Geometric Models in Stereoscopic Video

Geometric Models in Stereoscopic Video

Cheng Hong Yang

Université du Québec
Institut national de la recherche scientifique
INRS-Télécommunications
16 Place du Commerce, Verdun
Québec, Canada, H3E 1H6

Août 1995

Rapport technique de l'INRS-Télécommunications no. 95-12

Summary

This report has be written within the scope of the course TEL-7403 summer 1995 under the direction of Prof. Janusz Konrad, my director of research. The main aim of this work is to lay out useful geometric models for further study in stereoscopic video processing. The fundamental concepts in stereoscopic processing are reviewed in company with geometric models. We try describe the concerned problems not only by giving simply a definition but also by giving a mathematical analysis. Another aim is to develop a technique to eliminate vertical parallax due to the fact that the sensor planes are not parallel in a toed-in camera setup. Finally some stereoscopic properties of human visual system are mentioned.

Contents

List of Figures

Introduction

The geometric models are important in stereoscopic video processing. It is not useful to itemize the concepts in stereoscopic processing without giving mathematical geometric models. The introduction of projective geometry into stereoscopic processing makes interpreting the properties easier and clearer. Keystone distortion is a well known problem in toed-in camera setup. It must be analyzed quantitatively if we want effect a restoration. The two main camera models, parallel and toed-in cameras, and their properties have been described under mathematical models. We think that the rectification technique could eliminate keystone distortion. This technique has been developed within this work. In the first section, the fundamental model in computer vision of a pinhole camera is described. We review the concepts of epipolar geometry, an important concept in stereo vision, and that of disparity and parallax in the second section. In the third section we lay out in greater detail the models of The two main camera models, parallel and toed-in cameras, and the properties. The mathematical computations to develop rectification technique are showed in the forth section. In the last section, based on recent articles, some stereoscopic properties of the human visual system are discussed.

1 Modeling and Calibrating Cameras

1.1 Pinhole Camera

A pinhole camera model is depicted in Figure 1. It consists of two screens. A small hole has been punched in the first screen, and through this hole some of the rays of light emitted or reflected by the object form an inverted image of that object on the second screen.

A geometric model of the pinhole camera can be directly built. It consists of a plane R called retinal plane in which the image is formed through an operation called *a perspective projection*; a point C, the optical center, located at a distance f , the focal length of the optical system, is used to form the image $\mathbf{m}$ in the retinal plane of 3D point $\mathbf{M}$ as the intersection of the line (C, M) with the plane R. The optical axis is the line going through the optical center C and perpendicular to R.

The point $\mathbf{m}$ is a vector in R^2, and $\mathbf{M}$ is a vector in R^3.

$$\mathbf{M} = \begin{pmatrix} X_o \\ Y_o \\ Z_o \end{pmatrix}, \mathbf{m} = \begin{pmatrix} x \\ y \end{pmatrix}$$

If we choose the coordinate system $\{C, (X_o, Y_o, Z_o)\}$, C being the origin (called *standard coordinate*), image points are interpreted by two R^2 coordinates $\{O, (x, y)\}$,

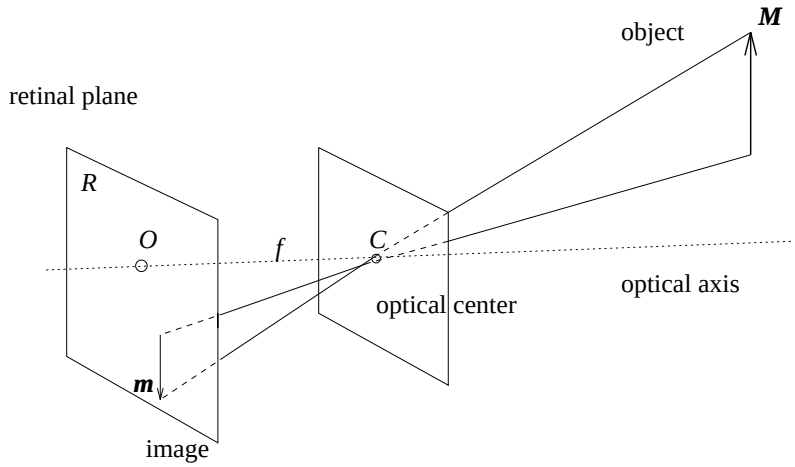

Figure 1: Pinhole camera model

where O is the intersection point of the optical axis and retinal plane. The relationship can be expressed by:

$$-\frac{f}{Z_o} = \frac{x}{X_o} = \frac{y}{Y_o}, Z_o \neq 0 \text{ or } \begin{cases} x &= \frac{-f}{Z_o} X_o, Z_o \neq 0 \\ y &= \frac{-f}{Z_o} Y_o \end{cases}$$

Unfortunately, this is not a linear relationship, therefore a powerful tool would be lost. Using a supplementary component, say s, an image point can be presented by a triplet $(\tilde{x}, \tilde{y}, s)$, with $\tilde{x} = -fX_o, \tilde{y} = -fY_o, s = Z_o$.

x, y could be obtained through this parameter, say $x = \frac{\tilde{x}}{s}, y = \frac{\tilde{y}}{s}$. The idea is that $(\tilde{x}, \tilde{y}, s)$ and $(x, y, 1)$ (or (x, y)) should represent the same point. Therefore we introduce the projective coordinate. A projective point $(x, y, 1)$ can be considered as an ordinary point (x, y). To distinguish we denote $\tilde{\mathbf{M}}$ a projective vector. If $\tilde{\mathbf{M}} = (\tilde{x}, \tilde{y}, z)$ is a projective point then $\forall k \neq 0$ all points $k\tilde{\mathbf{M}}$ are equivalent, i.e, standing for the same projective point. A projective point like $(\tilde{x}, \tilde{y}, 0)$ is a point at infinity in the projective space. This concept can be easily generalized to a space of higher dimension (R^3).

Hence, the relationship mentioned earlier is expressed in linear form in projective coordinate system if we denote U the supplementary component of a 3D point

$$\begin{pmatrix} \tilde{x} \\ \tilde{y} \\ s \end{pmatrix} = \begin{pmatrix} -f & 0 & 0 & 0 \\ 0 & -f & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \begin{pmatrix} \tilde{X}_o \\ \tilde{Y}_o \\ \tilde{Z}_o \\ U \end{pmatrix}$$

or simply

$$\tilde{\mathbf{m}} = \tilde{\mathbf{P}}\tilde{\mathbf{M}} \tag{1}$$

And the relationship will hold in the same form with different matrixes $\tilde{\mathbf{P}}$ and for any choice of 3D and retinal plane coordinate systems.

