular to plane xoy. When looking at the visual image, the optical axis of vision may be perpendicular to the two-dimensional image plane, which is perpendicular to plane xoy; with the change of gaze angle, the optical axis of vision is no longer perpendicular to the image plane. In addition, to further demonstrate the important role of the vanishing point and vanishing lines to form a three-dimensional visual perception on the plane, please note the z-axis and $\hat{z}$ -axis in Figure 2.

From the picture in Figure 1C, it can be seen that the orientation of the optical axis of the camera points to the distant vanishing point, but it is very close to the orientation of the $z$ -axis of Figure 2, rather than the $\hat{z}$ -axis; the ground is clearly tilted in the direction toward the vanishing point. For this reason, there is an effect of upward-sloping for the visual perception, and the sense of reality of a three-dimensional scene in pictures is weakened. However, in an actual scene, when looking at a straight road leading into the distance, this effect is much less obvious. If the direction of the camera's optical axis is consistent with the $\hat{z}$ -axis in Figure 2, it will greatly reduce the effects of upward-sloping and increase the sense of reality of three-dimensional visual perception, which also confirms that the interpretations of Figures 1 and 2 made in this paper are reasonable.

Incidentally, the image-forming process of the camera's optical system and the imaging process of optical system of the human vision have some similarities (we shall not consider the differences between them here in detail), and therefore, the expression of photographs and paintings is based on method of Figures 1 and 2.

### *2.2. Important Role of the Vanishing Point in 3D Space and on the 2D Plane*

Why can the technique shown in Figure 1 express three-dimensional structure on the two-dimensional plane? The presence of the vanishing point (infinity point) of visual perception is an important reason. In particular, it is closely related to the formation of psychological coordinates, which in the past did not attract people's attention.

When watching parallel railway tracks fade into the distance, the human visual system perceives the intersection of the parallel lines at a faraway point, which is determined by the basic optical characteristics of the visual system [27-31]. There is no corresponding infinity point in the Cartesian rectangular coordinate system; however, it is possible to add a new coordinate point ($a$) in homogeneous coordinates, thus establishing a mapping between the Cartesian coordinate system $R^n$ and the affine coordinate system $P^n$:

$$R^n \to P^n:$$

$$\begin{aligned} &(x_1, x_2, \cdots, x_n)^{\mathrm{T}} \to (x_1, x_2, \cdots, x_n, 1)^{\mathrm{T}} \\ &(x_1, x_2, \cdots, x_n, 0)^{\mathrm{T}} \to (x_1/a, x_2/a, \cdots, x_n/a, 1)^{\mathrm{T}}, a \to 0 \end{aligned} \tag{3}$$

The infinity point $(x_1, x_2, \cdots, x_n, 0)^{\mathrm{T}}$ is just the limiting case of $(x_1/a, x_2/a, \cdots, x_n/a, 1)^{\mathrm{T}}$ when $a \to 0$, corresponding to the intersection of parallel railway lines at a

point in the distance [1, 4, 7, 32, 33]; the infinity point mapped onto the retina (the imaging plane) will form a visual image as shown in Figure 3.

The optical axis of vision points to a distant focus, that is, the fixation point. Straight parallel lines converge at the focus. The focus and its vanishing line project onto the retina, i.e., the imaging plane. Through the vanishing line and vanishing point in the retina, the visual system can perceive a distant intersection in scenes from the outside world. Figure 2 shows the Cartesian coordinate system, in which the z-axis is consistent with the optical axis of vision. Of particular note is that the parallel lines from infinity are focused on the retina by means of the lens of vision, which is similar to watching parallel railroad tracks converging at one point in the distance, so that the vanishing point is closely related to the parallel lines by affine transformation and its homogeneous coordinates. The relationship between the Cartesian coordinate system, projective coordinate system and affine coordinate system is shown in Figure 3.

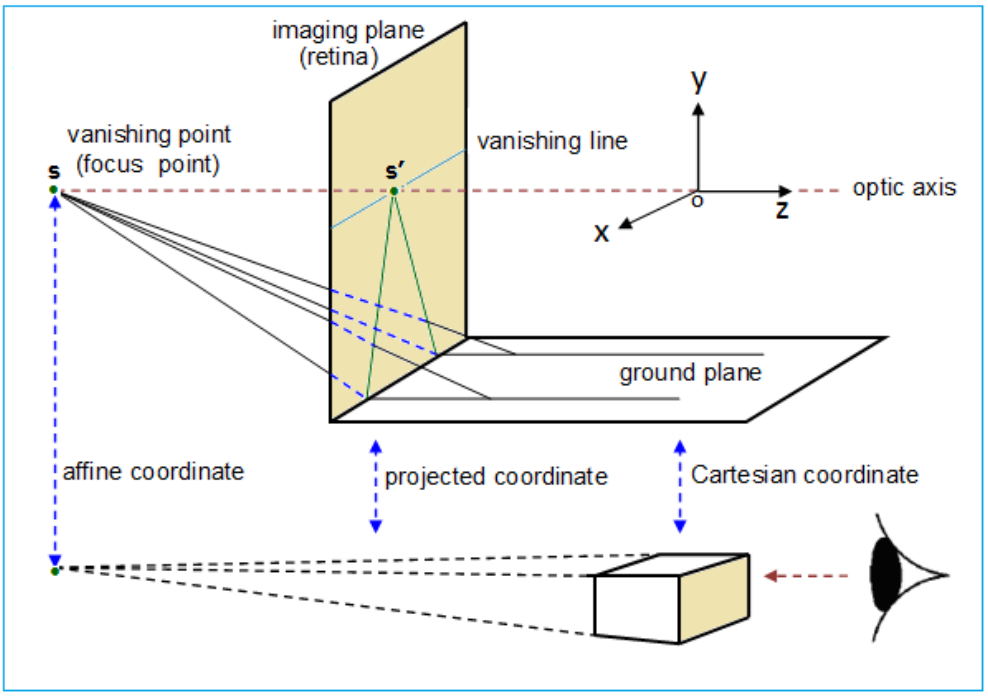


***Figure 3.*** *Optics model of the affine transformation of parallel lines implemented by vision*

Worthy of particular note is equation (3), which contains the conversion between the Cartesian and affine coordinate systems, in which the key is the limit $a \to 0$, i.e., when the distance between an observer and his/her fixation point or spatial range of visual gaze is very small, and the Cartesian coordinate system plays a major role. As $\alpha \to 0$, or the fixation point goes into the distance, an affine coordinate system instead of a Cartesian coordinate system comes into play, and parallel lines gradually converge to a point that is simply the vanishing point (Figure3).

