2. Stereoscopic digital video

A stereoscopic digital video signal consists of two sequences of digital images. One for the left eye and one for the right eye. Digital images are briefly introduced in Section 2.1.

A camera for stereoscopic video capture consists of two monoscopic cameras. The two basic configurations for the arrangement of two such cameras, parallel and converging, are discussed in Sections 2.2.1 and 2.2.2. In Section 2.2.3 the concept of the epipolar line is introduced, completing the model for stereoscopic video capture to the point where relevant numerical calculations can be made for any particular camera configuration. An example of such calculation is given in Section 2.3. The subject of Section 2.4 is occlusion in stereoscopic image pairs.

Stereoscopic display or projection in general is discussed in Section 2.5 along with the implications of displaying material recorded by either parallel or converging camera configurations. Techniques for the actual realisation of such a display are discussed in Section 2.6.

Because of differences between the two camera's making up the stereoscopic camera left and right view image (sequences) demonstrate differing grey-scale statistics which must be compensated for in the luminance signal in order for disparity estimation to be successful. This is discussed in section 2.7.

Finally section 2.8 contains a description of the adaptation of the MPEG-2 video coding scheme to stereoscopic video: the DISTIMA S-MPEG codec.

2.1 Digital image nomenclature

Digital images consist of a discrete number of so called picture elements, usually referred to as pixels. These pixels are arranged in a rectangular grid as illustrated in Figure 2.1. This grid has a width of W_{pix} and a height of H_{pix}. So that the total number of pixels is W_{pix} x H_{pix}.

For each pixel luminance and, usually, chrominance data is stored. This data is quantised and stored in a digital fashion. In the typical (4:2:2) luminance/chrominance (Y/UV) storage scheme 8 bits are dedicated to the luminance component for each pixel while 8 bits are allocated for each of the two components of the chrominance signal for two adjacent pixels. The net number of bits per pixel (bpp) for this scheme is 8+2x8/2=16. For all of the image sequence material referred to in this work the luminance component was stored in 8 bits so that it could take one of 2^8=256 discrete values.

The pixel belonging to frame t from a video sequence s at co-ordinate (x,y) is denoted as $s_t(x,y)$.

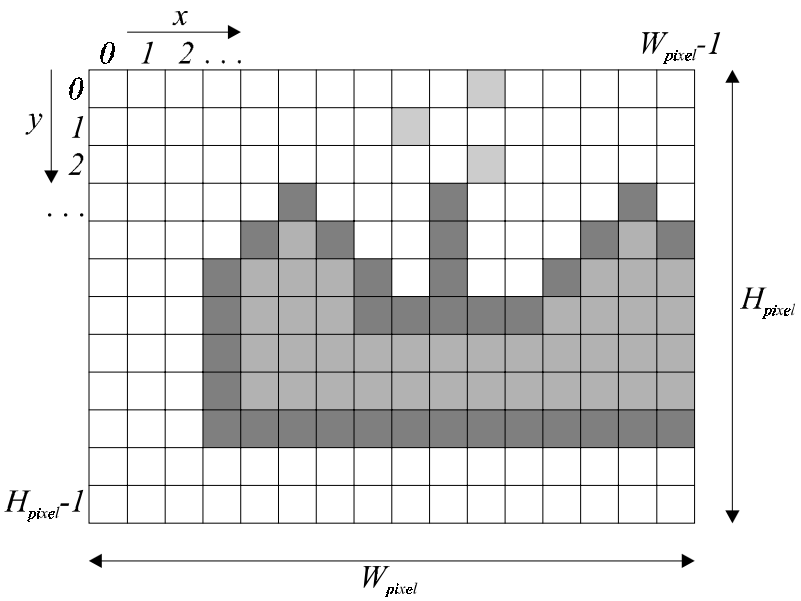

Figure 2.1: Digital image pixel grid.

All the material from the DISTIMA project which was used in this research had luminance and chrominance stored at a 4:2:2 ratio as described above. Image dimensions are W_{pix}=720, H_{pix}=576 at 25 frames per second per video channel. There are two video channels: left and right.

Each frame consists of two interlaced fields as illustrated in Figure 2.2. These fields

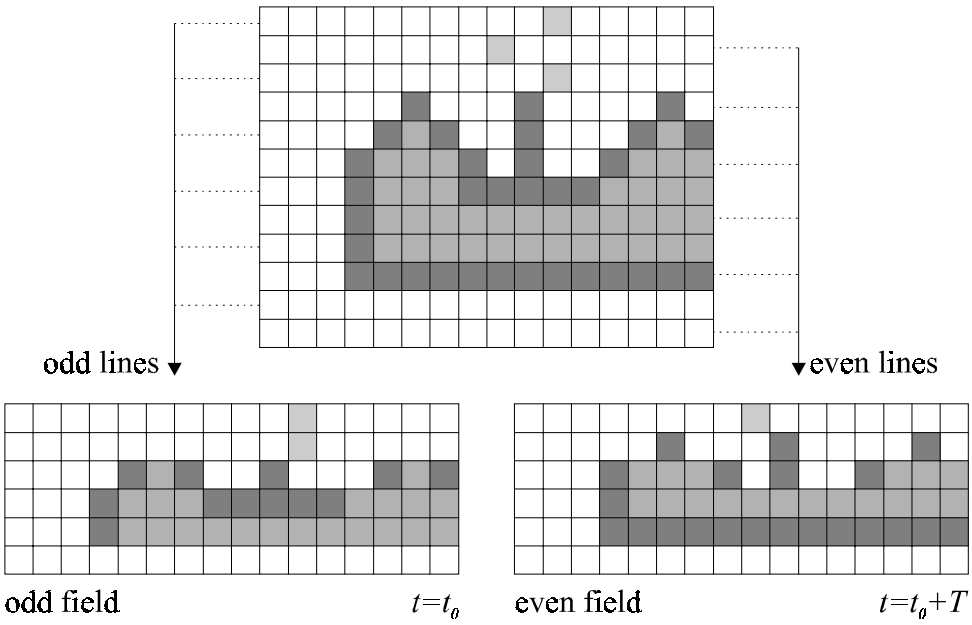

Figure 2.2: Decomposition of a frame into fields.

correspond to instances in time 1/50th of a second apart so that a frame actually contains data from two instances in time. The fields are displayed time sequentially at 50 fields per second at their proper spatial positions.

2.2 Camera geometry

A stereoscopic video camera is a co-planar configuration of two monoscopic video cameras. Within such a co-planar configuration two sub configurations may be distinguished. The parallel camera configuration where the optical axes of both cameras run parallel and the converging camera configuration where the optical axes of the cameras intersect at a convergence point. The parallel camera configuration is in fact a degenerate version of the converging configuration with an infinitely distant convergence point.

2.2.1 Parallel cameras

The parallel camera configuration is illustrated in Figure 2.3. *X*,*Y* and *Z* denote real world co-ordinates of the real word point *w* which is being observed. x_l and y_l denote the co-ordinates of the projection of this point onto the left camera sensor

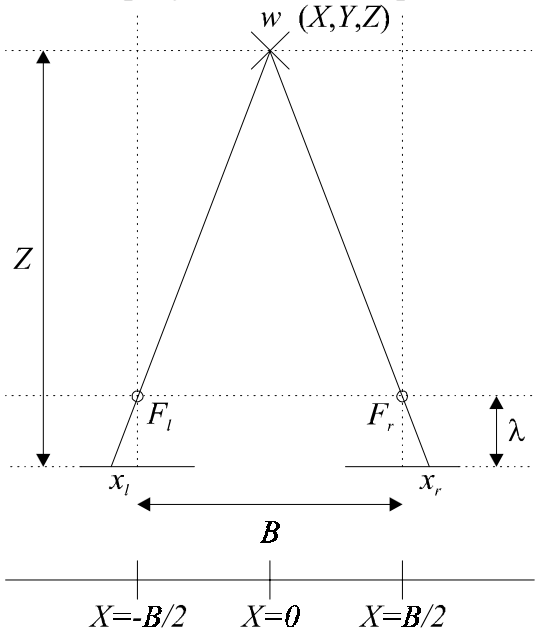

Figure 2.3: Parallel camera geometry.

and x_r, y_r denote the co-ordinates of the projection onto the right camera sensor. F_l and F_r are the focal points of the left and right camera lenses respectively. The horizontal distance, *B*, between the two cameras is referred to as the *baseline*. The image capture process for each of the two cameras now consists of a translation along the *X* axis followed by a perspective transformation [Gonz92]. The perspective transformation has a single parameter, λ, the focal length.