1.2 Changing Coordinate Systems: Intrinsic and Extrinsic Parameters

First, we consider the effects by changing the origin of the image coordinate and the units on the x, y axes. The translation of the origin is represented by a translation vector $\mathbf{v}$ written in the new system. The ratio of the change of the units is represented by a 2×2 matrix $\mathbf{S}$. The old system is standard coordinate system we discussed before.

$$\mathbf{m}_{new} = \mathbf{S}\mathbf{m}_{old} + \mathbf{v}$$

where

$$\mathbf{v} = \begin{pmatrix} v_x \\ v_y \end{pmatrix}, \mathbf{S} = \begin{pmatrix} k_x & 0 \\ 0 & k_y \end{pmatrix}.$$

All affine transformations can be written in linear form if we adopt projective coordinate system.

$$\tilde{\mathbf{m}}_{new} = \tilde{\mathbf{H}}\tilde{\mathbf{m}}_{old}$$

where

$$\tilde{\mathbf{H}} = \begin{pmatrix} \mathbf{S} & \mathbf{v} \\ \mathbf{0} & 1 \end{pmatrix}.$$

According to the relationship (1),

$$\tilde{\mathbf{m}}_{old} = \tilde{\mathbf{P}}_{old}\tilde{\mathbf{M}}$$

Therefore

$$\begin{aligned} \tilde{\mathbf{m}}_{new} &= \tilde{\mathbf{H}}\tilde{\mathbf{P}}_{old}\tilde{\mathbf{M}} \stackrel{\text{def}}{=} \tilde{\mathbf{P}}_{new}\tilde{\mathbf{M}} \\ \tilde{\mathbf{P}}_{new} &= \begin{pmatrix} -fk_x & 0 & v_x & 0 \\ 0 & -fk_y & v_y & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix} \end{aligned}$$

$\alpha_x = -fk_x, \alpha_y = -fk_x, v_x$, and v_y dot not depend on the position and orientation of the camera in space, and are called intrinsic parameters.

The *normalized coordinate* system which is widely used in motion and stereo applications is defined as a retinal coordinate system obtained by a $\tilde{\mathbf{H}}$, such that

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} & \mathbf{0} \end{pmatrix}$$

By contrast, the world coordinate system could suffer from a rotation $\mathbf{R}$ and a translation $\mathbf{T}$ without magnification. It corresponds to the fact that we do not always describe the position of an object according to the position of the camera.

$$\mathbf{M}_{old} = \mathbf{T} + \mathbf{R}\mathbf{M}_{new}$$

Also it gives a linear relationship in projective coordinates

$$\tilde{\mathbf{M}}_{old} = \tilde{\mathbf{K}}\tilde{\mathbf{M}}_{new}.$$

Then

$$\begin{aligned} \tilde{\mathbf{m}} &= \tilde{\mathbf{P}}_{old}\tilde{\mathbf{K}}\tilde{\mathbf{M}}_{new} \stackrel{\text{def}}{=} \tilde{\mathbf{P}}_{new}\tilde{\mathbf{M}}_{new} \\ \tilde{\mathbf{P}}_{new} &= \tilde{\mathbf{P}}_{old}\tilde{\mathbf{K}} \end{aligned}$$

The six parameters describing the rotation and translation are called extrinsic parameters.

Therefore the overall projective matrix is obtained after the adjustment of the coordinate systems on the sensor and in the real world:

$$\tilde{\mathbf{P}} = \tilde{\mathbf{H}}\tilde{\mathbf{P}}_{old}\tilde{\mathbf{K}}.$$

We can write the general form of the matrix P as a function of the intrinsic and extrinsic parameters

$$\tilde{\mathbf{P}} = \begin{pmatrix} \alpha_u \mathbf{r_1} + u_0 \mathbf{r_3} & \alpha_u t_x + u_0 t_z \\ \alpha_v \mathbf{r_2} + v_0 \mathbf{r_3} & \alpha_u t_y + v_0 t_z \\ \mathbf{r_3} & t_z \end{pmatrix}$$

$\mathbf{r}_1, \mathbf{r}_2$ and $\mathbf{r}_3$ are row vectors of matrix R.

Calibration is the process of estimating the four intrinsic and the six extrinsic parameters of a camera. The projective matrix $\tilde{\mathbf{P}}$ should satisfy certain constraints. The linear method is to write the relationship between coordinates of 3D and 2D points linearly so that least-squares method can be used. Nonlinear method is to minimize a certain quantity (chosen according to a constraint) subject to the condition that $\tilde{\mathbf{P}}$ must satisfy. Linear method is very simple and gives results that are just as good as those obtained by a minimizing criterion [2].

2 Stereo Vision

Our eyes see the world from a different point of view. Two slightly different images are obtained on our left and right retinas. The mind combines two different, although similar, images into one image (*fusion*) and the depth can be perceived due to the difference between the images(*disparity*).

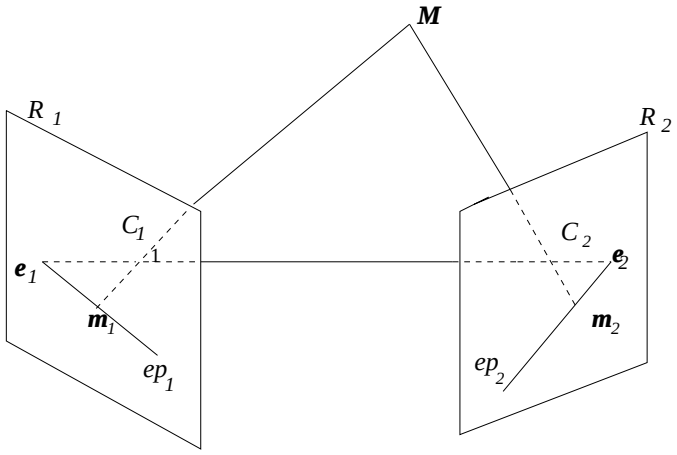

Figure 2: Stereo vision system: Epipolar constraint

2.1 Correspondence Problem

A stereo vision system consists of two or more cameras(Figure 2). To reconstruct 3D coordinates from a pair (or more) of given 2D images obtained by the cameras, we must first deal with the *correspondence problem*: given a token in the image 1 what is the corresponding token in the image 2 (see Figure 2). Because there are too many potential candidates to choose, some properties must be exploited. The fundamental constraint is the *epipolar* constraint.

From Figure 2 we reason that all possible matches $\mathbf{m_2}$ in R_2 of $\mathbf{m_1}$ must be located on the images of the half line ep_2 (*epipolar line*). The point $\mathbf{e}_2(\mathbf{e}_1)$, which is the intersection point of the line C_1C_2 (*base line*) and the plane $R_2(R_1)$, is called the epipole of the second (first) camera with respect to the first(second) camera. The plane C_1MC_2 is called *epipolar plane* defined by $\mathbf{M}$. The epipolar lines are intersections of epipolar plane and R_1 and R_2.

To obtain the epipolar constraint, we compute the epipolar line ep_2. In projective geometry, a line is represented by a vector, for example $\mathbf{n}$. All points $\tilde{\mathbf{m}}$ on the line should satisfy the condition

$$\tilde{\mathbf{m}}\tilde{\mathbf{n}} = 0.$$

We know that $\mathbf{e}_2$ is on the epipolar line, and that it's simple to compute its coordinates from those of C_1 and C_2. The point at infinity on the line C_1M has the form $(\mathbf{D}^T, 0)^T$ In order to compute this point, we write matrixes $\tilde{\mathbf{P}}_\mathbf{i} = (\mathbf{P}_i, \mathbf{q}_i)$. As we have seen, $\mathbf{q}_i = 0$ in the standard coordinate. But they are not zero in general. Therefore

$$\tilde{\mathbf{m}}_1 = [\mathbf{P}_1, p_1] \begin{pmatrix} \mathbf{D} \\ 0 \end{pmatrix}.$$

We obtain

$$\mathbf{D} = \mathbf{P}_1^{-1}\tilde{\mathbf{m}}_1.$$

This point has his image at position $\mathbf{P_2}\mathbf{P}_1^{-1}\tilde{\mathbf{m}}_1$ on ep_2 on the R_2 .