We know that the affine coordinate system is introduced mainly to reflect the structure of the human retina and to describe the features of vision's optical system (that is, the vanishing point and size constancy). The process of viewing an actual scene via human vision is essentially (linear) perspective projection from 3D visual space $P^3$ $(x, y, z)$ to 2D retinal imaging plane $P^2$ $(fx/z, fy/z, f)$, where the focal length $f$ is the distance (about 18–22 mm) between the imaging plane and the optical center (the projection center). It is also an indirect measure of the $z$-axis, because the scaling factor for the projection is $f/z$; as $z$ changes, the scaling factor also changes, and the cross sections, which are in different points of the $z$-axis, are just the samples of the visual image at different depths of field (or at different distances, see Figure 7 for details). A visual image on the retina is formed by corresponding projection lines from all points in a cross section projected onto the same point of the imaging plane (while all points situated on the same projection line means that the "affine coordinate is multiplied by a non-zero constant, its projection unchanged"; this is the root of all problems of 3D reconstruction and its physical meaning). However, different points on a different cross-section will be projected along different lines, thereby forming the depth of field and depth of focus, that is, different samples $S_k$ of a different scene corresponding to the different objective distance $l_k$ in Figure 7 can be formed. Then, the depth of field of the actual scene can be expressed using the following projection equation of an

affine coordinate system, where $\mathbf{M} = (x, y, z, 1)$ is a point in 3D space, $\mathbf{m} = (fx / z, fy / z, 1)$ is a point on the2D plane, $P$ is the projection matrix, and the projection equation has the following form:

$$\begin{bmatrix} fx \\ fy \\ z_k \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z_k \\ 1 \end{bmatrix} \quad \text{or} \quad z_k \mathbf{m} = P^3 \mathbf{M} \qquad (4)$$

If $z_k$ is the depth of field, it is also the distance from the focal plane (or the principal plane) to the point $\mathbf{M}$ (or an object at distance $L$ in Figure 9, or $l$ of equation (20)).

For the "Picture" mode, the human visual system takes input from the picture on the two-dimensional plane, and must therefore perform a projective transformation from the 2D image to the 2D retina. This is the mapping of the pictures to the retina. Here, the projection equation is simplified to the following form:

$$\begin{bmatrix} fx \\ fy \end{bmatrix} = \begin{bmatrix} f & 0 \\ 0 & f \end{bmatrix} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix}$$

or

$$\hat{\mathbf{z}}\mathbf{m} = P^2 \hat{\mathbf{M}} \qquad (5)$$

For the imaging plane, the difference between z and $\hat{\mathbf{z}}$ is caused by the included angle $\beta$ between them, as shown in Figures 1 and 2. Obviously, equations (4) and (5) already show one of the important differences between scene and picture modes.

As an example of the application of the vanishing point, the inversing of a Necker cube, which is a known problem in stereoscopic perception, can be explained by the alternating of a Cartesian coordinate system and an affine coordinate system. The Necker cube has a constant perspective angle; i.e., each of the four sides of a Necker cube (see Figure 4) in the vertical direction, horizontal direction and tilt direction are parallel to each other.

There seems to be no vanishing point in Figure 4. In fact, each of the four parallel sides extend to infinity in the left and right, up and down, and forward and backward directions. The parallel sides converge together and inevitably form vanishing points, all of which form a closed circle. This closed circle is the vanishing line. The circular vanishing line is the fundamental reason why human visual perception can invert opposite sides for the front and back of the cube in Figure 4. In the actual scene, visual perception does not need the condition "under the limit case $a \to 0$" to perceive the emergence of the vanishing point; in human vision, parallel railway lines usually seem to converge at a point at a certain distance, but not at an infinite distance. This situation provides practical possibilities and specific effective ways that the 3D depth of an image can be displayed on a 2D plane.

In computer vision, the vanishing point is called the infinity point. The vanishing point has an important role in implementing the three-dimensional Cartesian coordinate system in the two-dimensional plane and in expressing stereo scenes in the 2D plane. There are three basic types of the vanishing point in the natural scene, namely, one, two and three vanishing points,

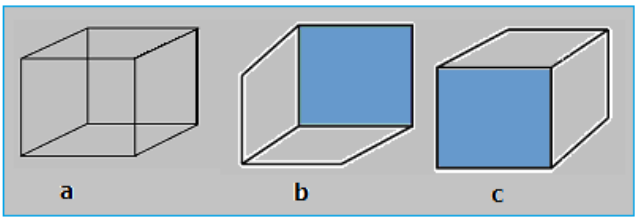


***Figure 4.*** *Inversion phenomenon and three-dimensional visual perception of a Necker cube. (a) Necker cube, which can be perceived as in (b), or as in (c)*

respectively [4,5,7]. In a variety of buildings and real landscapes, both in photography and during actual viewing under various visual angles, these types of vanishing points can be found.

The coordinate axes (x, y, $\hat{z}$ ) of the three-dimensional Cartesian coordinate system in a plane. as can be seen from Figure 1, all three kinds of Cartesian coordinate systems that can be drawn in a plane (that is, from which stereoscopic visual perception can be obtained), without exception, have a mapping with the vanishing point. In other words, the existence of the vanishing point is a necessary condition for drawing up a three-dimensional Cartesian coordinate system in the plane.

From Figure1(c), because of visual perception effects, the axes *x*, *y* and $\hat{z}$ appear to be orthogonal to each other; they are in fact unable to be orthogonal in the two-dimensional plane. In the manner shown in Figure 1, there is an arbitrary angle $\beta$ between the $\hat{z}$ -axis and the x–y plane or between axes x, y and $\hat{z}$ ; however, $\beta \neq 0^o$ (or $\beta \neq 180^o$ ) and $\beta \neq 90^o$ (or $\beta \neq 270^o$ ).

Thus, in three-dimensional space, the depth axis z is perpendicular to the imaging plane (xoy) of the retina, and the optical axis of vision is consistent with the z-axis; that is, both the optical axis and the z-axis are, at the same time, perpendicular to the imaging plane. However, in the two-dimensional plane, the $\hat{z}$ -axis is not perpendicular to the (xoy) plane (see Figure 1). The optical axis may be perpendicular to the (xoy) plane; however, the optical axis for vision is not consistent with the $\hat{z}$ -axis, which is one of the fundamental differences between actual three-dimensional scenes and stereoscopic images on the two-dimensional plane.

This is an important distinction between the basic characteristics of picture mode and scene mode. This difference indicates that, in scene mode—since the optical axis is consistent with the *z*-axis and both of them are orthogonal to the imaging plane—when the gaze point changes, a new visual image is formed on the retina. For picture mode, with a change of gaze direction, if the included angle between the optical axis and the plane (xoy) is also changed, a new visual image is not formed but visual

perception of the original gazed image may vary greatly, especially the change in stereoscopic visual perception. For example, as long as we are look at the same picture (such as an advertising picture) from a different perspective, we can obtain the effect of a different visual perception of the same picture.

### *2.3. Size Constancy of Visual Perception in 3D Space and on the 2D Plane*

How can the 3D structure of an image be expressed in a 2D plane? In addition to the existence of the vanishing point in visual perception, this problem is also closely related to size constancy. From the viewpoint of visual information processing, the vanishing point is a basic condition for the existence of size constancy. Figure 1 is just a reflection of visual characteristics, that is, in Figure 1 the constraints (namely $\beta \neq 0^o$ or $\beta \neq 90^o$) are consistent with the actual presence of the vanishing point (see also Figure 5).