The camera projection co-ordinates, ordinates x_l, y_l, x_r and y_r, as a function of real world co-ordinates are given by:

$$x_l(X,Z) = \lambda \frac{X + \frac{B}{2}}{\lambda - Z} \ , \qquad (2.1)$$

$$y_l(Y,Z) = \lambda \frac{Y}{\lambda - Z} \ , \tag{2.2}$$

$$x_r(X,Z) = \lambda \frac{X - \frac{B}{2}}{\lambda - Z} \ , \tag{2.3}$$

$$y_r(Y,Z) = \lambda \frac{Y}{\lambda - Z} \ . \tag{2.4}$$

The unit of measure for the above equations is metres. Inspection of the equations shows that a real world point $(X,Y,Z)^T$ is projected onto the same y co-ordinate in both left and right cameras but that the x components of the projections differ. The amount by which the projection of a point in the right camera is shifted to the right with respect to the projection of the same point in the left camera is the horizontal *disparity*, $d_{p,h}(Z)$:

$$d_{p,h}(Z) = x_r(X,Z) - x_l(\mathrm{X},\mathrm{Z}) \tag{2.5}$$

Substituting (2.1) and (2.3) into (2.5) we get:

$$d_{p,h}(Z) = \lambda \frac{-B}{\lambda - Z} \ . \tag{2.6}$$

This function is illustrated in Figure 2.4 for a camera configuration with parameters as shown in Table 2.1.

Table 2.1: Example parallel camera parameters.

Camera parameter	Value
B	0.100 m.
λ	0.050 m.

As Z nears λ from below or above $d_{p,h}(Z)$ approaches minus or plus infinity respectively. Also for points positioned infinitely far away along the Z axis we note that:

$$\lim_{Z \to \infty} d_{p,h}(Z) = 0 \tag{2.7}$$

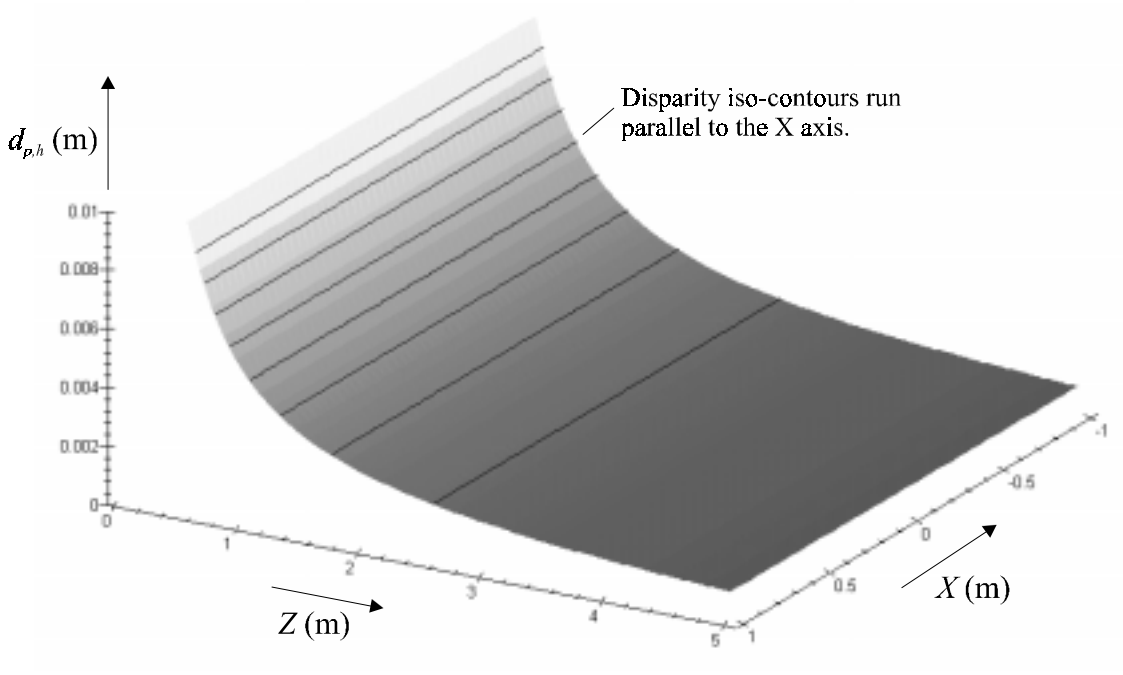

Figure 2.4: Disparity for sample parallel camera configuration.

2.2.2 Converging cameras.

In a converging camera configuration the optical axes of the two cameras converge at a convergence point:

$$Z_{conv} = \frac{B}{2\tan(\beta)} \ . \tag{2.8}$$

The set-up is illustrated in Figure 2.5.The projection onto the camera imaging plane is a little more complex than for the parallel camera situation (in fact the parallel camera configuration is a degenerate version of the converging camera configuration). Because the camera optical axes are now no longer parallel to the *Z* axis projection on the imaging plane now consists of horizontal translation followed by rotation and perspective projection. For the left camera translation moves the real world point (*X*,*Y*,*Z*) to:

$$\begin{bmatrix} X \\ Y \\ Z \end{bmatrix} \xrightarrow{translation} \begin{bmatrix} X + \frac{B}{2} \\ Y \\ Z \end{bmatrix} . \tag{2.9}$$

Rotation around the Y axis by an angle of +β of the translated point gives:

$$\begin{bmatrix} X+\frac{B}{2} \\ Y \\ Z \end{bmatrix} \xrightarrow{rotation} \begin{bmatrix} \cos(\beta)(X+\frac{B}{2})-\sin(\beta)Z \\ Y \\ \sin(\beta)(X+\frac{B}{2})+\cos(\beta)Z \end{bmatrix}. \tag{2.10}$$

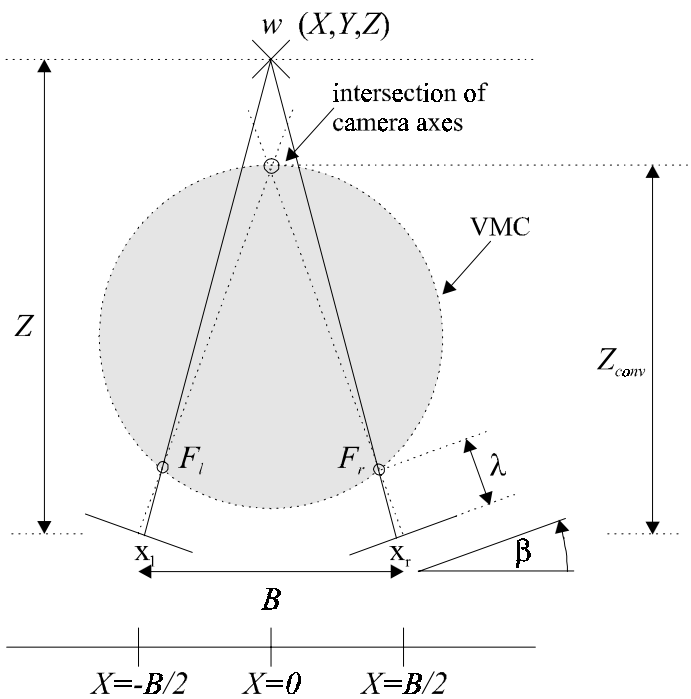

Figure 2.5: Converging camera geometry.