According to the properties of projective geometry, the representation the line passing through two points is the cross product of those points:

$$\tilde{\mathbf{e}}_2 \wedge \mathbf{P_2}\mathbf{P}_1^{-1}\tilde{\mathbf{m}}_1$$

In order to show the epipolar constraint, we use an antisymmetric matrix $\mathbf{F}_2$ such that $\forall \mathbf{X}, \quad \tilde{\mathbf{e}}_2 \wedge \mathbf{X} = \mathbf{F}_2\mathbf{X}$ holds true.

All points $\tilde{\mathbf{m}}_2$ on the line ep_2 should satisfy

$$\tilde{\mathbf{m}}_2\mathbf{F}_2\mathbf{P}_2\mathbf{P}_1^{-1}\tilde{\mathbf{m}}_1 = 0$$

or simply

$$\tilde{\mathbf{m}}_1\tilde{\mathbf{F}}\tilde{\mathbf{m}}_2 = 0.$$

This constraint is called Longuet-Huggins equation.

2.2 Disparity and Parallax

Retinal disparity describes the difference between the image obtained on the left and right fovea. Retinal disparity can be defined as the difference between the convergence angle associated with the objet and the convergence angle associated with the fixated point [3] In Figure 3, it equals to $\alpha - \phi$. The locus of the points having images at the same position on the fovea in each eye is the *horopter*. Horopter usually describe a locus of points that should result in zero disparity. Objects slightly in front of or behind the horopter produce small disparities. They are located within a region surrounding the horopter called *fusional area*. Inside this area, fusion is possible, and the two monocular images give rise to a single three-dimensional percept. Objects farther away from the horopter, outside the fusional area produce large disparities and give rise to monocular half-images that are generally not fusable.

A stereoscopic display is one that differs from a planar display in only one respect: it is able to display parallax values of images points. A pair of images (left and right) are displayed on a single screen by some method so that left image can be only seen by the left eye of the viewer and the right image can be only seen by the right eye. The *parallax* (distance between the homologous points of the right and left images; usually we confuse disparity and parallax)in the two images produces disparity in the eyes thus providing the stereoscopic cue.

Four types of parallax can be classified (Figure 4):

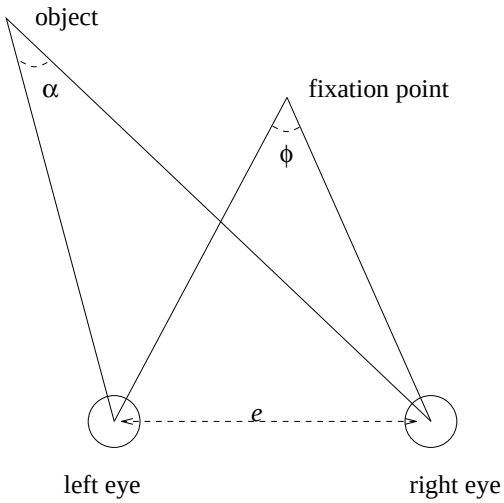

Figure 3: Definition of retinal disparity

- Zero parallax: the homologous image points of the two images exactly correspond to each other.

- Uncrossed or positive parallax (Figure 4(a)). In one type of positive parallax the axes of the left and right eyes are parallel, this happens in the visual world when looking at an object at a great distance from the observer. Any uncrossed value of the parallax between zero and e produces image appearing to be within the space behind the screen.

- Another type of a uncrossed parallax happens when the images are separated by some distance greater than the interocular distance e. When this occurs, the lines of sight from the eyes to the image points are diverging, and the eyes must diverge, or outward, in order to fuse such image points. If the eyes are called upon to fuse stereoscopic images with large angular values of divergence, the fatigue and discomfort will result [5].

- Crossed or negative parallax (Figure 4(b)). Any 3D images generated by crossed parallax appear to be closer than the plane of the screen.

The amount of screen parallax my be computed with respect to the geometric model of the scene as

$$P = \frac{e(Z_i - V)}{Z_i}$$

where Z_i is the depth of the 3D image of the object, and V is the viewing distance [3] (see Figure 4(c)).The parallax is also very usually measured in degrees of arc of the angle it produces, say $2 \arctan \frac{P}{2V}$. It has been suggested that as a rule for stereoscopic graphics the parallax must not exceed 1.5 degrees [6].

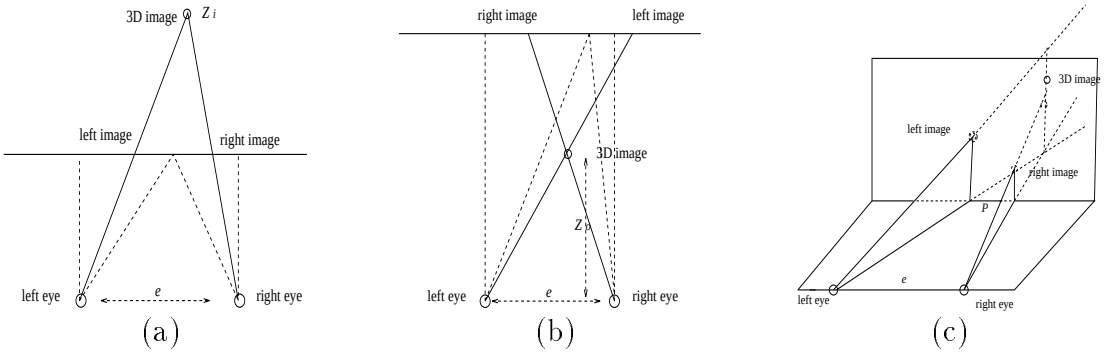

Figure 4: Classification of parallax: (a) uncrossed parallax (b)crossed parallax (c) position of 3D image

2.3 Motion Estimation

It seems somewhat strange to present the motion estimation problem in a stereo vision section. However, the scenario of moving images is similar to the scenario of stereoscopic images; there are several (at least two) projections of the same object.

Let us assume that a camera is moving in a static environment and following an unknown trajectory and that we know several (e.g., 2) images $\mathbf{m_1}, \mathbf{m_2}$ of a an object $\mathbf{M}$. C_1 is the position of the optical center at the first time instant, C_2 is the position of the optical center at the second time instant (considering R_1 as the retinal plane at the first instant and R_2 as the that at the second instant, see Figure 2). We describe the motion by rotation $\mathbf{R}$ and translation $\mathbf{v}$. Let M_1 and M_2 be the 3-D points, which correspond to the 2-D image points $\mathbf{m}_1$ and $\mathbf{m}_2$ on the retinal planes, in the standard coordinate systems associated with the each camera respectively. . It is clear that the vectors C_1C_2, C_1M_1 and C_2M_2 are coplanar.If we interpret the vector in same coordinate system (e.g., that of the first instant), we would have

$$\mathbf{M}_1(\mathbf{v} \wedge \mathbf{R}\mathbf{M}_2) = 0, \tag{2}$$

or in a matrix form:

$$\mathbf{M}_1\mathbf{E}\mathbf{M}_2 = 0.$$

Note that the constraint above delimits only the direction of $\mathbf{v}$ not the scale.The matrix $\mathbf{E}$ is called the *essential matrix*. Some properties of $\mathbf{E}$ have been investigated in [2]. Two parameters determine the translation and three parameters determine the rotation. Therefore, the minimum number of the correspondences to resolve the problem is five.