We know that when observing the outside world, human vision shows size constancy for visual properties. In short, the sensory perception of the observer is that the same object is far away if it is small and near if it is large. Size constancy is closely related to the affine transformation of the vanishing point, which appears in visual properties. Size constancy is determined by the characteristics of the optical system of the visual pathway. We know that the height of an object on the base plane can be used to measure the size constancy of visual perception; also, important depth cues can be calculated using the following equation [6,8]:

$$S = \alpha AD \tag{6}$$

Here, $S$ is the height of the object on the fundamentals, $\alpha$ is the visual angle of the picture being taken, $D$ is the distance from the photographer to the object, (i.e., the depth information), and $A$ is the scaling factor of the retina. Clearly, equation (6) deals with the two-dimensional case, in which $S$ and $D$ form a plane such as the ($x$–$y$) plane in Figure 1, which is perpendicular to the base plane ($\hat{z}$ -$x$). For a 3$D$ scene expressed on a 2$D$ plane, equation (6) must be modified according to equation (1). As a result, the degree of depth perception obtained from the 2$D$ plane is smaller than that from the 3D scene, and the actual reduction in the degree of depth perception can be calculated according to the following equation (see the Method section for an example):

$$S = \alpha\cos(\beta - 90^o)AD \tag{7}$$

Psychophysical experiments have shown that size constancy is roughly constant within a certain distance (for example, tens of meters). If a vanishing point in the visual perception of the outside world does not exist, then naturally size constancy also does not exist. As a result, the image structure of a 3D scene cannot be expressed on a 2D plane according to Figure 1. Because the vanishing point is determined by the optical characteristics of the human visual system, size constancy is an embodiment of this feature; affine transformation is its mathematical description, in which the most valuable cue for visual perception is parallel lines intersecting at infinity, thereby forming a theory of vanishing point.

The example in the Method section shows that to determine the depth of different objects in a picture, size constancy must be combined with a calculation of the vanishing point. In computer vision and computer graphics, these calculations are implemented by corresponding algorithms and proprietary software; however, it is still important to note that in Figure 1, when calculating the included angle between the $\hat{z}$-axis and the ox-axis, both the vanishing point and the horizontal vanishing line must be displayed in the picture (see Figures 1 and 2), which allows the determination of $\beta$ [34-36]. The horizontal vanishing line corresponds to the ox-axis; the $\hat{z}$-axis corresponds to such a straight line, which has the fastest change in depth, and is determined using the fitting method of linear least squares between the linear perspective and the texture gradient in the figure [36].

In the Methods section, we calculate the depth of the image by means of size constancy mainly to illustrate that the depth of visual perception and the calculation results are consistent with each other.

## 3. Mathematical Models

### *3.1. Description of the Ambient Light Field Function*

According to the method given in Figure 1, since we can draw a 3D scene graph in the 2D plane, the next question is: how do we properly describe the three-dimensional structure of an image? What about it is different from the actual scene?

For the actual scene, imagine that at the fixation point, objects reflect light beams to the surrounding area (including radiation of all kinds of lighting sources). At different angles, the light beams pass through the observer's pupil, and then form an image on the retina. In fact, the reflected light from the gaze point to the surrounding environment is called "structured light", which carries various environmental information that can be described using the light intensity $V_x, V_y, V_z$, wavelength $\lambda$, position (x, y, z) in Cartesian coordinates and time $t$:

$$P_W = P_W(x, y, z, \lambda, t, V_x, V_y, V_z) \tag{8}$$

The state function $P_W$ is an objective description of the characteristics of the light field in the actual scene and has nothing to do with the observer. An observer looking at a particular point has their optical axis of vision parallel to the z-axis, and thus the angle $(\theta, \phi)$ of light into the pupil can be calculated. The distance $l_k$ between the retina and fixation point is also completely determined. When the fixation point moves forward or backward along the z-axis, the value of $l_k$ will accordingly change, which can be calculated in terms of the biological structure of the eye and

its optical parameters (e.g., the pupil, cornea, and focal length). By means of size constancy and around various reference objects, human vision can very easily perceive the distance of the fixation point, so that the z-dimension will be removed from equation (8), and a function with seven variables is obtained. The resulting form of the function, in the spherical coordinate frame, can be expressed as follows:

$$P_V = P_V(\theta, \phi, \lambda, t, V_x, V_y, V_z) \tag{9}$$

The z-axis representing the depth is eliminated from equation (9). However, when clearly focusing, $V_z$ is known, which carries information about the distance between the observer and the gaze point of sharp focus. When the fixation point moves forward or backward along the z-axis, the distance between the gaze point and the focal point on the retina changes with it, and the visual depth perception also changes with it; see the schematic in Figure 7. It can be seen as essentially a layered depth image.

The light entering the eye can also be calculated using the coordinates (x, y), where the image plane (that is, the retina) is perpendicular to the optical axis; then, the imaging function $P_R$ on the retina can be expressed as follows:

$$P_R = P_R(x, y, \lambda, t, V_x, V_y, V_z) \tag{10}$$

The image on the retina, $P_R$, is just the mathematical description of the ambient structured light. The basic idea stems from Adelson and Bergen's full-optical (or plenoptic) function concept [37, 38]. It is an idealized concept, used to describe natural scenes that seem too complicated and difficult to deal with. However, in this paper, we use full-optical functions that can better analyze the information structure of the incident light, which carry information about the surrounding environment.

In this case, the input representation for vision is just the full-optical function $P_R$ formed by the visible light-illumine-

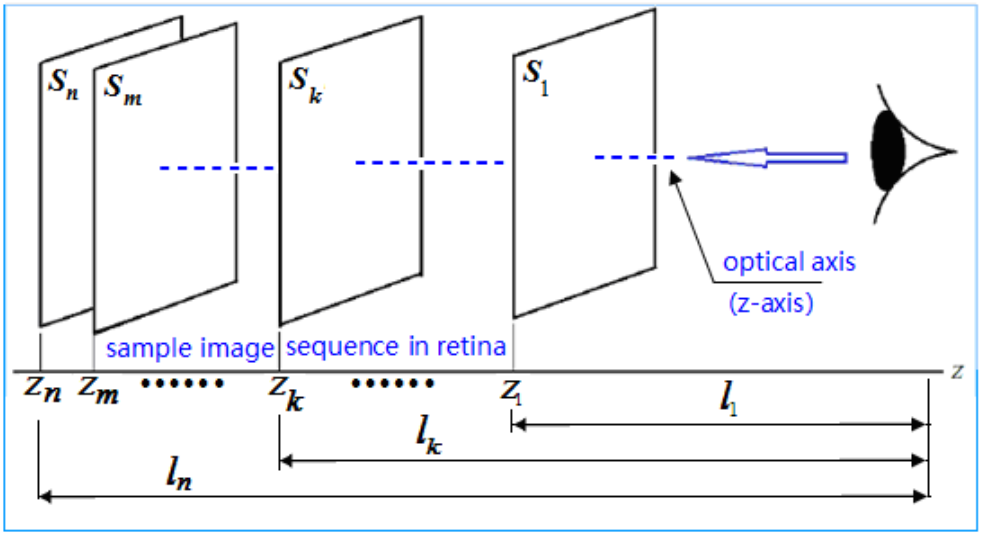


***Figure 5.*** *Visual sample sequence formed in the retina by different fixation points*

-tion in the environment, and the output representation for vision is a complete image $I_R$. In other words, the observer forming a visual image on the retina by looking at the actual scene, resulting in visual depth perception, is the "scene mode". Conversely, seeing the actual scene's photographic images also allows the formation of a visual image on the retina, likewise causing visual depth perception, but in this case using the "picture mode". Now we ask: what is the difference between these two? From the perspective of mathematical transformations, are they both equivalent? We will discuss this issue in the following.

Note that when viewing the actual scene, the observer can see only a sample of the function $P_W$. Similarly, when viewing a picture or painting in a plane, the viewer can see only a sample of the light intensity array of the image. In other words, the image $P_R$ on the retina and the image $I_R$ in the V1 cortex are just a sample of the full-optical function $P_W$ of the ambient light field. Along with the movement of the fixation point, for scene mode, the sample sequence $l_1$, $l_2$, …, $l_k$, …, $l_n$ is formed. For picture mode, only one sample determines when this picture was taken instead of a sample sequence. It can be seen that in scene mode, the sample sequence contains depth information front and rear of the fixation point (see Figure 5); for picture mode, it is fixed. From this, the difference between the two modes is clearly in evidence.