Finally perspective projection [Gonz92] is performed giving x_l and y_l.

$$x_l(X,Z) = \lambda \frac{\cos(\beta)(X+\frac{B}{2})-\sin(\beta)Z}{\lambda-\sin(\beta)(X+\frac{B}{2})-\cos(\beta)Z}, \tag{2.11}$$

$$y_l(X,Y,Z) = \lambda \frac{Y}{\lambda-\sin(\beta)(X+\frac{B}{2})-\cos(\beta)Z}. \tag{2.12}$$

Similarly x_r and y_r can be found:

$$x_r(X,Z) = \lambda \frac{\cos(\beta)(X-\frac{B}{2})+\sin(\beta)Z}{\lambda+\sin(\beta)(X-\frac{B}{2})-\cos(\beta)Z}, \tag{2.13}$$

$$y_r(X,Y,Z) = \lambda \frac{Y}{\lambda + \sin(\beta)(X - \frac{B}{2}) - \cos(\beta)Z} \ . \tag{2.14}$$

Given these expressions it is possible to determine the horizontal disparity for the converging camera configuration:

$$d_{c,h}(X,Z) = x_r(X,Z) - x_l(X,Z) = -2\lambda \frac{N_d(X,Z)}{D_d(X,Z)} \ . \tag{2.15}$$

where:

$$\begin{aligned} N_d &= 4\sin(\beta)\cos(\beta)Z^2 + \left\{2B - 4\lambda\sin(\beta) - 4B\cos^2(\beta)\right\}Z \\ &\quad + B\cos(\beta)\{2\lambda - B\sin(\beta)\} + 4\sin(\beta)\cos(\beta)X^2 \end{aligned} \tag{2.16}$$

and

$$\begin{aligned} D_{d=} &4\cos^2(\beta)Z^2 + 4\cos(\beta)\{B\sin(\beta) - 2\lambda\}Z \\ &+ \left\{4\lambda^2 - 4\lambda B\sin(\beta) + B^2\sin^2(\beta)\right\} - 4\sin^2(\beta)X^2 \ . \end{aligned} \tag{2.17}$$

For β=0 the converging camera configuration reverts to the parallel camera configuration as we can confirm by substituting β=0 into (2.15):

$$d_c(X,Z) = -2\lambda \frac{-2BZ + 2\lambda B}{4Z^2 - 8\lambda Z + 4\lambda^2} = -2\lambda \frac{2B(\lambda - Z)}{4(\lambda - Z)^2} = -\lambda \frac{B}{\lambda - Z} \ . \tag{2.18}$$

The points in the X-Z plane which have 0 disparity, $d_c(X,Z)=0$ lie on the so called Vieth-Müller circle [Lipton91] with centre:

$$\begin{bmatrix} X_c \\ Z_c \end{bmatrix} = \begin{bmatrix} 0 \\ \dfrac{B\cos(2\beta) + 2\lambda\sin(\beta)}{2\sin(2\beta)} \end{bmatrix} \tag{2.19}$$

and radius

$$r_{VMC} = \frac{B - 2\lambda\sin(\beta)}{2\sin(2\beta)} \qquad B > 2\lambda\sin(\beta) \ . \tag{2.20}$$

The Vieth-Müller circle (abbreviated as VMC) is marked in Figure 2.5. It is in fact a circle which passes through the focal points of left and right camera as well as through the intersection of the optical axes, Z_{conv}. The condition that $B>2\lambda \sin(\beta)$ means that the focal point of the right camera may not lie to the left of the focal point of the left camera. The far intersection of the Vieth-Müller circle and the Z axis is given by:

$$Z_{far} = Z_c + r_{VMC} = \frac{B(1+\cos(2\beta))}{2\sin(2\beta)} = \frac{B}{2\tan(\beta)} \ . \tag{2.21}$$

Which means that this farthest point on the Vieth-Müller circle coincides with the convergence distance Z_{conv} of the parallel camera configuration:

As Z approaches infinity the disparity approaches a limit:

$$\lim_{z\to\infty} d_{c,h}(X,Z) = -2\lambda\tan(\beta) \ . \tag{2.22}$$

In Figure 2.6 disparity is plotted for a camera configuration with the example parameters given in Table 2.2.

Table 2.2: Example converging camera parameters.

Camera parameter	Value
B	0.100 m.
λ	0.050 m.
β	1°

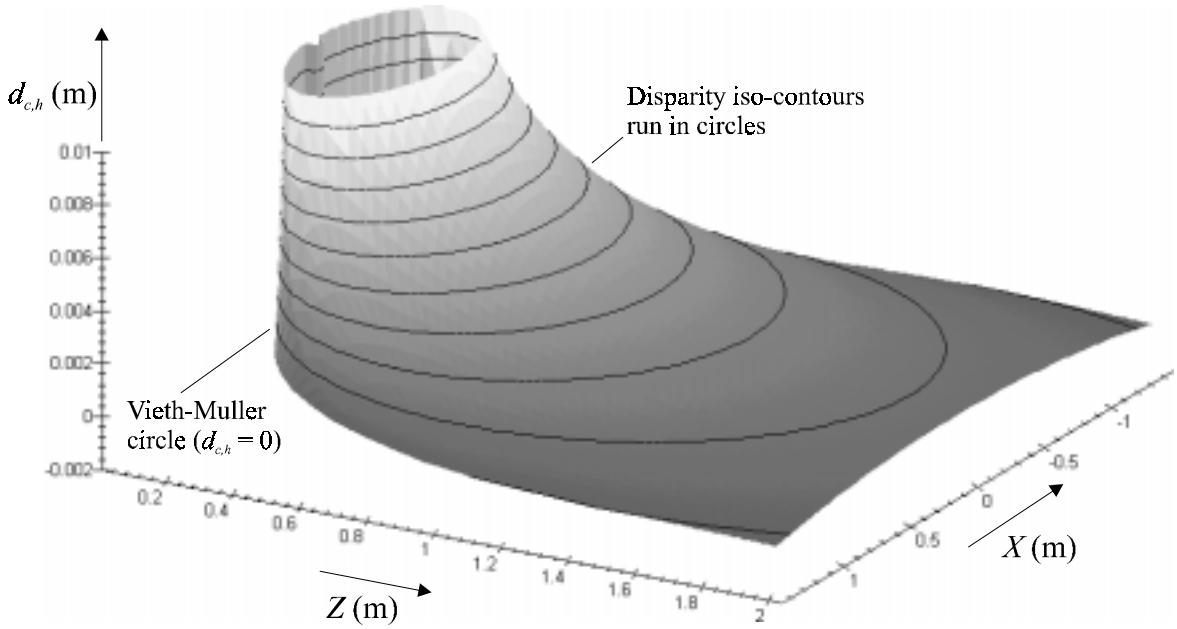

Figure 2.6: Disparity for sample converging camera configuration.

Unlike for parallel cameras a converging camera configuration also results in a vertical disparity component:

$$d_{c,v}(X,Y,Z) = y_r(X,Y,Z) - y_l(X,Y,Z) = -8\lambda\frac{XY\sin(\beta)}{D_{d,v}(X,Y,Z)} \tag{2.23}$$

where

$$D_{d,v}(X,Y,Z) = 4\lambda\{\lambda - B\sin(\beta) - 2Z\cos(\beta)\} + \sin^2(\beta)\{B^2 - 4X^2\} + 2BZ\sin(2\beta) + 4\cos^2(\beta)Z^2. \quad (2.24)$$

Inspection of (2.23) and (2.24) shows that the vertical disparity component is equal to zero in the plane defined by X=0 as well as the plane defined by Y=0. Outside of these two planes the vertical disparity is not equal to zero. The magnitude of the vertical disparity decreases with increasing Z.

2.2.3 Epipolar lines.

The set of points which projects onto $P = (x_p,y_p)$ on the left camera sensor is given by the line passing through the point (x_p,y_p) on the sensor and the focal point F_l of the left lens. This line will appear on the right camera sensor as another line, the so-called *epipolar* line. It is along this line that we should look when trying to find a match for the point at $(x_p,y_p)^T$, when estimating disparity.