The five-point algorithm is based on the properties of projective geometry. The rigidity constraint for five pairs of correspondence points described in projective terms permits the computation of the epipoles are computed. Another approach to the determination of motion try to determine the essential matrix from a set of eight pairs of points in the first step. The condition above is expressed as a linear system whose

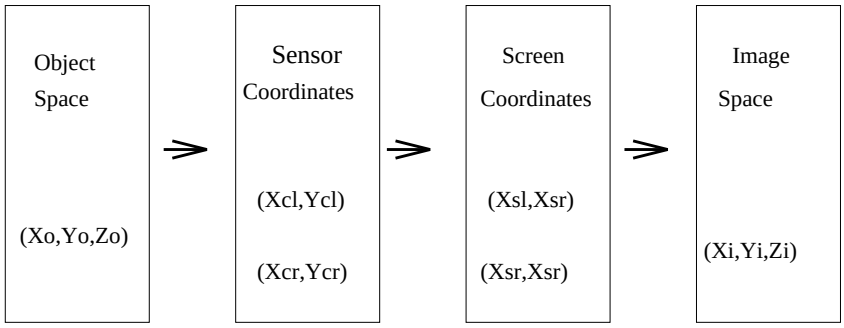

Figure 5: Processing of the transforms

variables are the matrix elements. The second step is to decompose the matrix into a translation vector and a rotation matrix. This is what we call the eight-point algorithm.

3 Geometry of Stereoscopic Video System

A stereoscopic system consists of a pair of video camera mounted side by side to obtain left and right images. A stereoscopic display consists of a single display surface on which the left and right images are displayed and separated by some method. The geometry of a stereoscopic video system can be determined by considering the imaging and display processes as three separate coordinate transforms (Figure 5). First, from 3D coordinates of the object in real world (object space) to the position on the two sensors. Secondly, from these coordinates to the coordinates on the display screen, and finally to a 3D position perceived by the viewer.

The relationship between the final 3D image coordinates and the position of 2D images on the display surface is given by Woods, Docherty and Koch[9] (the drawing in Figure 4(3) makes easy to obtain these equation)

$$X_i = \frac{e(X_{sl} + X_{sr})}{2(e - P)}, \tag{3}$$

$$Y_i = \frac{e(Y_{sl} + Y_{sr})}{2(e - P)}, \tag{4}$$

$$Z_i = \frac{Ve}{e - P} \tag{5}$$

Refer to the following listing for the explanation of the symbols.

The transform from the sensor coordinates to the screen coordinates is simply a magnification process. The ratio of the magnification is denoted by M.

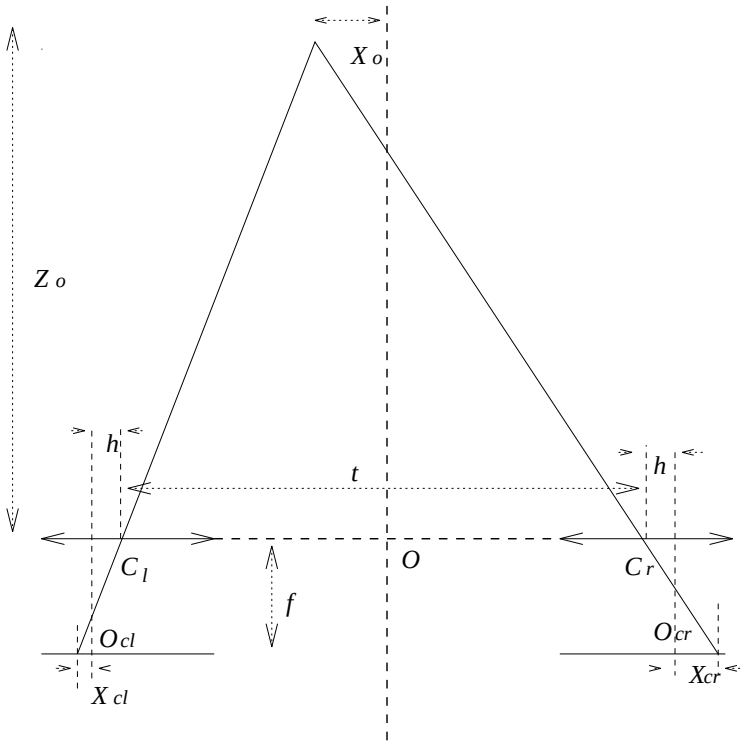

Figure 6: Convergence attained by a lateral shift

There are two types of camera setup with respect to the convergence: parallel cameras and toed-in cameras [1].

Since some symbols will be cited quite frequently in the following, we itemize them before going on with greater detail.

t	Camera separation, the distance between the two optical centers.
f	Local length
w	Sensor width
2ϖ	Horizontal angle of view of the camera.
M	Frame magnification, ratio of screen width to sensor width.
P	Image parallax
C_l, C_r	Optical centers of cameras.
$O_{sl}O_{sr}$	Origins of the sensor planes.
2ϕ	Angle of convergence produced by the optical axes.
V	Viewing distance.
h	Sensor axial offset.

3.1 Parallel Cameras

The convergence is ensured by the external shift h of each sensor with respect to the optical axis of its lens. The situation is illustrated in the Figure 6.

$$X_{cl} = \frac{f(t+2X_o)}{2Z_o} - h, \tag{6}$$

$$X_{cr} = -\frac{f(t-2X_o)}{2Z_o} + h \tag{7}$$

$$Y_{cl} = Y_{cr} = \frac{Y_o f}{Z_o} \tag{8}$$

Combining the above equations with equations (2-4), we can get the overall coordinates of 3D image perceived by the viewer. We find that

$$X_i = \frac{MfeX_o}{Mft - Z_o(2Mh - e)},$$

$$Y_i = \frac{MfeY_o}{Mft - Z_o(2Mh - e)},$$

$$Z_i = \frac{VeZ_o}{Mft - Z_o(2Mh - e)}$$

In order to estimate distortion in 3D image, we compute the derivatives of the length perceived by the viewer to the real length in the three directions:

$$M_x = \frac{\partial X_i}{\partial X_o} = \frac{Mfe}{Mft - Z_o(2Mh - e)}, \tag{9}$$

$$M_y = \frac{\partial Y_i}{\partial Y_o} = M_x \tag{10}$$

$$M_z = \frac{\partial Z_i}{\partial Z_o} = \frac{MVfet}{(Mft - Z_o(2Mh - e))^2} \tag{11}$$

Obviously there is no form distortion in the XY plane because $M_x = M_y$. And there is no depth plane curvature because Z_i depends clearly only on Z_o.