Then, the problem for the visual system is how to get this sample from the full-optical function. This problem is closely related to visual depth perception.

### *3.2. Seeing Mode and Obtaining the Visual Sample*

We know that color information is transmitted by the parvocellular pathway through the lateral geniculate nucleus to the visual cortex V1, in which color, shape, and motion are also separated from each other [39-41]. Therefore, the wavelength $\lambda$ can be separated out from the full-optical function. In addition, according to the structure and function of the retina and the primary visual cortex, the position information is mainly responded to by the photoreceptor cells

(rods and cones) in the retina, while orientation information is mainly processed by simple cells, complex cells and orientation functions in the visual cortex V1 [5].

Therefore, position and orientation information can also be separated. Without considering time-varying information for the present, we can use the variable separation approach to simplify the full-optical function with seven variables: $P_v = P_v(\theta, \phi, \lambda, t, V_x, V_y, V_z)$, and also process the visual information.

### *3.2.1. Scene Mode*

When the observer's eye gazes at a point in a scene, the reflected light beam from the point (x, y, z) carries information to the eye about the intensity $V_x$, $V_y$ and $V_y$. The light axis is consistent with the z-axis, and the light intensity (as a stimulus signal for vision) excites photosensitive cells in the retina into firing, so that the intensity of the light stimulus is transformed into the activity strength of a photosensitive cell. Therefore, we only need to record the position and orientation θ, $\varphi$ of the light emitted from the point (x, y). For this reason, full-optical functions can be expressed using a dual-plane parameterized method, which can be described using the intersection coordinates $(\theta, \varphi)$ and (x, y) of a light and space plane (retina) $P(x, y)$ as well as an angled plane (the visual cortex) $P(\theta, \varphi)$, as shown in Figure 6 [42, 43].

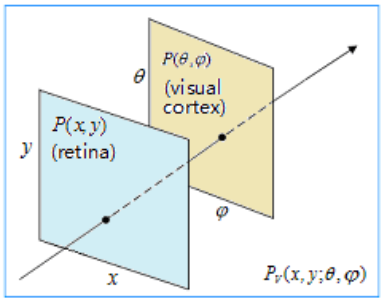


***Figure 6.*** *Nested expression of dual-plane parametric representation*

The visual image can be expressed mathematically as the Kronecker inner product of the retina $P(x, y)$ and visual cortex $P(\theta, \varphi)$, namely: $I_R \Rightarrow P(x, y) \otimes P(\theta, \varphi)$, that is, through the dual-parameterized methods of the plenoptic functions, we can easily understand the information processing functions of the retina $P(x, y)$ and the visual cortex $P(\theta, \varphi)$ in the visual pathways. If each primitive of visual image is denoted by $R_{u,v}(a)$, then the visual image $I_R$ can be expressed as the following matrix:

$$[R_{u,v}(a)]_{U\times V} = \begin{bmatrix} r_{1,1}(a) & r_{1,2}(a) & \cdots & r_{1,V}(a) \\ r_{2,1}(a) & r_{2,2}(a) & \cdots & r_{2,V}(a) \\ \vdots & \vdots & \vdots & \vdots \\ r_{U,1}(a) & r_{U,2}(a) & \cdots & r_{U,V}(a) \end{bmatrix}$$

$$u = 1, 2, \cdots, U, \quad v = 1, 2, \cdots, V \tag{11}$$

Similarly, the receptive fields of a functional column in primary visual cortex V1 can be expressed as the following matrix:

$$[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} = \begin{bmatrix} b_{1,1}(g) & b_{1,2}(g) & \cdots & b_{1,\Phi}(g) \\ b_{2,1}(g) & b_{2,2}(g) & \cdots & b_{2,\Phi}(g) \\ \vdots & \vdots & \vdots & \vdots \\ b_{\Theta,1}(g) & b_{\Theta,2}(g) & \cdots & b_{\Theta,\Phi}(g) \end{bmatrix}$$

$$\theta = 1, 2, \cdots, \Theta, \quad \varphi = 1, 2, \cdots, \Phi \tag{12}$$

According to the mapping $I_R \Rightarrow P(x, y) \otimes P(\theta, \varphi)$, the neural computation of cortex V1 to the retinal image is a coincidence operation. That is, each primitive retinal image (i.e., line segments, corners and other features) is detected by receptive fields of various bandwidths and orientations of simple and complex cells in V1 cortical columns in the manner of a compliance operation, which is the Kronecker product operation.

$$\begin{bmatrix} r_{1,1}(a)[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} & r_{1,2}(a)[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} & \cdots & r_{1,V}(a)[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} \\ r_{2,1}(a)[B_{=\theta,\varphi}(g)]_{\Theta\times\Phi} & r_{2,2}(a)[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} & \cdots & r_{2,V}(a)[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} \\ \vdots & \vdots & \vdots & \vdots \\ r_{U,1}(a)[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} & r_{U,2}(a)[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} & \cdots & r_{U,V}(a)[B_{\theta,\varphi}(g)]_{\Theta\times\Phi} \end{bmatrix}\Big|_{\max}$$

$$= \begin{bmatrix} r_{1,1}(a) & r_{1,2}(a) & \cdots & r_{1,V}(a) \\ r_{2,1}(a) & r_{2,2}(a) & \cdots & r_{2,V}(a) \\ \vdots & \vdots & \vdots & \vdots \\ r_{U,1}(a) & r_{U,2}(a) & \cdots & r_{U,V}(a) \end{bmatrix} \otimes \begin{bmatrix} b_{1,1}(g) & b_{1,2}(g) & \cdots & b_{1,\Phi}(g) \\ b_{2,1}(g) & b_{2,2}(g) & \cdots & b_{2,\Phi}(g) \\ \vdots & \vdots & \vdots & \vdots \\ b_{\Theta,1}(g) & b_{\Theta,2}(g) & \cdots & b_{\Theta,\Phi}(g) \end{bmatrix}\Big|_{\max} \tag{13}$$

**Science Research**
2014; 2(5): 135-149
Published online November 20, 2014 (http://www.sciencepublishinggroup.com/j/sr)
doi: 10.11648/j.sr.20140205.20
ISSN: 2329-0935 (Print); ISSN: 2329-0927 (Online)

Obviously, we can obtain the following mappings:

$$I_R \Rightarrow P(x,y)\otimes P(\theta,\varphi) \rightleftharpoons [R_{u,v}(a)]_{U\times V} \otimes [B_{\theta,\varphi}(g)]_{\Theta\times\Phi} \Rightarrow I_{V1} \quad (14)$$

where $I_{V1}$ is the image in V1. The above distributed parallel computing of $[R_{u,v}(a)]_{U\times V} \otimes [B_{\theta,\varphi}(g)]_{\Theta\times\Phi}$, from the retinal image $I_R$ to the cortical image $I_{V1}$, can be achieved within milliseconds of time, which is consistent with recent psychophysical experiments [44].

Light with different angles should show up as different perspectives or visual angles in the image, so that such a method of representing light field data can be associated with a neural representation of the human retina in primary visual cortex. A large number of neurobiological experiments have demonstrated that there is a one-to-one mapping between the retina and the visual cortex, and topological connections have been established through ganglion cells. In the retina, photosensitive cells record position and intensity information of incident light; V1, through simple cells, complex cells and function columns, deals with direction or orientation information [45-49].