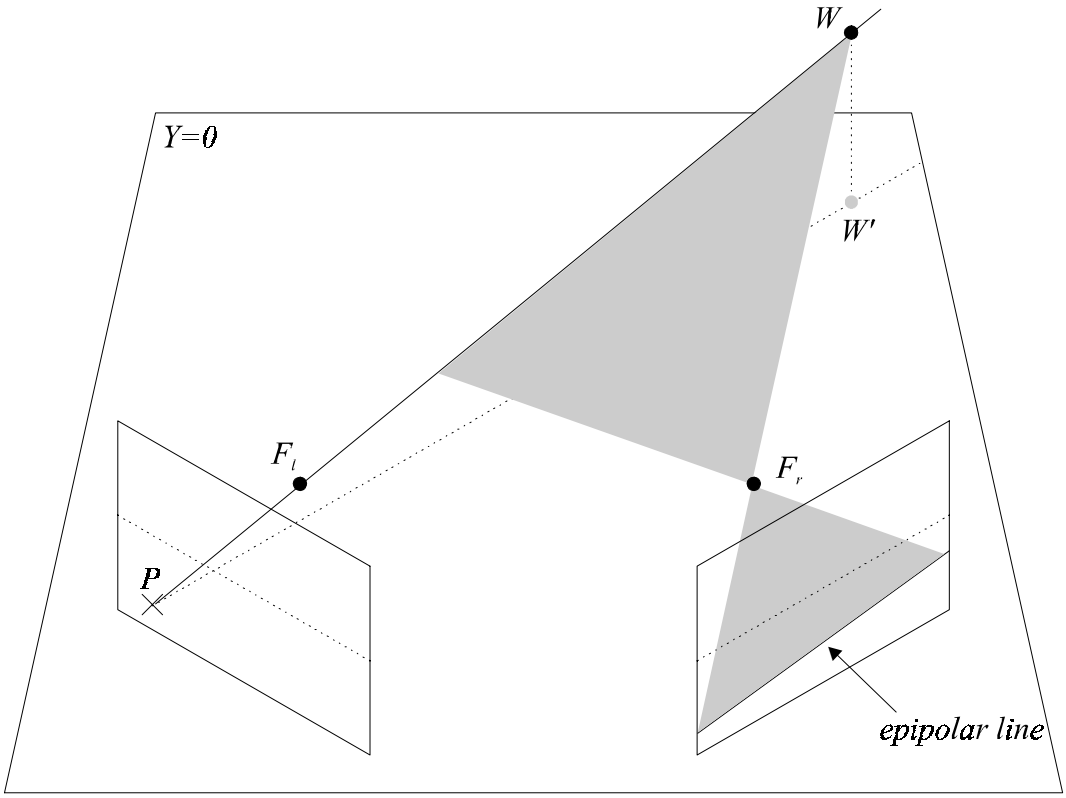

Figure 2.7: Epipolar line in converging camera configuration.

The idea is illustrated in Figure 2.7. The real world point W (located at a positive Y, above the Y=0 plane) projected through F_l onto the point P on the left image plane. Any of the points on the line which connects W, F_l and P would project onto P. When this line is projected onto the right image plane (through the right focal point F_R) the result is the epipolar line corresponding to P.

So estimating disparity, looking for matches along an epipolar line corresponds to depth estimation in the real world, as will be clear from the above figure.

The right image epipolar line corresponding to (x_p,y_p) in the left image is given by:

$$y_r = a_{epi} \cdot x_r + b_{epi} \tag{2.25}$$

where:

$$a_{epi}(x_p, y_p) = -\frac{\sin(\beta)\left(\sin(\beta)x_p - \cos(\beta)\lambda\right)y_p}{\cos(\beta)^2\lambda^2 - \sin(\beta)^2 x_p{}^2} \tag{2.26}$$

and

$$b_{epi}(x_p, y_p) = \frac{\left(\cos(\beta)\lambda - \sin(\beta)x_p\right)\cos(\beta)\lambda \cdot y_p}{\cos(\beta)^2\lambda^2 - \sin(\beta)^2 x_p{}^2}. \tag{2.27}$$

The epipolar lines intersect on the $y_r = 0$ axis (in the right image) at the so called epipolar point:

$$epi_R = \frac{\lambda}{\tan(\beta)}. \tag{2.28}$$

This epipolar point is common to all epipolar lines. Therefore it must be the projection of a point in the real world which lies on all possible lines of projection for the left camera. There is only one such point: epi_R is the projection onto the right camera image plane of the focal point of the left lens. Similarly epi_L is the projection onto the left camera plane of the focal point of the right lens.

Inspection of (2.26) and (2.27) shows that epipolar lines run horizontally, i.e. $a_{epi}=0$, for $\beta=0$ or $y_p=0$. As was already mentioned before disparity is a purely horizontal phenomenon only for the parallel camera configuration: all epipolar lines have slope zero. For small β the value of a_{epi} will be such that the epipolar lines will only slightly deviate from the horizontal in the image corners.

2.3 Camera parameters for horizontal disparity

Digital images have only finite resolution. For converging camera configurations where β is not too large the maximum value of a_{epi} is such that the maximum occurring vertical disparity is smaller than half the height of such a pixel so that vertical disparity will have no significant effect on the image captured and disparity may be modelled as a purely horizontal phenomenon.

Parameters for a practical example converging camera configuration [Pana95] are given in Table 2.3. The convergence distance for this camera configuration is:

$$Z_{conv} = \frac{B}{2\tan(\beta)} = \frac{0.100\ \text{m}}{2\tan(1)} \approx 2.864\ \text{m}.$$

Table 2.3: Parameters for example converging camera configuration.

Name	Value	Description
B	0.100 m	baseline
β	1°	convergence angle
λ	0.050 m	focal length
W_{ccd}	10.16 mm	width of CCD
H_{ccd}	7.62 mm	height of CCD
W_{pix}	752	number of pixels along horizontal
H_{pix}	582	number of pixels along vertical

Using (2.26) we evaluate the maximum value for a_{epi}, in the image corner:

$$a_{epi,\max} = a_{epi}\left(\frac{-10.16\ \text{mm}}{2}, \frac{7.62\ \text{mm}}{2}\right) \approx -0.00133.$$

The maximum disparity of interest is the disparity of far away objects. Objects very near to the camera could theoretically have a greater magnitude of disparity (but with a different sign) but in well composed stereo scenes this will not occur and the greatest magnitude of disparity can be calculated using (2.22):

$$d_{\max,h} = -2\lambda\tan(\beta) = -2\cdot 0.050\ \text{m}\cdot\tan(1) \approx -1.746\ \text{mm}.$$

So the maximum vertical disparity becomes:

$$d_{c,v,\max} = d_{c,h,\max}\cdot a_{epi,\max} \approx -1.746\ \text{mm}\cdot -0.00133 \approx 0.00232\ \text{mm}.$$

We want to compare this maximum vertical disparity to the height of a pixel:

$$h = \frac{H_{ccd}}{H_{pixel}} = \frac{7.62\ \text{mm}}{582} \approx 0.0131\ \text{mm}.$$

So for this example camera configuration the maximum vertical disparity is almost 6 times smaller than the pixel height. Stereoscopic video material captured by this camera configuration can be modelled as having purely horizontal disparity. Since all the stereoscopic video reference material used in this work has similarly negligible vertical disparity we will model disparity as being a strictly horizontal phenomenon.

The two factors which influence the slant of the epipolar lines are the convergence angle, β, and the focal length, λ. For practical focal lengths it will be the

convergence angle which determines whether the (worst case) slant of the epipolar lines is non-negligible or not. For most applications where β lies in the neighbourhood of 1° or so the worst case epipolar slant will be negligible as in the example above. However, a stereoscopic video-conferencing application where a participant is sitting in front of a monitor with camera's attached at either side of this monitor can easily result in a β of 30° or so thus forcing image rectification [Papa95] or modelling of disparity as a non horizontal phenomenon.

