

Emeryville Camera Setup

Daniel Lyddy (daniell@cs.berkeley.edu)

December 8, 1997

Contents

| | | |
|----------|--|-----------|
| 1 | Introduction | 4 |
| 2 | Executive Summary | 4 |
| 2.1 | Camera Setup | 4 |
| 2.2 | Cost Estimate | 4 |
| 3 | Site Layout | 4 |
| 3.1 | Northbound Lanes | 4 |
| 3.2 | Southbound Lanes | 7 |
| 4 | Design Goals | 7 |
| 4.1 | Coverage | 7 |
| 4.2 | Image Quality | 7 |
| 4.2.1 | Spatial Resolution | 7 |
| 4.2.2 | Image Compression | 7 |
| 4.2.3 | Color Separation | 8 |
| 4.2.4 | Cabling | 8 |
| 4.3 | Durability | 8 |
| 4.4 | Control | 9 |
| 4.5 | Cameras for Human Surveillance | 9 |
| 5 | Design Details | 9 |
| 5.1 | Camera Parameters | 9 |
| 5.1.1 | Extrinsic Parameters | 9 |
| 5.1.2 | Intrinsic Parameters | 11 |
| 5.2 | Calculation of Coverage and Resolution | 11 |
| 5.2.1 | The image coordinate system | 11 |
| 5.2.2 | The Projection Matrix | 12 |
| 5.2.3 | Obtaining the Projection Matrix from Camera Parameters | 13 |
| 5.2.4 | Resolution Calculations | 14 |
| 5.2.5 | Calculation of Coverage Area | 16 |
| 5.3 | Design Strategy | 20 |
| 5.3.1 | Keeping Resolution Constant | 20 |
| 5.3.2 | Determining Coverage Area | 20 |
| 6 | Final Design | 21 |
| 6.1 | Camera Setup | 21 |
| 6.1.1 | Camera CCD sizes | 21 |
| 6.1.2 | Design Calculations | 21 |
| 6.2 | Cost Estimates and Specifications | 23 |
| 6.2.1 | The Platinum Plan | 23 |
| 6.2.2 | The Gold Plan | 23 |
| 6.2.3 | The Silver Plan | 24 |
| 6.2.4 | The Bronze Plan I | 24 |
| 6.2.5 | The Bronze Plan II | 25 |
| 6.3 | Summary of Options | 25 |

| | |
|---------------------------|-----------|
| 7 Conclusion | 25 |
| 8 Acknowledgements | 25 |

1 Introduction

The Pacific Park Tower is a 320-foot skyscraper located near the intersection of Powell Street and Interstate 80 (I-80) in Emeryville, California. The roof of this tower offers views of several interesting sections of I-80, including: a weaving section on the southbound approach to the split between I-580 and the San Francisco-Oakland Bay Bridge, on and off ramps feeding both directions, and highway segments that are equipped with loop detectors. We wish to place two sets of cameras to observe vehicles travelling in both directions, and we want the fields of view of these cameras to overlap so we may measure and record the behavior of individual vehicles over long stretches of this highway.

2 Executive Summary

2.1 Camera Setup

We present two sets of designs, one using cameras with 2/3-inch sensor arrays, the other using cameras with 1/2-inch sensor arrays. Both systems use six cameras in each direction, and both provide a resolution of 2 pixels per foot at camera image center. The 1/2-inch design provides continuous ground-plane coverage from 0 to 3500 feet in the southbound direction and from 0 to 7000 feet in the northbound. The 2/3-inch design provides coverage from 0 to 2900 feet in the southbound direction and from 0 to 5900 feet in the northbound.

2.2 Cost Estimate

| Plan | Sensor Size | Price |
|-----------|-------------|-----------|
| Gold | 1/2-inch | \$224,500 |
| Silver | 1/2-inch | \$137,500 |
| Bronze I | 2/3-inch | \$80,600 |
| Bronze II | 1/2-inch | \$77,500 |

We present four designs, three of which use 1/2-inch technology and one of which uses 2/3-inch technology. The table above summarizes the cost estimates for these four designs, ranked in order of final video image quality (best first). All prices are approximate, based on manufacturers' list prices. Volume and educational discounts will most likely apply.