The visual system can process all-optical functions in a similar manner as dual-plane parameterization, indicating that the optical information to the outside world can be aptly described by an all-optical function. After separation of a visual input signal by the retina and the primary visual cortex, two-parameter processing is achieved, essentially converting an all-optical function $P_{\mathrm{w}}(x,y,z;V_x,V_y,V_z;\lambda,t)$ to a light field function $P_R(x,y,\theta,\varphi)\big|_{z_k}$ as shown in the following equation:

$$P_{\mathrm{w}}(x,y,z;V_x,V_y,V_z;\lambda,t) \rightarrow P_R(x,y,\theta,\varphi)\big|_{z_k} \quad (15)$$

This combination of position, intensity, direction and orientation is just a neural representation of $I_R$ in V1; that is, a sample of the actual visual scene obtained by vision. Viewing the actual scene or seeing a stereoscopic image or drawing of the scene expressed in the 2D plane according to Figure 1 can be regarded as a sample of visual perception.

That is, the input signal to vision is $P_{\mathbf{w}}(x,y,z;V_x,V_y,V_z;\lambda,t)$, and the input signal to V1 is $P_R(x,y,\theta,\varphi)\big|_{z_k}$ for the scene mode.

#### *3.2.2. Picture Mode*

As already noted, when vision is clearly focused, $V_{z_k}$ is known and carries information about the distance between the observer and the sharply focused fixation point. In this sense, looking at the actual scene and looking at a photographic image of the scene will produce the same effect in three-dimensional perception. However, the photographic pictures are presented in the manner of Figure 1, the input representation is $P_I(x,y,V_x,V_y;\lambda)$, compared with scene mode, and there is no information related to the angle $(\theta,\varphi)$ of incident light.

That is, the input signal to vision is $P_I(x,y,V_x,V_y;\lambda)$, and the input signal to V1 is $P_{V1}(x,y)$ for picture mode.

Now, the reduction of the effect of depth perception caused by angle $\beta$ (included angle between the z -axis and the $\hat{\mathrm{z}}$ -axis in Figure 1) and the change of intensity $V_{z_k}$ should also be considered; here, $V_{z_k}$ is incident light intensity in 3D physical space. For this reason, if the stereoscopic image and the drawing are expressed according to the method in Figure 1, then their intensity information relative to $V_{z_k}$ has to be corrected. Let $V_{z_k-2D}$ denote the corrected value of $V_{z_k}$ in the Cartesian coordinate system introduced into the 2D plane, as shown in Figure 1.

$$V_{z_k-2D} = \frac{V_{z_k}}{\cos(\beta-90^{\mathrm{o}})} \quad (16)$$

Human vision is not sensitive to slight changes in visual depth perception; generally, it is difficult to distinguish the difference. Of course, whether the human visual system adopts this strategy to obtain sample information requires further in-depth study.

### *3.3. Mapping Relationships between Scenes, Pictures and Visual Images*

An actual three-dimensional (3D) scene can be described using full-optical functions, as shown:

$$I_{3D} = P_{\mathrm{w}}(x,y,z;V_x,V_y,V_z;\lambda,t) \quad (17)$$

As described above, in the psychological coordinate system drawn on a two-dimensional plane, a scene picture, shown in accordance with the method in Figure 1, can be expressed as

$$I_{p-3D} = P_{\mathrm{v}}(x,y,\hat{z};V_x,V_y,V_{\hat{z}};\lambda,t) \quad (18)$$

Here, $I_{p\text{-}3D}$ represents the stereoscopic image of the psychological coordinate system in the plane. According to the physical dimensions of the object being looked at, this paper introduces two viewing modes, “scene mode” and “picture mode”. The former relates to the viewing of the actual scene, and its physical dimensions are 3D; the latter relates to viewing the image on a plane whose physical dimensions are 2D. However, whether when looking at the actual scene, or three-dimensional images on the plane, we can perceive the depth structure of the image, thereby causing stereoscopic visual perception.

This difference between the two modes lies in two aspects. There is the difference in cognition—for example, when viewing the actual scene, the viewer is in the actual scene, there are differences in aspects of visual perception, and the viewer feels immersed in the reality of the scene; whereas while looking at pictures of a scene, the viewer and the scene are separated, with a lack of reality. There is also a difference in terms of depth perception. From equations (1) and (15) we can see that the difference between "scene mode" and " picture mode" is actually the difference between the $\hat{z}$ and the z, namely: $z=\hat{z}\cos(\beta-90^{o})$. This equation shows that depth perception in the "picture mode" of stereoscopic perception is clearly lower than in the "scene mode" of three-dimensional perception. However, for the human visual system, it is difficult to distinguish quantitative depth differences between these two modes (absolute depth information), but fairly straightforward to process and estimate the kind of information, such as front and back, right and left, up and down, size, and distance and other location information of objects in scenes and pictures (all this belongs to relative depth information). When pictures are dynamic (such as TV), analysis, judgment and direction of movement of objects are more important.

Vision is a causal system, that is, an observer viewing a three-dimensional scene $I_{3D}$ of the outside world will form a visual image $\psi_{r(p-3D)}$ on their retina; and vice versa, when forming a visual image $\psi_{r(3D)}$ on the retina, the scene of the external world will be perceived to be similar to $I_{3D}$. Likewise, if the image viewed is $I_{p\text{-}3D}$, the retina will form a visual image $\psi_{r(p-3D)}$, and if the retina contains the visual image $\psi_{r(p-3D)}$, the observer will perceive the scene in the external world as $I_{p\text{-}3D}$. According to the optical characteristics of the visual system, the object point and the image point can be interchanged. Therefore, both $I_{3D}$ and $\psi_{r(3D)}$ are mutually conjugate mappings, as are $I_{p\text{-}3D}$ and $\psi_{r(p-3D)}$, and vice versa:

$$I_{3D} \rightleftharpoons \psi_{r(3D)}$$

$$I_{p-3D} \rightleftharpoons \psi_{r(p-3D)}$$

From this, we can draw an important mapping:

$$P_{w}(x,y,z;V_x,V_y,V_z;\lambda,t) \leftrightarrow P_{v}(x,y,\hat{z};V_x,V_y,V_z;\lambda,t) \quad (19)$$

The reason for the above relationship is that the object and its picture have the same dimensions in terms of cognitive psychology; that is, the visual image $\psi_{r(3D)}$ is formed on the retina by the full-optical functions $I_{3D} = P_{w}(x,y,z;V_x,V_y,V_z;\lambda,t)$. Since the physical dimensions of the retina are 2D, the retinal image must be formed in the manner shown in Figure 1, which is a reflection of the optical characteristics of the visual system. Thus, we have reason to believe that whether using "scene mode" or "picture mode"; i.e., whether visual images on the retina are represented using $\psi_{r(3D)}$ or $\psi_{r(p-3D)}$, their display mode is consistent with the full-optical function $I_{p\text{-}3D}$, complying without exception with the pattern in Figure 1. This is also the basic meaning in cognitive psychology of equation (19).