2.4 Occlusion

Assuming limited β so that the disparity shift can be modelled as horizontal it should be noted that this horizontal shift does not hold for all pixels in left and right

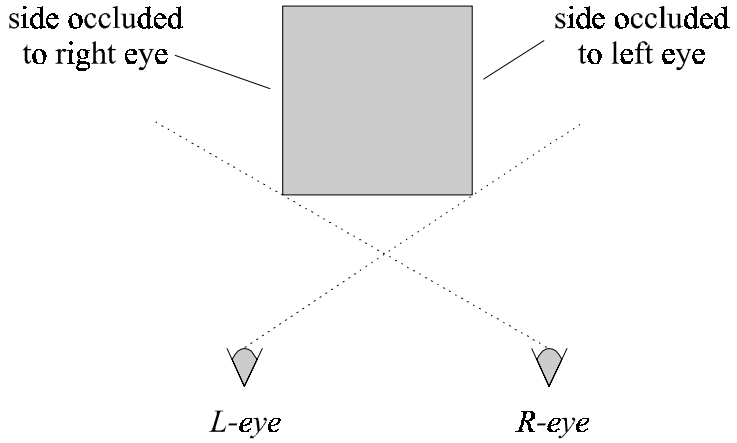

Figure 2.8: Occlusion.

view images some pixels will be occluded in either left or right view. The effect is illustrated in Figure 2.8 where the front aspect of the square object is visible in both views. The left side of the object however, is visible only in the left view and occluded in the right view. Conversely the right side of the object is visible only in the right view and occluded in the left view. As occluded pixels are present only in either left or right view they cannot be linked by disparity, whether it be horizontal or not.

2.5 Display of recorded stereoscopic video

Stereoscopic display on a flat screen is shown schematically in Figure 2.9. Here the horizontal displacement of the projection of the right eye view of a point with regard to the projection of the left eye view of that same point is referred to as parallax and is denoted by p. The observed depth Z_{obs} is the sum of the distance from the observer to the screen, Z_s, and an offset term Z_m. Similarity of triangles gives:

$$\frac{Z_m}{\left(\frac{p}{2}\right)} = \frac{Z_m + Z_s}{\left(\frac{E}{2}\right)} \Leftrightarrow Z_m = \frac{p \cdot Z_s}{E - d} \tag{2.29}$$

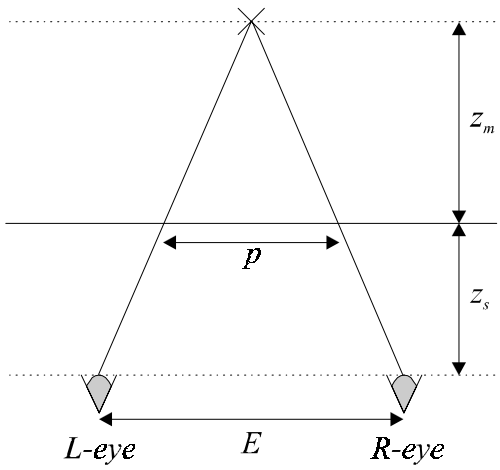

Figure 2.9: Stereoscopic projection.

So for Z_{obs} we get:

$$Z_{obs} = Z_s + Z_m = Z_s \frac{E}{E - p} \tag{2.30}$$

This function is illustrated in Figure 2.10.

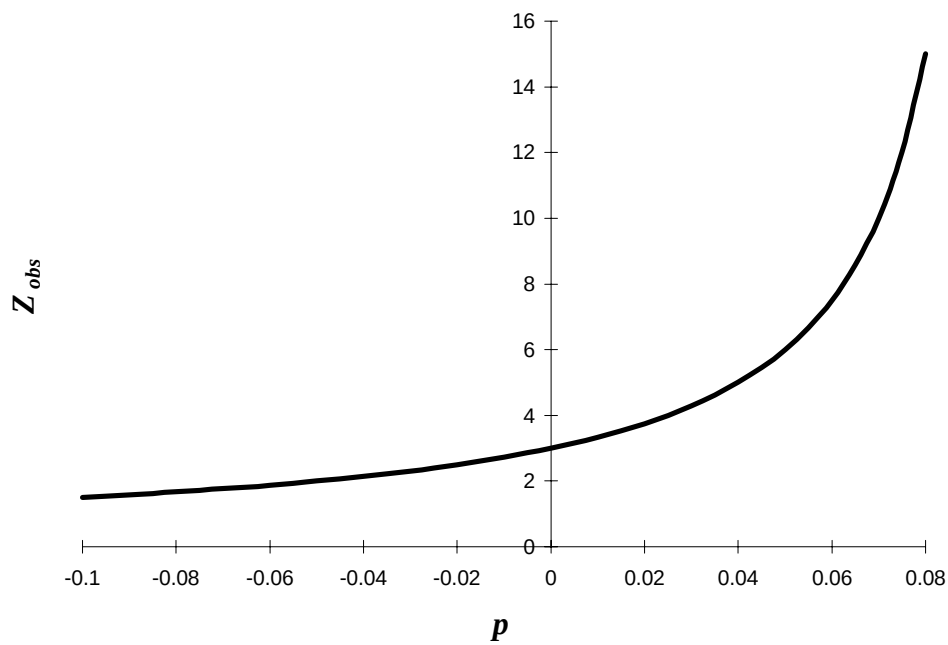

Figure 2.10: Z_{obs} vs. p, Z_s=3, E=0.10

For p = 0 the perceived distance is equal to the screen distance Z_s. For negative p Z_p goes to zero. It should be noted that as p becomes more negative the observer will increasingly have to look in a cross-eyed fashion and binocular fusion will become

uncomfortable and subsequently impossible. The highest sensible value for p is equal to the eye separation, E. When $p=E$ both of the observers eyes are looking straight ahead which corresponds to looking at a point which is infinitely far away. In fact:

$$\lim_{p \uparrow E} Z_{obs} = \infty \tag{2.31}$$

The relation between parallax and camera disparity is a simple one. Parallax has the opposite sign of disparity because projection onto the camera's image plane happens in mirror image. The magnitude of parallax is magnified by a factor M which is the proportion of (projection) screen width to camera image plane width. Also most displays allow the introduction of an additional offset term onto the parallax [Kutk94, Lipt91]. This term is commonly referred to as the horizontal image translation, H. Thus:

$$p_p(Z) = -M \cdot d_p(Z) + H \,. \tag{2.32}$$

Assuming a parallel camera configuration and substituting (2.6) into (2.32) we obtain:

$$p_p(Z) = \frac{M\lambda B}{\lambda - Z} + H \,. \tag{2.33}$$

Which can be substituted into (2.30) to give the relation between the observed distance Z_{obs} and the real world distance Z_w originally recorded by the stereoscopic camera:

$$Z_{obs}(Z_W) = Z_s \frac{E}{E - \dfrac{M\lambda B}{\lambda - Z_w} - H} \,. \tag{2.34}$$

Inspection of $Z_{obs}(Z_w)$ shows that, for the parallel camera configuration, all objects have a projected depth which locates them between the observer and the screen. Objects very near to the observer (small Z_w) are perceived to be disproportionately near while objects which are infinitely far away (Z_w very large) are perceived to be at the same distance as the projection screen itself (Z_s) when H equals zero. A positive H makes far objects look further away while a negative H makes them look nearer:

$$\lim_{z_w \uparrow \infty} Z_{obs}(Z_w) = Z_s \frac{E}{E - H}, \quad H < E \,. \tag{2.35}$$

Obviously increasing H beyond E would not be sensible as this would increase the zero parallax to a point where the observer's eyes can no longer converge and stereo perception becomes impossible. Even values of H which are smaller than E but near to it may already result in uncomfortable viewing.

For a converging camera configuration the situation is analogous. Now:

$$\lim_{z_w \uparrow \infty} Z_{obs,conv}(Z_w) = Z_s \frac{\cos(\beta) \cdot E}{\cos(\beta) \cdot (E - H) - 2\sin(\beta)\lambda \cdot M} \quad . \tag{2.36}$$

2.6 Display techniques

The display of binocular stereoscopic imagery requires that distinct and appropriate images are displayed for each of the observer's eyes. Several techniques which achieve this exist. What distinguishes these techniques is the method by which the left and right eye images are separately presented to the left and right eye of the viewer. If this separation is not perfect and the left eye perceives some of right eye signal and vice versa there will be crosstalk. Other factors which distinguish stereoscopic display techniques are:

- How many viewers can watch at the same time: one or more?
- Freedom: can the viewer move his head freely? Can the viewer walk about the room freely?
- Comfort: does the viewer have to wear cumbersome headgear?