Let's look at the ratio between $M_x(M_y)$ and M_z, which would indicate the degree of form distortion.

$$\frac{M_z}{M_x} = \frac{Vt}{Mft - Z_o(2Mh - e)}$$

This value depends on Z_o, i.e., the compression or magnification in depth is not constant. A reasonable hypothesis would be to assume that $\frac{M_z}{M_x}$ is independent of Z_o, that is :

$$2Mh = e, \qquad \frac{M_z}{M_x} = \frac{V}{Mf}$$

If $V = Mf$, the ratio is unitary. That indicates no form distortion in image space. This constraint is well known to lens designers and photographers [6]. When

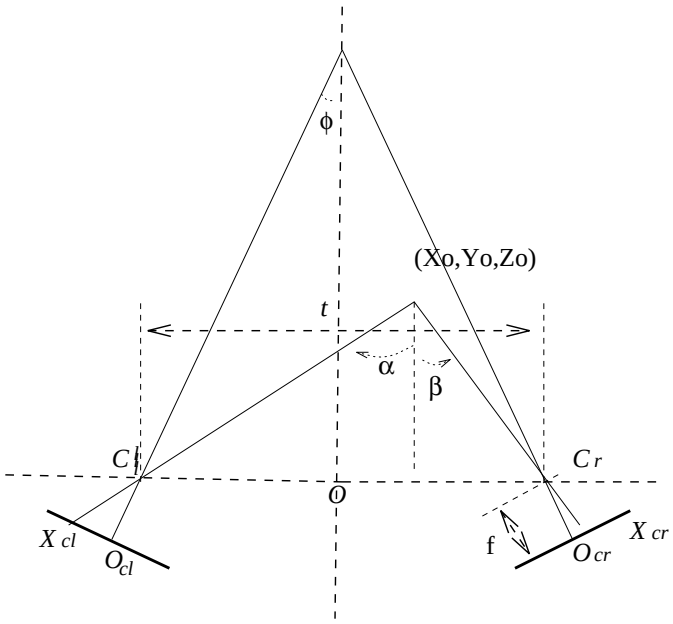

Figure 7: Model of toed-in camera

this constraint is fulfilled, the angle subtended by display screen at the viewer's eye, equals that subtended by the camera's lens on the sensor.

If all these hypotheses hold true, then

$$Z_i = \frac{eZ_o}{t}$$

That means if the eye separation is equal to lens separation, viewer would perceive exactly on the display what we can see in the real world.

3.2 Toed-in Cameras

The situation of a toed-in camera setup is illustrated in Figure 7. The coordinates in the two sensors are given by Woods, Docherty and Koch [9]. To simplify the expressions we note

$$\alpha = \arctan\frac{t + 2X_o}{2Z_o}, \qquad \beta = \arctan\frac{t - 2X_o}{2Z_o}$$

To obtain the relationship between the object coordinates and the sensor coordinates, we refer to Figure (8).

The computation for determining X_{cl}, X_{cr} is direct because $\angle ICO_l = \angle FC_lG = \alpha - \phi$. Notice that

$$\frac{Y_{cl}}{Y_o} = \frac{IC}{CG}$$

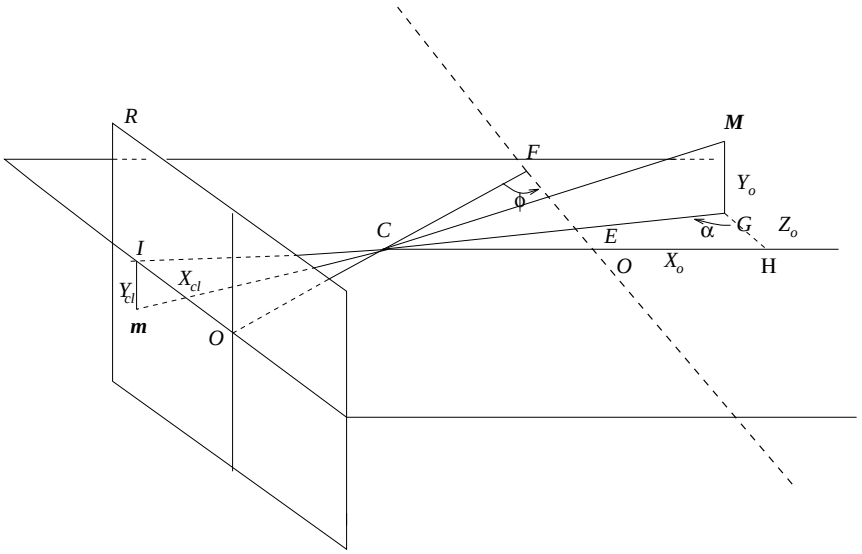

Figure 8: Computing the sensor coordinates on the left sensor $CH = X_o, MG = Y_o, GH = Z_o, \angle CEO = \angle CGH = \alpha$

and

$$CG = \frac{Z_o}{\cos\alpha}$$

We can write:

$$X_{cl} = f\tan(\alpha - \phi), \tag{12}$$

$$X_{cr} = -f\tan(\beta - \phi) \tag{13}$$

$$Y_{cl} = \frac{fY_o\cos\alpha}{Z_O\cos(\alpha - \phi)} \tag{14}$$

$$Y_{cr} = \frac{fY_o\cos\beta}{Z_O\cos(\beta - \phi)} \tag{15}$$

We notice that the more farther from the center the greater the vertical parallax. This phenomena is symmetric and caused by *keystone distortion*. We look in greater detail at keystone distortion. Assume that a straight line located horizontally in front of the camera system, is precisely defined by

$$\begin{cases} Z_o = C_z \\ Y_o = C_y \end{cases}$$

For example, we consider the image in the left sensor (X_{cl}, Y_{cl}).

$$\frac{d(X_{cl})}{d(\alpha)} = f\sec^2(\alpha - \phi)$$

$$\frac{d(Y_{cl})}{d(\alpha)} = -\frac{fC_y\sin\phi\sec^2(\alpha - \phi)}{C_z}$$

$$\frac{d(Y_{cl})}{d(X_{cl})} = -\frac{C_y\sin\phi}{C_z}$$

The last equation means that the image of the horizontal line will not be horizontal, but will have a slope that increases with ϕ and C_y. Only when $\phi = 0$ the image has no keystone distortion. When the line on the horizontal epipolar plane, its image would not be affected by this kind of distortion.

Finally we give the overall 3D image coordinates.

$$\begin{aligned} X_i &= \frac{Mfe(\tan(\alpha-\phi)-\tan(\beta-\phi))}{2e-4Mh+2Mf(\tan(\alpha-\phi)+\tan(\beta-\phi))} \\ Y_i &= (\frac{1}{\cos(\alpha-\phi)}+\frac{1}{\cos(\beta-\phi)})\frac{MfeY_o\cos\phi}{Z_o(e-2Mh+Mf(\tan(\alpha-\phi)+\tan(\beta-\phi))} \\ Z_i &= \frac{Ve}{e-2Mh+Mf(\tan(\alpha-\phi)+\tan(\beta-\phi))} \end{aligned}$$

Z_i depends on both X_o and Z_o because α and ϕ depend on X_o and Z_o. It produces an effect called depth plane curvature. Roughly, X_o increases Z_i decreases for constant Z_o. Second, the relationship between Z_o and Z_i in not linear (*depth non-linearity*). The non-linearity of depth in the display can lead to incorrect depth perception on the monitor and if the camera system is in motion it can lead to false estimation of velocity.