Because the retina is two-dimensional, the visual image on the retina will be expressed in the manner of Figure 1, that is, instead of the coordinate system $(x,y,z)$ by the coordinate system $(x,y,\hat{z})$. Regardless of whether the dimensions of the visual input representation are three-dimensional or two-dimensional, visual images on the retina are two-dimensional. At first glance, the visual images of these two modes on the retina are exactly the same in terms of dimensions, which appears to be a contradictory result. We know whether we are looking at an actual scene or a photograph, even though both of them are, essentially, (linear) projection transformations from 3D visual space $P^3$ to the 2D imaging plane $P^2$ of the retina. However, in scene mode, the transformation is $z_k\mathbf{m} = P^3\mathbf{M}$, while in picture mode, it is $\hat{z}\mathbf{m} = P^2\mathbf{M}$; i.e., the points in the 2D plane are projected onto the retina, which is $m_{picture} \rightarrow m_{retina}$. The loss of depth information of the image $\psi_{r(p-3D)}$ with respect to $\psi_{r(3D)}$ is $\cos(\beta-90^{o})$. As for the output representation of V1, its situation is similar to that mentioned above, so we will not repeat it.

For the two modes, the above mentioned is another difference obtained from the viewpoint of affine transform.

Notably, in scene mode, $z_k$ means that the fixation point is moved back and forth in the z-axis in the gaze direction, to form a sample sequence shown in Figure 5, thereby continuously changing depth can be perceived. However, in picture mode, when shooting pictures, the focal length and object distance have been specified, and can no longer be changed with different fixation points. Therefore, in scene mode, depth perception includes more information than the in picture mode in the z-axial direction, i.e., sequence $z_k$.

This study shows the following differences in the basic characteristics between "scene mode" and "picture mode". The optical axis of the former is consistent with the z-axis of the Cartesian coordinate system, the retinal imaging plane is perpendicular to the optical axis and the z-axis, and the imaging plane of the latter may be perpendicular to the optical axis. However, the z-axis of the psychological coordinate system on the plane is not consistent with the optical axis.

### *3.4. Depth of Field of Retina Pattern*

When the human visual system perceives an external scene,

the optical axis is consistent with the *z*-axis. The imaging plane is perpendicular to the optical axis and the *z*-axis, which is an inherent optical characteristic of the visual imaging system. A very small distance fore-and-aft in the focal plane (visual image on the retina), in which the diameter of the light spot (circle of confusion) is small (0.005 mm) [50], still forms a clear visual image; this very small distance is known as depth of focus (see Figure 7). The conjugate relationship between the image point and the object point for the objects being viewed shows the same situation. In front of and behind the object at a certain distance, when the light spot formed is small, the near point and far point together form the depth of field.

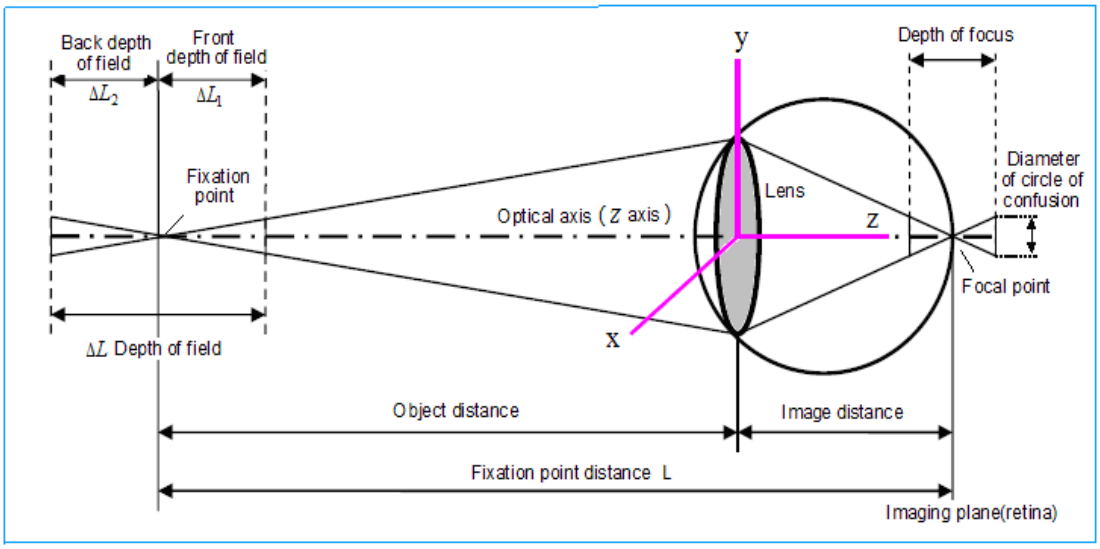


***Figure 7.*** *Schematic diagram of depth of field and depth of focus*

It is through the optics of the human visual system that depth of field can be perceived in the external world. Depth of field on the retina is determined by $\delta$ (permissible circle of confusion diameter); $f$ (lens focal length); $F$ (pupil size); $L$ (focusing distance); $\Delta L_1$ (front depth of field) and $\Delta L_2$ (back depth of field). The equation for calculating the depth of field is as follows:

$$\Delta L_1 = \frac{F\delta L^2}{f^2 + F\delta L}, \quad \Delta L_2 = \frac{F\delta L^2}{f^2 - F\delta L},$$

$$\Delta L = \Delta L_1 + \Delta L_2 \Delta L = \frac{2f^2 F\delta L^2}{f^4 - F^2\delta^2 L^2} \tag{20}$$

The scene image can be reconstructed based on $\Delta L_1$ (front depth of field) and $\Delta L_2$ (back depth of field). The visual image on the retina does not lose depth information because the three-dimensional scene forms a visual image on the two-dimensional retina. The main reason is that the optical axis is consistent with the *z*-axis. Thus, the distance $z_k$ along the *z*-axis can be used instead of the focusing distance $L$ (see Figures 3 and 5), which makes equation (20) describing the optical characteristics of the human visual system easier to understand. Therefore, equation (20) can be rewritten as:

$$\Delta L = \Delta L_1 + \Delta L_2 \Delta L = \frac{2f^2 F\delta l_k^{\,2}}{f^4 - F^2\delta^2 l_k^{\,2}} \tag{21}$$

Here, the $l_k$ represents the distance along the z-axis from the object being observed to the retina when in focus. Simply put, the distance between the observer and the object reflects the scene depth information of the sample image ( $l_1$, $l_2$, …, $l_k$, …, $l_n$ ) at the focal point (see Figure 5). It is thus clear that making the optical axis consistent with the *z*-axis is a very effective constraint; it is not artificially imposed, but is determined by the optical characteristics of the visual system.

Sometimes, the maximum depth of field can also be expressed with the diameter $d$ of a photoreceptor (photosensitive cell) in the retina, the pupil diameter $D$ of the eye, the image distance $l'$ and the object distance $l$, i.e.:

$$\Delta L = 2\frac{l^2}{l'^2}\cdot\frac{\delta}{D} \approx 2\frac{l^2}{l'^2}\cdot\frac{d}{D} \tag{22}$$

After obtaining the equation for vision depth (19), we can make a comparison between the two modes, namely "scene mode" and "picture mode", and compare the depth of field of an actual scene with one of an image. In equation (21), $\Delta L$ represents the depth of field perceived by the viewer when looking at the actual scene. An observer looking at a photograph of the same actual scene sees the picture displayed according to Figure 1; that is, the stereoscopic picture in the psychological coordinate system, and its depth of field $\Delta\hat{L}$, needs to be corrected using $\Delta L$ of equation (1):

$$\Delta\hat{L} = \Delta L \cos(\theta - 90^{\circ}) \quad (23)$$

# 4. Conclusions and Discussion

In this paper, the problems studied and discussed are very old. According to historical records, in 430 AD, Chinese painter Zong Bing realized that the same object appears smaller when far away and larger when near to an observer. On these grounds, he proposed the perspective method and applied it in his paintings. In the 15th century, during the Renaissance, Leonardo da Vinci also studied the same perspective principle, which is notable in his famous painting "The Last Supper". This issue is now a hot topic in the field of information processing. To better express stereoscopic animated scenes on a plane, this problem is also now under investigation in the field of computer graphics [51]. In computer vision, this issue is studied to determine corresponding points from dual camera pictures to reconstruct the three-dimensional structure of the scene itself [52]. In his work, Marr mainly explores how the neural mechanisms of cognitive psychology work in the perception of three-dimensional structure from a two-dimensional image [2,3]. Street and wall painters comprehensively use perspective, texture gradients, shading, shadows, and color and occlusion methods, so scene drawing on the plane has a very vivid hierarchy [5]. Comparatively speaking, however, image rendering is at a higher level [23,42,43].