2.6.1 The stereoscope

The classical *stereoscope* may be best known in its incarnation as the children's viewmaster toy. It is illustrated in Figure 2.11. The left and right eye views are stored on and displayed from photographic slides. Each slide is presented to the intended recipient eye through a lens system. The eyes/slides are separated by a

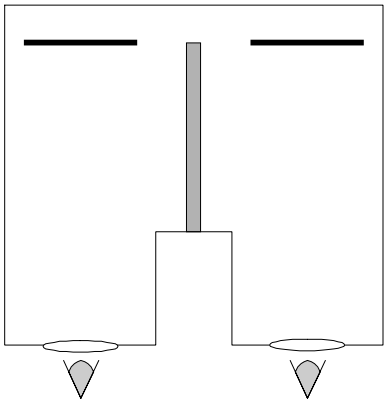

Figure 2.11: The stereoscope.

septum so that there is absolutely no crosstalk between left and right eye views.

Given well composed slides the stereoscopic effect is good. Unfortunately only one observer at a time can view the scene and this observer must hold the stereoscope, which can be relatively large, in front of his eyes while viewing.

2.6.2 The anaglyph method

Another popular and inexpensive stereoscopic display technique is the *anaglyph* method. This method is used both in print (still images) and on projection screens (sequences). Left and right views are reduced to monochromatic images. They are superimposed on the same display, the left image in the green component and the right image in the red component (or vice versa). The viewer wears a pair of glasses of which the right lens is red and the left lens is green. The green lens in front of the left eye lets the green image intended for the left eye pass while it blocks the red image intended for the right eye. The red lens in front of the right eye performs the complementary function.

This inexpensive system produces a good stereoscopic effect that can be viewed by multiple viewers (all wearing uncumbersome light glasses). If the projection and lens colours are well tuned then there is very little crosstalk between the left and right views. A severe limitation of this system is the fact that colour information is lost because colour is used as a selection mechanism. Also viewing of the real world, surrounding the display area, through red/green glasses can be tiring as strange perceptual effects may be experienced.

2.6.3 Polarised display

Another technique for separating left and right eye views involves the use of polarised light. Left and right view images are projected through perpendicular polarising filters. The viewer wears glasses with polarising lenses which let one image through to the left eye and the other to the right eye.

Left and right views may be displayed concurrently (i.e. they may be superimposed) or they may be time-multiplexed in a left - right - left - right fashion.

This kind of display is almost as simple as the anaglyph system. Multiple viewers may observe the same scene. The glasses which the viewer wears are light and uncumbersome. Unlike with red/green glasses there are no problems with perception of the area around the display. Also colour is fully maintained.

Because the selection mechanism depends on the relative orientation of polarised light and lenses it is quite sensitive to crosstalk. If the viewer inclines his head slightly crosstalk immediately results.

2.6.4 LCD shutterglasses

Image selection can also be performed with LCD shutterglasses. The idea is illustrated in Figure 2.12. Left and right view images are time sequentially multiplexed and projected in left - right - left -right order. The video synchronisation signal is passed to the shutter glasses worn by the viewer. When a left view image is projected the LCD in front of the right eye darkens while the left eye LCD

becomes translucent. Similarly when a right view image is projected the left LCD becomes dark and the right LCD becomes translucent.

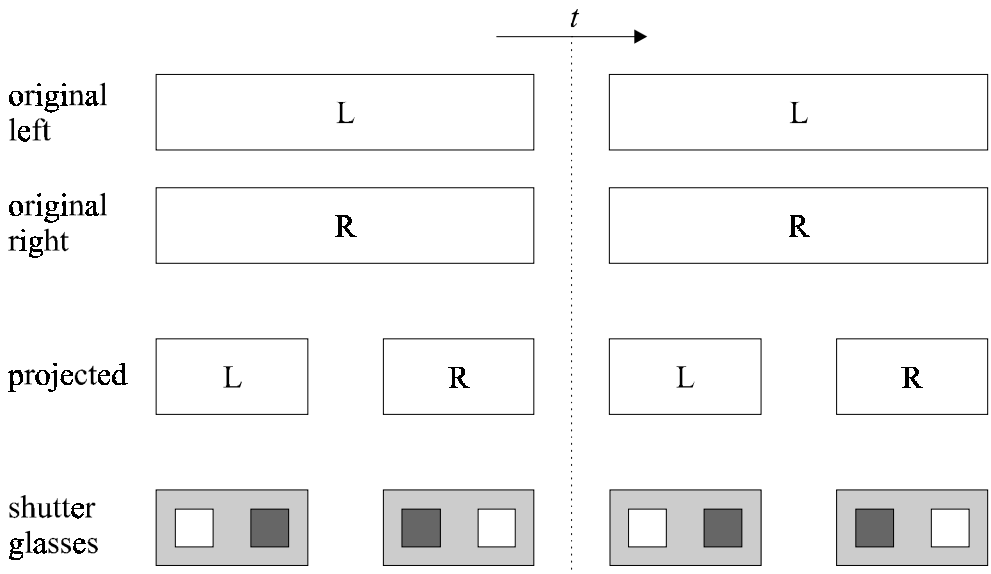

Figure 2.12: L/R image selection using shutterglasses .

Because of the time sequential multiplexing the monitor must run at double the frame rate of the original sequence material: to display a stereoscopic 50 Hz sequence by time multiplexing requires a 100 Hz display.

It is important that the transient characteristics of the LCDs be such that the light-dark and dark-light transitions are made sufficiently quickly, preferably within the field blanking time.

If the LCD speed is high enough the separation is good and there is little crosstalk. Colour is not lost and viewer position and/or head inclination has little or no effect on the separation. Multiple viewers may watch the same scene but they must all wear the LCD shutterglasses which are relatively large and heavy and also expensive.

2.6.5 Lenticular autostereoscopic displays

Stereoscopic display techniques which do not require the viewer to wear special glasses or headgear are known as autostereoscopic displays [Hama95, Raoul95]. The most prevalent autostereoscopic display techniques utilise lenticular screens. The principle an autostereoscopic display utilising such a lenticular screen is illustrated in Figure 2.13. Pixels for left and right view images are displayed contemporaneously in alternate columns. By optical lens effects of the lenticular lens sheet the columns displaying left view pixels are visible only to the left eye while the columns displaying right view pixels are visible only to the right eye [Börn93] without the viewer requiring glasses.

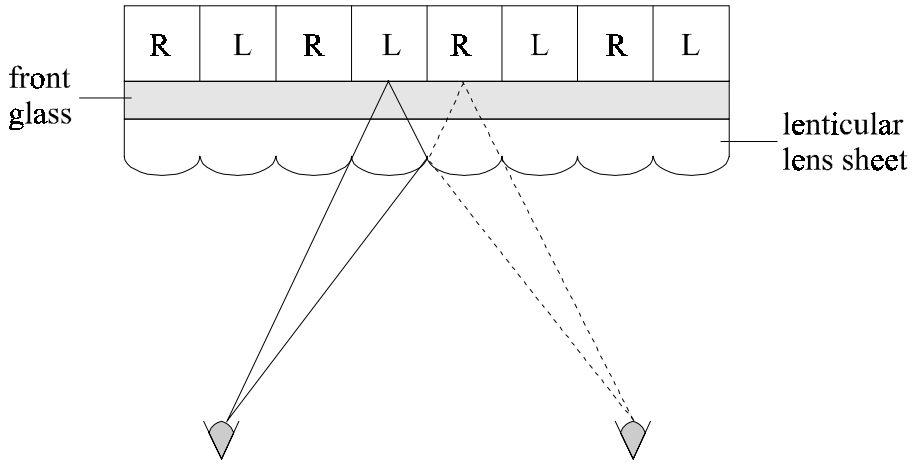

Figure 2.13: Lenticular display.