3 Site Layout

3.1 Northbound Lanes

The layout for the northbound direction is shown in Figure 3.1. The origin for both northbound and southbound directions is a point on the ground directly beneath the mounted cameras. In the northbound direction, the road runs fairly straight until it reaches the Ashby Avenue overpasses, after which it bends slightly to the west. Since this bend occurs more than 2000 feet away from the origin, it has little effect on camera angles or distances and can be safely ignored. There are loop detectors installed in both directions at 1300, 2400, 4100, and 5900 feet from the origin. The detectors at 2400 feet are not visible from the roof because the Southern Ashby Ave overpass occludes them.

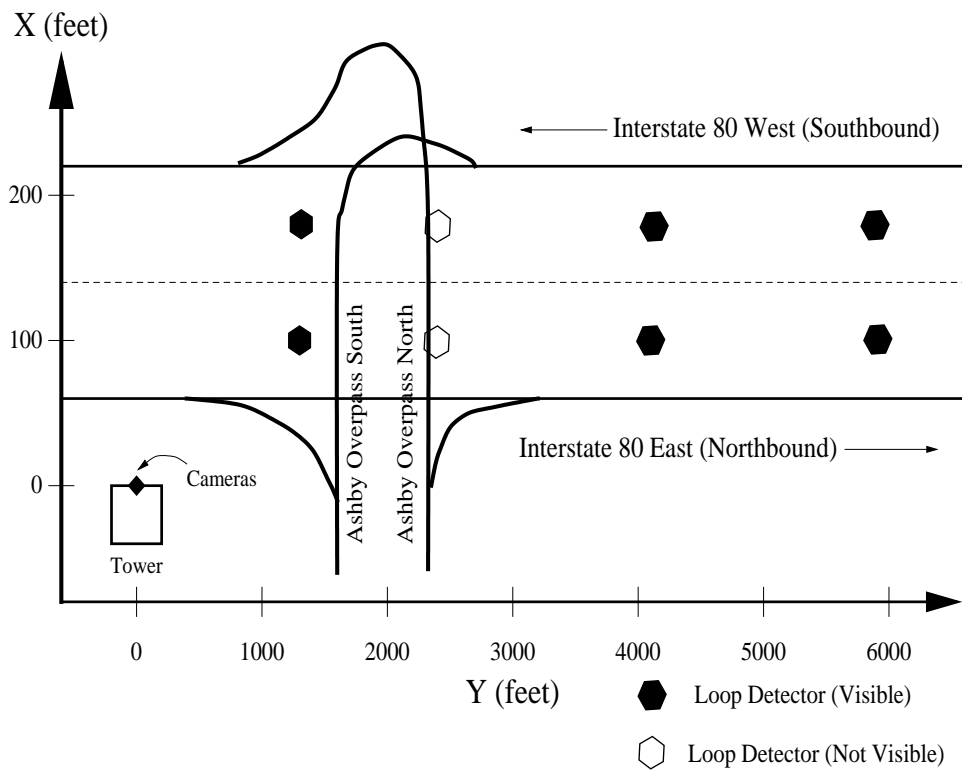


Figure 1: Site Layout, Northbound

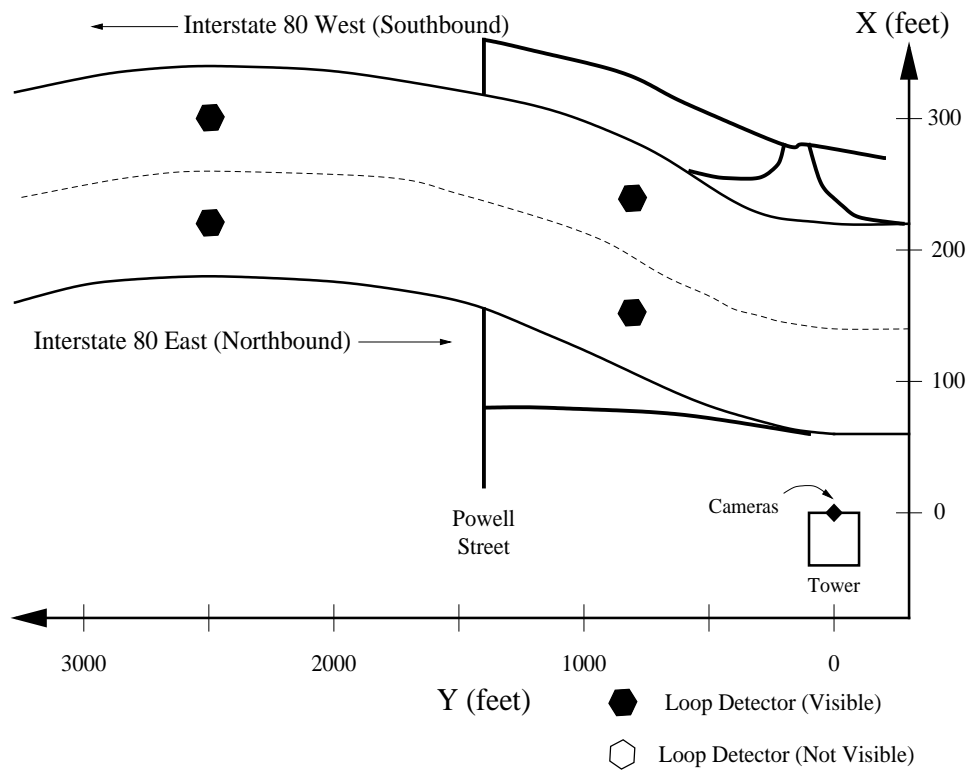


Figure 2: Site Layout, Southbound

3.2 Southbound Lanes

The southbound layout is shown in Figure 3.2. Here, the bend in the road is significant at about 300 feet from the origin, so its effects are included in our calculations. There are two sets of loop detectors installed in the southbound direction; one set is located 800 feet from the origin and measures traffic in both directions while the other is located 2500 feet from the origin and measures southbound traffic only.

4 Design Goals

4.1 Coverage

As we mentioned in the Introduction, we wish to install two sets of cameras to measure individual vehicle trajectories over long distances in both directions. Our video tracking algorithms require that our cameras observe departing traffic, so we set one set of cameras up to observe northbound traffic north of the tower, while the second set of cameras observes southbound traffic south of the tower. We sequence the cameras pointing