However, we can only know the how and not the why. The intent of this paper is to explore the primary information processing mechanisms that the visual system uses to perceive depth information from stereoscopic scenes on the two-dimensional plane. When excluding factors of atmospheric perspective, texture gradient, image hierarchy, shading, shadow, color and occlusion and binocular disparity, the problem can be naturally attributed to the following problem: Why can we introduce three-dimensional Cartesian coordinates onto the two-dimensional plane to express the three-dimensional structure of the image (see Figure 1)? Why can we draw a line cube (Necker cube, see Figure 4) on the plane? Why can we draw a winding curve into the distance, which can be perceived as a river? What is the basis for this in cognitive psychology?

First, we use the basic situation in Figure 1 as a starting point for research. We point out that the manner of representing 3D objects in Figure 1 is the only way that three-dimensional Cartesian rectangular coordinates can be introduced into a two-dimensional plane to express the three-dimensional structure of the image, and give the calculation formula by which we can evaluate the loss of depth information from a 3D scene. This is followed by in-depth analysis of size constancy in the 3D scene and the 2D image. Differences in vanishing point between the 3D scene and the 2D image also show that vanishing point and size constancy are the basis of the way that the visual system perceives the outside world, as well as introducing the 3D Cartesian coordinate system into the 2D plane. We further study the role of perceptual constancy and vanishing point in forming a visual image, which reveals a corresponding mapping between the object point and image point in the optics of the visual system. In particular, we propose a neural computational method showing how to separate a full-optical function by the retina and visual cortex, called the dual-parametric approach. This paper also provides a method of obtaining an output sample of visual perception, giving a calculation and measurement of size constancy to show its important role in stereoscopic visual perception and cognitive processing.

Second, in this paper, the problem discussed initially seems simple and clear, but answering it presents unexpected difficulties and complexity. The reader may also to try to answer this question, to see whether there is a clear result. We place so much emphasis on this issue, pointing out its features and complexity, because the study of this problem for stereoscopic visual perception is important and the results have much practical application value for computational vision.

To summarize, the main differences in visual depth perception between scene mode and picture mode consist of the following three points:

1. For the former, the depth axis (z-axis) of a Cartesian coordinate system is consistent with the optical axis of vision, and the imaging plane (or retina) is perpendicular to the optical axis and thus perpendicular to the z-axis. For the latter, because of the different viewing angle, the imaging plane is either perpendicular to the optical axis, or cannot be perpendicular to the optical axis. However, the z-axis (denoted by the symbol $\hat{z}$ in the text) of the three-dimensional Cartesian coordinate system expressed on the plane is not consistent with the optical axis of the image, thus causing a loss of depth information. This paper presents a method to estimate the amount of such loss.
2. In scene mode, the visual input signal is a seven-dimensional plenoptic function, and in picture mode, the visual input representation is the light intensity array, i.e., a four-dimensional function of the light field [23], which is obtained by simplifying and separating the seven-dimensional all-optical function. This distinction is very important for the study of visual depth perception for these two modes.
3. In scene mode, the distance of the fixation point is variable; that is, the vanishing point is able to change with a changing fixation point. Also, in picture mode, the vanishing point is fixed, because when shooting the picture, its vanishing point is determined by the optical system of the camera.

In addition, we provide a method to obtain a sample sequence of the output representation of visual perception, and the relationship between the sample and the depth information [26], which relate the size constancy and vanishing point with each other. We point out that both of these have an important role in imaging of the optical system in vision, thus preliminarily explaining why visual images on the human retina contain two-dimensional depth information.

One of our aims is to answer Marr's question, and we believe that the comprehensive results of these studies will give a preliminary answer to Marr's question; of course, it is

also an issue of great concern to many biological vision researchers.

The following table compares the main characteristics of two depth perception patterns; relevant research results have been discussed in detail in the above sections.

***Table 1.*** *Comparison of visual depth perception between scene mode and picture mode*

| Item compared | Scene mode | Picture mode |
|---|---|---|
| Coordinate system | $(x, y, z)$ | $(x, y, \hat{z})$ |
| Visual depth perception | $z$ | $z_{\perp} = \hat{z}\cos(\beta\text{-}90^{\circ})$ |
| Input representation of vision | Plenoptic function | Array of light intensity |
| Mathematic expression | $P_W = P_W(x, y, z, \lambda, t, V_x, V_y, V_z)$ | $P_R = P_R(x, y, \lambda, V_x, V_y)$ |
| Image of cortex V1 | $P_{V1} = P_{V1}(x, y, \theta, \varphi)$ | $P_{v1} = P_{v1}(x, y)$ |
| Projection equation | $z_k \mathbf{m} = \mathrm{p}^3\mathbf{M}$ | $\hat{z}\mathbf{m} = \mathrm{p}^2\mathbf{M}$ |
| Affine coordinate | $\begin{bmatrix} fx \\ fy \\ z_k \end{bmatrix} = \begin{bmatrix} f & 0 & 0 & 0 \\ 0 & f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} x \\ y \\ z_k \\ 1 \end{bmatrix}$ | $\begin{bmatrix} fx \\ fy \end{bmatrix} = \begin{bmatrix} f & 0 \\ 0 & f \end{bmatrix} \begin{bmatrix} x \\ y \end{bmatrix}$ |
| Property of geometric optics | Optic axis is perpendicular to image plane | Optic axis is not perpendicular to image plane |
| Image plane | Retina or xoy-plane | Retina or xoy-plane |
| Image features | Vanishing point change with fixation point | Vanishing point has been fixed |
| Open problem | | |
| how to measure the image | $P_{V1} = P_{V1}(x, y, \theta, \varphi)$ | $P_{v1} = P_{v1}(x, y)$ |

## Acknowledgments

This work is supported by the Natural Science Foundation of China (grant number: 61271425). The authors would like to thank Li Shuzhong and Sung Guangyu for giving their permission for us to use the photograph shown in Figures 1(c).

## Authors' Contributions

Zhao Songnian conceived the study and wrote the first draft. Yu Yunxian , Zhao Yunping and Cheng Wenjun took part in designing the study and contributed to the comparative analysis. Zhao Yunping and Jin Xi took part in the numerical calculations, verification and analysis of the data and drew all the illustrations. All authors discussed and modified the revised manuscript, and all authors have accepted the final version.

## Conflict of Interest

The authors declare that they have no conflicting interests.