With the basic lenticular display technique the number of possible viewer positions will be limited and the horizontal image resolution is impaired because of the necessity of displaying left and right view pixels in alternate columns. More sophisticated versions of the lenticular display technique can avoid these problems [Raoul95].

2.7 Difference compensation - stereoscopic balancing.

The two component cameras of a stereoscopic camera are frequently distinct units: two different CCDs, two sets of camera electronics. Because of this there is a potential luminance signal difference between the images captured by the left and right cameras. This imbalance varies from scene to scene and with camera calibration.

Disparity estimation algorithms which commonly match luminance values can be seriously disturbed by this imbalance. Therefore it is necessary to compensate for it. In principle a disparity field would be needed to perform such a compensation because it is necessary to match objects in left and right views in order to analyse the camera imbalance. Because this difference actually hampers the disparity estimation itself a simple, disparity independent, compensation technique is required for an initial compensation.

A simple model for the imbalance between the two cameras is that of a scaling term, a, and an offset term b which relate the luminance pixels $s_r(x,y)$ in the right view image to the pixels $s_r(x+d(x,y),y)$ in the left view image. The disparity, $d(x,y)$ is still present in this model:

$$s_r(x,y) = a \cdot s_l(x + d(x,y), y) + b. \tag{2.37}$$

Taking the expectation of this expression while assuming s_l and s_r to be weakly stationary we get:

$$E\{s_l(x,y)\} = a \cdot E\{s_r(x + d(x,y), y)\} + b = a \cdot E\{s_r(x,y)\} + b. \tag{2.38}$$

And similarly the variance of s_r is a function of the variance of s_l :

$$\sigma_{rp}^2 = E\left\{(s_r(x,y) - E\{s_r(x,y)\})^2\right\} = E\left\{s_r^{\,2}(x,y)\right\} - E^2\left\{s_r(x,y)\right\}, \tag{2.39}$$

$$\sigma_{rp}^2 = a^2\left(E\left\{s_l^2(x,y)\right\} - E^2\{s_l(x,y)\}\right) = a^2\sigma_l^2. \tag{2.40}$$

So that the unknowns a and b can be calculated from the mean and variance of the left and right view images:

$$a = \frac{\sigma_{rp}}{\sigma_{lp}}, \tag{2.41}$$

$$b = E\left\{s_r(x,y)\right\} - \frac{\sigma_{rp}}{\sigma_{lp}} E\left\{s_l(x,y)\right\}. \tag{2.42}$$

Having calculated a and b it is possible to apply scaling and offset to all left view pixels so that mean and variance of left and right view images become equal.

The a and b for the first fields of some of the DISTIMA test sequences used in this work are given in Table 2.4:

Table 2.4: Balance parameters for DISTIMA test sequences.

Sequence Name	a	b
Aqua	1.015	0.794
Jardin	0.986	2.438
Manege	0.850	2.295
Piano	0.902	-1.052
Tunnel	0.971	5.331

Some of the sequence material used in this work exhibited luminance differences of such a magnitude that the performance of existing motion estimators used as disparity estimators broke down completely. Applying the compensation described above re-established the satisfactory function of these estimators.

On the test sequences listed in Table 2.4 there is no significant temporal variation in the a and b.

2.8 Stereoscopic hybrid video coding

Hybrid video coding is the basic video coding technique used by, for instance, the H261 and MPEG coding schemes. It is called hybrid video coding because two techniques for the removal of redundancy and irrelevancy are used:

- Predictions are made for images from the video stream using motion vector information. These predictions are subtracted from the corresponding original images. Where the prediction fails there will be significant content in the difference image which must be encoded and transmitted.
- Difference images or original images which are to be encoded and transmitted are subjected to the decorrelating DCT transform which makes enables the balanced distribution of bits according to visual relevancy.

The decoder essentially receives difference images and motion vector information. These are combined with the previous decoded image to form a predicted image to which the decoded difference signal is then added. In order to prevent error propagation a whole original image is periodically transmitted.

A first, most obvious way, of adapting the hybrid coding scheme to stereoscopic video sequences is to simply code them separately using two distinct hybrid codecs. This is illustrated schematically in Figure 2.14. Prediction is performed purely on the basis of motion vector information within each of the two sequences.

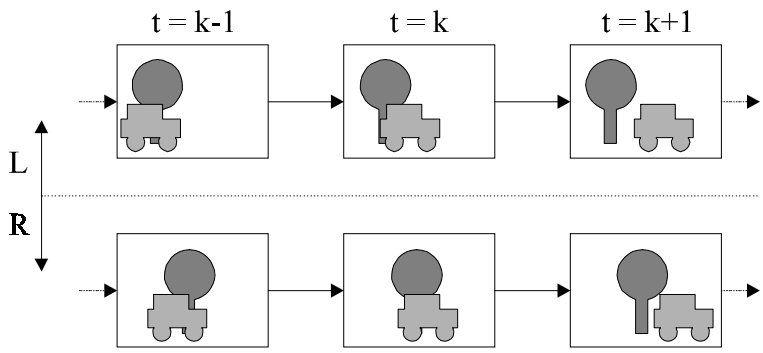

Figure 2.14: Temporal prediction.

The two resulting bitstreams can be decoded by conventional monoscopic decoders and viewed either monoscopically (downward compatibility) or displayed in a stereoscopic fashion. Because the independent coding errors which are made in the left sequence have no influence on the right sequence or vice versa this is also a robust way of performing hybrid stereoscopic video coding. A cluster of bit-errors causing a complete breakdown in either of the two sequences will have no influence whatsoever on the other sequence.

There is however one argument which counts against this way of coding stereoscopic video: the fact that the extra redundancy which is present in stereoscopic video is in no way exploited. No use is made of the similarity between contemporaneous left and right images. The bit-rate required for transmission of the two sequences is double the bit-rate required for the transmission of a single monoscopic sequence.

As left and right view images are recorded contemporaneously by closely spaced cameras the similarity between left and right view images could be greater than the

similarity between subsequent images from either sequence. This leads to a second prediction scheme as illustrated in Figure 2.15. The left sequence is still coded using a conventional codec and can be decoded for compatible monoscopic viewing. The right sequence is coded based on predictions made from contemporaneous left view images and can only be decoded using an appropriate stereoscopic decoder.

The disadvantage of this technique lies in the fact that predictions for right view images made from left view images (using block matching techniques common to hybrid codecs) are simply not as good as temporal predictions made purely from

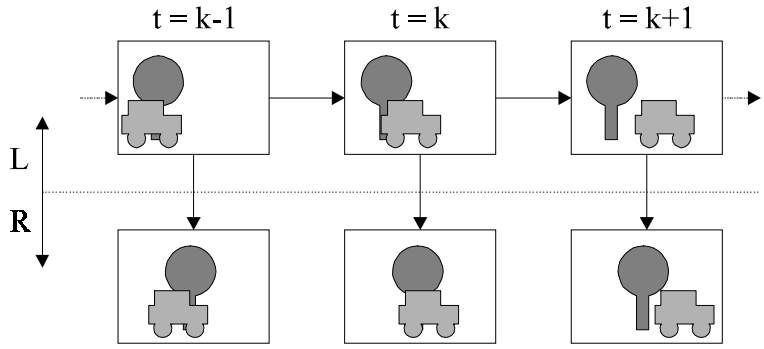

Figure 2.15: Disparity prediction.

previous right view images. Instead of making a bit-rate saving (or quality gain) a bit-rate overhead (or quality loss) is incurred. Therefore this is not a good way to adapt the hybrid coding scheme to stereoscopic video.