4 Rectification

As we mentioned before the toed-in camera setup suffers from vertical parallax due to keystone distortion. A restoration technique could be imagined: rectification projecting the sensor image onto a third plane parallel to the base line of the camera.

4.1 Position of the plane of rectification

If we know the angle of convergence, the angle between the sensors and the plane of rectification equals to the angle of convergence.

To determine position of the plane of rectification, the distance between the old plane and the plane of rectification, must be computed. When we simply rotate the sensor plane, the width of the image would be changed. A parallel translation can adjust the width, i.e., we move it to a position such that the width is the same.

Let w be the width of the sensor (Figure(10), and 2ϖ the angle of view of the camera. They have the relationship $\varpi = \arctan\frac{w}{2f}$.The final plane of rectification and the optical axis CO intersect at O'. The position of the plane is characterized by length of $f' = O'C$.

$$GH = \frac{w}{2}\cos\varpi(\frac{1}{\cos(\varpi+\phi)}+\frac{1}{\cos(\varpi-\phi)})$$

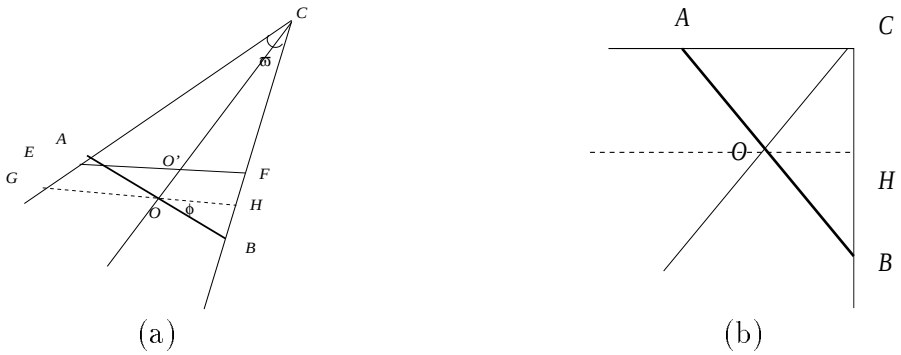

Figure 9: (a): Determining the position of the plane of rectification $OG = \frac{w\cos\varpi}{2\cos(\phi+\varpi)}, OH = \frac{w\cos\varpi}{2\cos(\phi-\varpi)}$ (b): Situation when a projection of an image onto a finite plane is impossible

The width of the image projected in the plane of rectification (EF) is equal to w, therefore

$$f' = \frac{2f}{\cos\varpi(\frac{1}{\cos(\varpi+\phi)} + \frac{1}{\cos(\varpi-\phi)})}. \tag{16}$$

Assume that the old sensor plane is S. The plane of rectification is S'. The angle between the two planes is ϕ (in an anti-clockwise direction) as showed in Figure 10. We always have:

$$\frac{y^r}{y} = \frac{x^r\cos\phi}{x} = \frac{f' - x^r\sin\phi}{f}$$

This induces

$$x^r = \frac{f'}{f\cos\phi + x\sin\phi}x \tag{17}$$

$$y^r = \frac{f'\cos\phi}{f\cos\phi + x\sin\phi}y \tag{18}$$

Alternatively, if the rotation is in a clockwise direction, the relation is similar:

$$x^r = \frac{f'}{f\cos\phi - x\sin\phi}x \tag{19}$$

$$y^r = \frac{f'\cos\phi}{f\cos\phi - x\sin\phi}y \tag{20}$$

This system is equivalent to a parallel system with the same optical center and a fictitious focal length $f^r = f'\cos\phi$. Therefore there will be no longer vertical parallax. We can compute the new coordinates Y^R_{cl}, Y^R_{cr} directly from the formula given for parallel cameras, or by substituting Y^R_{cl}, Y^R_{cr} by the relationships given above.

$$Y^R_{cl} = \frac{Y_o f'\cos\phi}{Z_o}$$

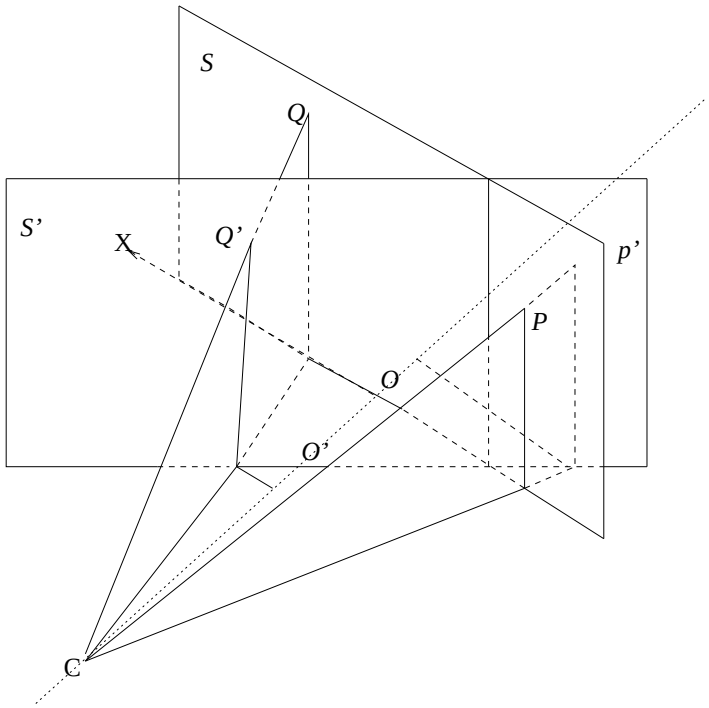

Figure 10: Rectification

Obviously the image center O'_{sl} will not be the center of the new image. The new coordinates of the image border $X^1_{sl} = -\frac{w}{2}$ and $X^2_{sl} = \frac{w}{2}$ can be computed:

$$\begin{aligned} X^1_{cl} &= -\frac{f'w}{2(f\cos\phi - \frac{w}{2}\sin\phi)} \\ X^2_{cl} &= \frac{f'w}{2(f\cos\phi + \frac{w}{2}\sin\phi)} \\ X^1_{cl} + X^2_{cl} &= -\frac{f'w^2\tan\phi}{2f^2\cos\phi(1+\tan\varpi\tan\phi)(1-\tan\varpi\tan\phi)} \end{aligned}$$

The fictitious system will have an equivalent lateral shift h^r, it defined by the position of the plane, focal length, the width of the sensor and the angle between the sensor plane and the plane of rectification.

$$h^r = f'\cos\phi + \frac{f'w^2\tan\phi}{4f^2\cos\phi(1+\tan\varpi\tan\phi)(1-\tan\varpi\tan\phi)} \tag{21}$$

From the equation (16), we can conclude that a necessary condition for rectification is

$$\phi + \varpi \neq \frac{\pi}{2}.$$

Because the angle ϕ is usually small, and the angle of view of a camera is also inferior to 90 degree, we can rewrite this condition

$$\phi + \varpi \quad < \quad \frac{\pi}{2} \tag{22}$$

otherwise some part of the image will be lost in the plane of rectification (Figure 10).