## References

[1] M Sonka, V Havac, and R Boyle. Image processing, analysis, and machine vision, Second edition, Thomson Learning and PT Press, 310-321, 1999

[2] D Marr. Vision: A computational investigation into the human representation and processing of visual information. New York: Freeman, 1982

[3] K A Stevens. The vision of David Marr. *Perception*, 41, 1061-1072, 2012

[4] H A Mallot. Computational vision: information processing in perception and visual behavior, The MIT Press, Cambridge, London, England, 23-46, 2000

[5] J P Frisby and J V Stone. Seeing, The computational approach to biological vision, second edition, The MIT Press London, England, 539-551, 2010

[6] M Hershenson. Visual space perception: A primer, Cambridge, MA: MIT Press, 78-91, 2000

[7] J Herrault. Vision: Images, Signals and Neural Networks, World Scientific publishing Co. Pte. Ltd, 2010

[8] S E Palmer. Vision Science, MIT Press, 186-193, 1999

[9] M Caradini, J B Demb, V Mante, et al., Do we know what the early visual system does? *J. Neurosci*., 25(46), 10577-10579, 2003.

[10] J Kornmeier and M Bach. Object perception: When our brain is impressed but we do not notice it, *Journal of Vision*, 9(1):7, 1-10, 2009

[11] J Kornmeier and M Bach. The Necker cube—an ambiguous figure disambiguated in early visual processing, *Vision Research*, 45, 955–960, 2005

[12] J J Koenderink. Solid shape, MIT press, 1990

[13] J J Koenderink, A. J. VAN Doorn, and A. M. L. Kappers. Surface perception in pictures. *Perception & Psychophysics*. , 52 (5), 487-496, 1992

[14] A J Van Doorn, J J Koenderink, and J Wagemans. Light fields and shape from shading, *International Journal for Numerical Methods in Engineering* , 2011

[15] M S Banks, R T Held, and A R Girshick Perception of 3-D layout in stereo displays. *Information Display* 25, 1, 12–16. 2009.

[16] E A Cooper, E A Piazza, and M S Banks. The perceptual basis of common hotographic practice. *Journal of Vision* 12, 5, 8:1–14. 2012.

[17] M Pharr and G Humphreys. Physically Based Rendering: From Theory to Implementation, 2nd ed. Morgan Kaufmann, 2010.

[18] D Vishwanath, A R Girshick, and M S Banks. Why pictures look right when viewed from the wrong place. Nature Neuroscience 8, 10, 1401–1410. 2005.

[19] F Steinicke, G Bruder, and S Kuhl. Realistic perspective projections for virtual objects and environments. ACM Transactions on Graphics 30, 5, 112: 1–10. 2011.

[20] S T Watt, K Akeley, M O Ernst, and M S Banks. Focus cues affect perceived depth. *Journal of Vision* 5, 10, 7:834–862. 2005.

[21] J Yu, L Mcmillan, and P Sturm. Multi-perspective modeling, rendering and imaging. *Computer Graphics Forum* 29, 1, 227–246. 2010.

[22] S M Banks, J C A Read, R S Allison, and J S Watt. Stereoscopy and the Human Visual System. *SMPTE Motion Imaging Journal* 26-43, 2012

[23] M Levoy and P Hanrahan. Light field rendering. In *Proc. siggraph*, 1996, 31-42

[24] J J Koenderink. The structure of images, Biological cybernetics 50 (5), 363-370, 1984

[25] K P Michael. Geometric, Physical, and Visual Optics Butterworth-Heinemann, 2001

[26] D Regan. Human perception of objects, Sinauer Associates, Inc. Sunderland, Mass. 116-120, 2000

[27] A J Jackson and I L Bailey. Visual acuity, Optometry in practice, 5, 53-70, 2004

[28] S H Schwartz. Geometrical and Visual Optics. McGraw-Hill Medical, 2013

[29] J J Koenderink and A J van Doorn. Representation of local geometry in the visual system, Biological cybernetics 55 (6), 367-375, 1987

[30] O Faugeras. Three-Dimensional Computer Vision: A Geometric Viewpoint, MIT Press,1993

[31] O Faugeras, QT Luong, and T Papadopoulo. The Geometry of Multiple Images: The Laws That Govern the Formation of Multiple Images of a Scene Andsome of Their Applications, MIT press, 2001

[32] O Stanley and P Nikos. Geometric Level Set Methods in Imaging, Vision, and Graphics Springer-Verlag New York Inc. 2012

[33] M K Bennett. Affine and Projective Geometry, John Wiley & Sons Inc 1995

[34] J A Shufelt. Performance evaluation and analysis of vanishing point detection techniques, IEEE Transaction on Pattern Analysis and Machine Intelligence, 21, 3, 282-288, 1999

[35] A Almansa, A Esolneux, and S Vamech. Vanishing point detection without any priori information, IEEE Transaction on Pattern Analysis and Machine Intelligence, 25, 4, 502-507, 2003

[36] M Clerc and S Mallat. Texture gradient equation for recovering shape from texture, IEEE Transaction on Pattern Analysis and Machine Intelligence, 24, 4, 536-549, 2002

[37] E Adelson and J Bergen, The plenoptic function and the elements of early vision, In Computational Models of Visual Processing. MIT Press, Cambridge, MA, 385-394, 1991

[38] E H Adelson and John Y A Wang. Single Lens Stereo with a Plenoptic Camera, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14, 2, 1992

[39] S Zeki. A vision of the Brain, Oxford: Blackwell Scientific Pub., 1993

[40] M S Livingstone and D H Hubel. Anatomy and physiology of a color system in the primate visual cortex, *Journal of Neuroscience*, 4, 309-356, 1984

[41] M S Livingstone and D H Hubel. Psychophysical evidence for separate channels for the perception of form, color, movement, and depth, J. Neurosci., 7, 3416-3468, 1987

[42] L McMillan and G Bishop. Plenoptic modeling: An image-based rendering system, Computer Graphics, 39-46, August 1995

[43] O Schreer, P Kauff, and T Sikjora, eds, 3D video communication: Algorithms, concepts and real-time systems in human centered communication, John & Sons, Inc., New York, 110-150, 2005

[44] M C Potter, B Wyble, C E Hagmann, and E S McCourt. Detecting meaning in RSVP at 13 ms per picture. *Attention, Perception, & Psychophysics* 12,2013

[45] J G Nicholls, A R Martin, B G Wallace, and P A Fuchs. From Neuron to Brain. Fourth

[46] Zhao Songnian, Zou Qi, Jin Zhen,Yao Guozheng, Yao Li. A computational model of early vision based on synchronized response and inner product operation, *Neurocomputing*, 73, 3229-3241, 2010

[47] Zhao Songnian, Zou Qi, Jin Zhen, Yao Guozheng, Yao Li. Neural computation of visual imaging based on Kronecker product in the primary visual cortex, *BMC Neuroscience*, 11: 43,1-14, 2010

[48] D H Hubel and T N Wiesel. Ferrier lecture, Functional architecture of macaque monkey visual cortex. *Proc R Soc Lond B Biol Sci*, 198: 1-59, 1977

[49] D H Hubel. Exploration of the primary visual cortex: 1955-1978. *Nature*, 299: 515-524, 1982

[50] D A Forsyth and J Pence. Computer vision: A modern approach (2ed), Prentice –Hall, 2002

[51] S Cunningham. Computer graphics: Programming in OpenGL for Visual communication, Prentice-Hall, 2007

[52] L G Shapiro and G C Stockman, Computer vision, Prentice-Hall, 2001