A third approach to hybrid coding scheme for stereoscopic video is illustrated in Figure 2.16. The left sequence is again coded using a conventional monoscopic coder thus giving monoscopic downward compatibility for the left channel. The

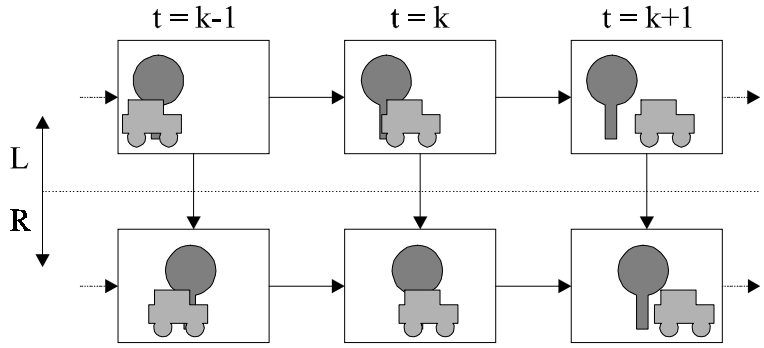

Figure 2.16: Combined motion/disparity prediction.

right sequence is coded using predicted images which are partially based on disparity (contemporaneous left view) or motion (other right views). The choice between motion and disparity prediction is made on a block by block basis. An attempted prediction is made for each block using both motion and disparity. Whichever of the two has the lowest error is incorporated into the composite predicted right view image.

In the graph of Figure 2.17 right channel prediction errors for motion prediction, disparity prediction and combined motion/disparity prediction are compared. The prediction error measure used is MSE (mean squared error). The MSE between an original image s_o and a predicted image s_p is given by:

$$MSE = \frac{\sum_{x=0}^{W_{pix}-1} \sum_{y=0}^{H_{pix}-1} \left[s_o(x,y) - s_p(x,y) \right]^2}{W_{pix} \cdot H_{pix}} \tag{2.43}$$

The material used was the first 20 frame images of the DISTIMA test sequence "train". The temporal predictions are made from previous original images while the disparity predictions are made from contemporaneous left channel original images.

Inspection of the graph shows that initially the prediction error for motion based prediction is lower than the prediction error for disparity based prediction. At frame 6 a moving train enter the foreground of the scene. This fast moving train causes the prediction error for motion prediction to rise to the level of the prediction error for disparity based prediction, sometimes even exceeding it.

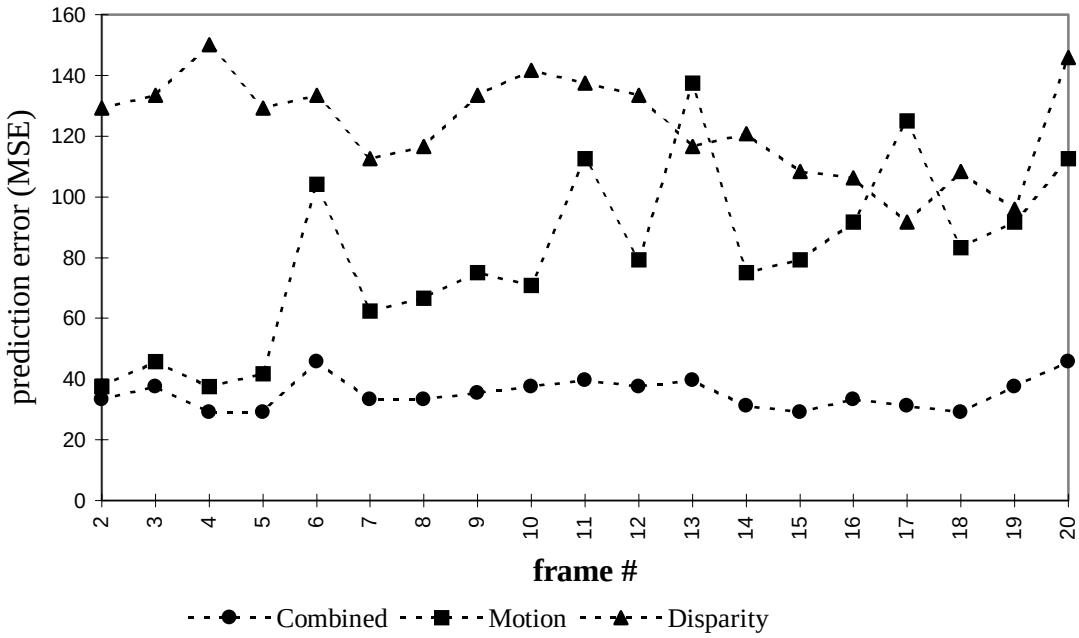

Figure 2.17: Comparison of prediction errors for sequence "train".

The disparity based prediction error is not sensitive to the change in motion in the scene as time and speed have no influence on disparity estimation which works on two images recorded at the same instant in time. For instance: the large jump in the motion based prediction error which occurs from frame 5 to frame 6 is caused by motion occlusion. In frame 6 there is a train wagon present which was not present in frame 5 therefore no good temporal prediction can be made for this wagon. In disparity this problem does not exist. The train wagon which is suddenly present in

the right channel frame 6 is also present in the left channel frame 6. Disparity based prediction for the train wagon is possible.

Combined prediction results in a lower error figure (for the whole frame) than either motion or disparity based prediction because blocks with the lowest error from the motion and disparity predicted frames are taken and combined to form the combined prediction. At the very worst this would result in a prediction error which is equal to the lowest error from the motion/disparity prediction pair. In practice glaring failures motion/disparity prediction are local and if one of the two fails the other still gives a good prediction (with a low squared error). The result is a predicted image which has few errors except for those inherent in working with 8x8 blocks. Also the prediction error for combined prediction is more stable in time than either motion or disparity based prediction.

The kind of codec which results is shown in Figure 2.18. This is the kind of codec realised for the hardware chain of the RACE-DISTIMA project. The DISTIMA codec [Fra92a,Fra92b] was in fact based on the MPEG-2 hybrid coding scheme.

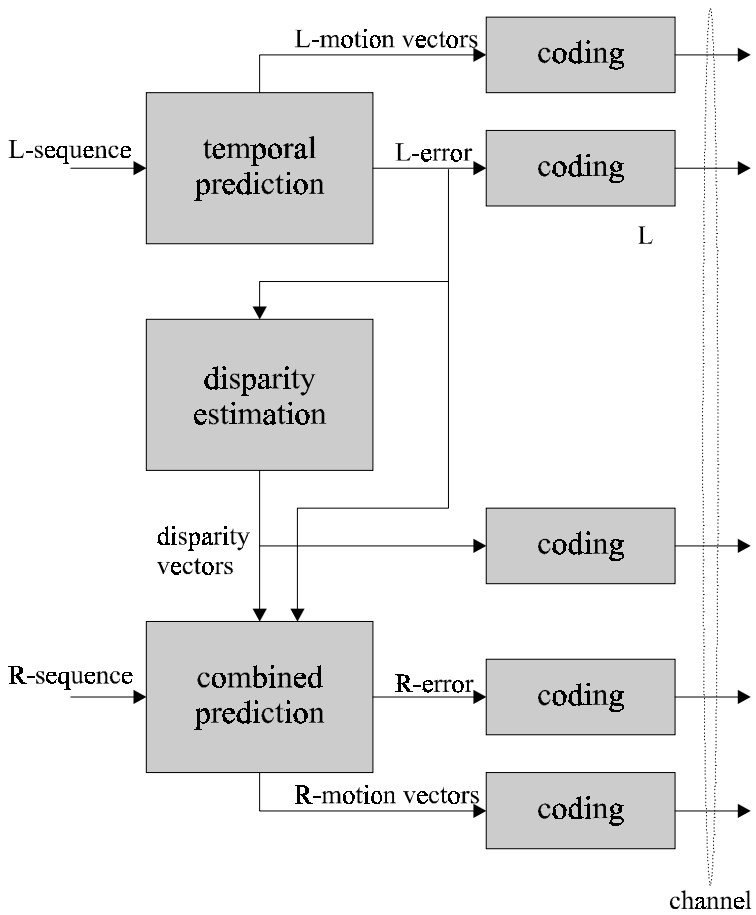

Figure 2.18: Stereoscopic hybrid video codec.