4.2 Angle of Convergence

In the last section, we discussed a rectification technique under the assumption that the angle of convergence is known. It may be, however, possible that the angle is unknown. Although the angle can be determined from the extrinsic parameters, we have to establish the correspondence between object space and sensor images (*calibration*). The fact is that we have no information on the object space.

We divide the task of computing the angle of convergence into two steps: first, we compute some pairs of homologous points, and then we determine the angle from these correspondence points. The second step seems similar to motion evaluation: the left camera takes the left image in the position of the left camera, and moves to the position of the right camera, and then takes the right image. The condition (2) can be rewritten in the coordinate system we have chosen (O is the origin)

$$(\mathbf{R}_l\mathbf{M}_l)(\mathbf{t} \wedge (\mathbf{R}_r\mathbf{M}_r)) = 0$$

where

$$\mathbf{R}_l = \begin{pmatrix} \cos\phi & o & \sin\phi \\ o & 1 & 0 \\ -\sin\phi & 0 & \cos\phi \end{pmatrix} \qquad \mathbf{R}_r = \begin{pmatrix} \cos\phi & o & -\sin\phi \\ o & 1 & 0 \\ \sin\phi & 0 & \cos\phi \end{pmatrix} \tag{23}$$

The vectors $\mathbf{M}_l$ and $\mathbf{M}_r$ can be expressed with respect to the standard coordinate systems associated with the cameras:

$$\mathbf{M}_l = \begin{pmatrix} x_l \\ y_l \\ -f \end{pmatrix}, \qquad \mathbf{M}_r = \begin{pmatrix} x_r \\ y_r \\ -f \end{pmatrix}, \qquad \mathbf{t} = \begin{pmatrix} t \\ 0 \\ 0 \end{pmatrix}$$

This induces

$$t(y_r(-x_l\sin\phi - f\cos\phi) + y_l(x_r\sin\phi - f\cos\phi)) = 0.$$

In fact, t never equals to zero. It gives a condition:

$$y_r(x_l\sin\phi + f\cos\phi) = y_l(x_r\sin\phi - f\cos\phi) \tag{24}$$

This relationship delimits the domain of search for fitting the correspondence point of a pair of given points (e.g., (x_l, y_l)).

$$y_r = \frac{y_l(x_r\tan\phi - f)}{f(\frac{x_l}{f}\tan\phi + 1)}.$$

The points on the only two epipolar lines that we know, i.e., $y_l = 0,\ y_r = 0$ cannot give any useful information for determining the motion. Let $F(\mathbf{m}_l, \mathbf{m}_r)$ be a measure of similarity between two points $\mathbf{m}_l, \mathbf{m}_r$, and let $\psi(\mathbf{m}_l, \mathbf{m}_r) = 0$ denote the equation (24). Given the angle of convergence, for each point $\mathbf{m}_l = (x_l, y_l)^t$, the correspondence point on the right sensor is in the set determined by the function $\psi(\mathbf{m}_l, \mathbf{m}_r) = 0$. The correspondence point of $\mathbf{m}_l$, $\mathbf{m}_r^*$ should be the solution of:

$$\begin{aligned} \mathbf{m}_r^*(\phi) &= \arg \min_{\mathbf{m}_r(\phi) \in R_r} F(\mathbf{m}_l, \mathbf{m}_r(\phi)) \\ \text{subject to} &\quad \psi(\mathbf{m}_l, \mathbf{m}_r(\phi)) = 0, \mathbf{m}_r \in R_r. \end{aligned} \tag{25}$$

The sum of the values of F over all image depends on the value of ϕ. The best value fitting the angle of convergence should minimize the sum of F. It suffices to solve the following optimization problem in order to determine the angle of convergence:

$$\phi^* = \arg \min_{\phi} \sum_{\mathbf{m}_l \in R_l} F(\mathbf{m}_l, \mathbf{m}_r^*(\phi)) \tag{26}$$

Since $\mathbf{m}_l$ and $\mathbf{m}_r^*(\phi)$ are projections of the same 3-D point $\mathbf{M}$, their intensity and colors should be similar. Certainly, due to the different scene illumination angles with respect to each camera, different imaging sensors, noise etc., these intensities will not be exactly the same. Nevertheless, we assume that the luminance of points $\mathbf{m}_l$ and $\mathbf{m}_r(\phi)$ are sufficiently close to justify the following similarity measure:

$$F(\mathbf{m}_l, \mathbf{m}_r) = \parallel I_l(\mathbf{m}_l) - I_r(\mathbf{m}_r) \parallel .$$

We can also imagine more complex similarity measure, such as that contains also the average or covariance of the luminance in a neighborhood of $\mathbf{m}_l$ and $\mathbf{m}_r^*(\phi)$. Another possibility would be to include color components (chrominances) in the above measure.

We can also reduce the number of points to examine in the problem (26). Instead of examining all points in the images, we could choose a special set of points. Note that a projection of a vertical line in front of the camera is also vertical. An edge detector could be employed to detect vertical edges in two images. Let VE_l, VE_r be the sets of vertical edges. For each $\mathbf{m}_l \in VE_l$:

$$\begin{aligned} \mathbf{m}_r^*(\phi) &= \arg \min_{\mathbf{m}_r(\phi) \in VE_r} F(\mathbf{m}_l, \mathbf{m}_r(\phi)) \\ \text{subject to} &\quad \psi(\mathbf{m}_l, \mathbf{m}_r(\phi)) = 0, \mathbf{m}_r(\phi) \in VE_r. \end{aligned} \tag{27}$$

We want to minimize only the chosen pairs of points in VE_l, VE_r.

$$\phi^* = \arg \min_{\phi} \sum_{\mathbf{m}_l \in VE_l} F(\mathbf{m}_l, \mathbf{m}_r^*(\phi)) \tag{28}$$

Another possible approach is to minimize the quantity of the left hand of the equation (24). It consists exactly of two separate steps. In the first step, a method

permits compute the positions of all homologous points in the two images, i.e., one $\mathbf{m}_r$ is determined by each $\mathbf{m}_l$ given. Given all $\mathbf{m}_l$ and $\mathbf{m}_r$, ϕ could be determined from:

$$\phi^* = \arg\min_{\phi} \sum_{\mathbf{m}_l \in R_l} (y_l(x_r \tan\phi - f) - y_r(x_l \tan\phi + f))^2 \tag{29}$$

The advantage is that ψ is a quadratic function of ϕ. Therefore ϕ can be computed analytically. The condition is that we have a robust algorithm for estimating correspondence points. The difficulty is moved to the step for estimating correspondence or disparity.

For determining the value of ϕ, the derivative of ψ is computed:

$$\frac{d\psi(\phi)}{d\phi} = \sec^2\phi \sum_{\mathbf{m}_l \in R_l} 2(y_r x_l - y_l x_r)((y_r x_l - y_l x_r)\tan\phi - f(y_r + y_l))$$

We know that $\sec\phi$ is not zero.

$$\sum_{\mathbf{m}_l \in R_l} 2(y_r x_l - y_l x_r)((y_r x_l - y_l x_r)\tan\phi - f(y_r + y_l)) = 0$$

We resolve ϕ:

$$\phi^* = \frac{\sum f(y_r x_l - y_l x_r)(y_r + y_l)}{(\sum(y_r x_l - y_l x_r))^2}$$

5 Human Factors

5.1 Disparity and Parallax

In optimal condition, the threshold of stereoscopic acuity is about 2 seconds of arc disparity [4]. Such fine differences in depth will be perceptible if the image can reproduce a horizontal spatial frequency of at least 24 cycles per degree [8].

The upper disparity threshold for binocular single vision depends on the stimulus exposure duration. Mean thresholds are 27 min of arc for crossed disparities, 24 min of arc for uncrossed disparities, given that convergence responses are impossible due to a short presentation period (e.g. 200 ms). With 2s stimulus duration, vergence response enables much large disparities to be brought within the range of fusion. Respective thresholds are 4.9 degree of arc for crossed and 1.6 degree (recall the constraint on the maximum parallax in the Section 2) of arc for uncrossed disparities[7, 10]. Experience with 3D film, however, suggests that large disparities may lead to increased visual strain. Further studies are necessary to arrive at final conclusion with respect to the upper disparity limit acceptable for prolonged viewing of stereoscopic displays [7].

Perception depends on the individual, some people can handle more parallax on a screen than others. Woods, Dochertty and Koch's experiment [9] revealed a wide range of responses. The result also suggested that depth range improved with increased exposure to stereoscopic display.

The same result indicated that homologous points should have less 7 mm (on a 16inch screen, viewing distance not reported) of vertical parallax for image fusion to be possible. It also reported that eye strain was apparent at higher values of vertical parallax.

5.2 Accommodation

There is strong evidence that the imbalance between accommodation and convergence is a major cause of increased visual fatigue with stereo display. A recent study by N.Hiruma on the accommodation response of observers viewing stereoscopic TV images demonstrated that the link between convergence and accommodation depends on the disparity of a displayed object. When the stereoscopic depth of an object is greater than the depth of focus of the observer's eyes, the accommodation response is suppressed regardless of the convergence angle [7].

5.3 Crosstalk

Crosstalk in a stereoscopic display results in each eye seeing an image of the unwanted perspective view.

In an ideal field-sequential stereoscopic display, the image of each field, made up of glowing phosphors, would vanish before the next field was written. In practice it's not what happens. After the right image is written it will persist while the left image is being written. The perception of ghosting varies with the brightness of the image color, and more importantly with parallax and image contrast. Given the present state of the art of monitors and their display tubes, the green phosphor has the largest afterglow and produce the most ghosting [6].

Yeh and Silverstrain [10] point out that the crosstalk has no significant effect on fusion limits for the contrast ratio and stimulus configuration used.

6 Conclusion

In projective coordinate system many properties, and geometric constraints can be described in a very brief form. The computation is direct. Many researchers noticed the various problems brought by toed-in camera and recommended the parallel camera. In practice, a direct application of a such camera is still difficult. Rectification

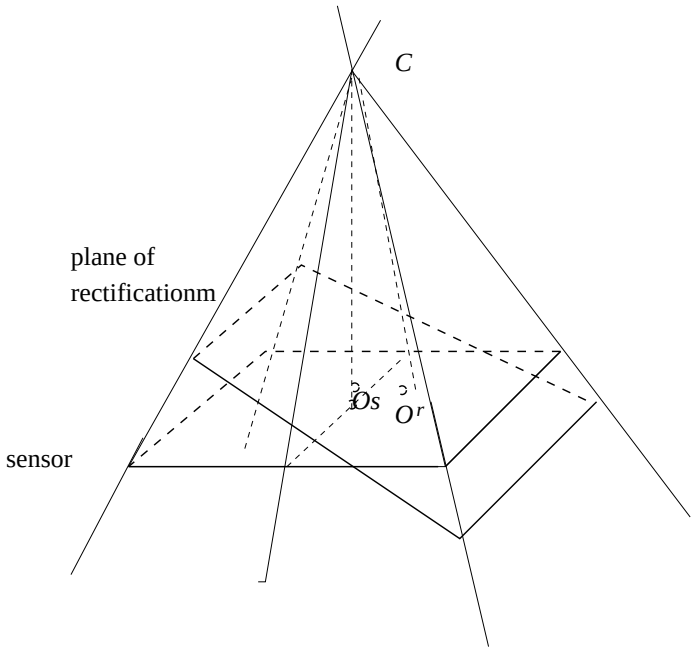

Figure 11: Comparison of the image form on the sensor and that of the image generated by rectification: O_{sr} center of sensor, O_r: new center on the plane of rectification.

seems to be a potential alternative to develop. We have also noticed that the height of the image would be changed; the consequence is the fine blanking interval appearing at the top and on the bottom (See Figure(11)). Because the rectification generally transforms a pixel position to subpixel position, the interpolation techniques will be employed to determine the value at each pixel on the plane of rectification.

It must be noted that keystone distortion is not the only resource of vertical parallax. Lens radial distortion resulting from the fact that lens has different local lengths at various radial distance from the center of the lens. can also induce vertical parallax. Rectification should alleviate keystone distortion but there is no guarantee that vertical parallax would be cured. Lens radial distortion can be significant source of vertical parallax, particularly when wide angle lenses are used on the camera system. When vertical parallax due to radical distortion is seen to be a problem, lens with low radical distortion should be chosen. Aspherical lens may meet this requirement [9].

References

[1] B. Choquet and J. Fournier, "Importance of opto geometrical adjustments for stereoscopic television," in *2nd International Conference on 3D Media Technologies*, 1992.

[2] O. Faugeras, *Three-dimensional Computer Vision: A geometric viewpoint.* The

MIT Press, 1993.

[3] L. Hodges and E. Davi, "Geometric consideration for stereoscopic virtual environment," *Presence*, vol. 2, no. 1, 1993.

[4] B. Julesz and B., "Foundations of cyclopean perception," *Uni of Chicago Press.*, 1971.

[5] L. Lipton, *Foundations of the Stereoscopic Cinema A Study in Depth.* Van Nostrand Reinhold Company, 1982.

[6] L. Lipton, *CrystalEyes handbook.* StereoGraphics Cooperation, 1991.

[7] S. Pastoor, "3d-television: A survey of recent research results on subjective requirements," *Signal Processing: Image Communication*, 1991.

[8] C. Schor and I. Wood, "Disparity range for local stereopsis as a function of luminance spatial frequence," *Vision Res.*, pp. 1649–54, 1983.

[9] A. Woods, T. Docherty, and R. Koc, "Image distortions in stereoscopic video syetem," in *Stereoscopic Display and Applications IV*, 1993.

[10] Y.-Y. Yeh and L. Silverstein, "Using electronic stereoscopic color display: Limits of fusion and depth discrimination," in *Tree-Dimensional Visualization and Display Technologies*, Proceeding of the SPIE, 1989.