in a given direction such that their fields of view overlap at the edges, so that vehicles can be reliably “handed off” from one camera to the next. To calculate fields of view on the ground, we must know the camera’s intrinsic parameters (focal length, image sensing area, skew between image axes) and extrinsic parameters (rotation and translation with respect to out world coordinate axes).

4.2 Image Quality

4.2.1 Spatial Resolution

In order to successfully detect and track vehicles, we must maintain a minimum bound on the number of pixels a vehicle occupies in a camera’s image plane. In addition, we want to match vehicles as accurately as possible by matching distinctive vehicle features such as size and color. It is easier to accurately compare these features if vehicles have approximately the same appearance from camera to camera. While it is impossible to maintain exact appearances from camera to camera, it is possible to design our system such that a given vehicle occupies approximately the same area on each camera’s image plane. The area covered by a given vehicle is not only a function of the intrinsic and extrinsic parameters, it is also a function of vehicle size and position relative to the camera. In general it is not possible to match vehicle sizes over the entire fields of view of two cameras. For this reason, we seek to match image sizes only at the center of each camera’s image plane; that is, if Vehicle X occupies N pixels when it appears at the image center of Camera 1, the same vehicle should occupy approximately N pixels when it appears at the image center of Camera 2.

4.2.2 Image Compression

There are several consumer-level digital video cameras available now that use compression algorithms or hardware Coder/Decoders (CODECs) to reduce the video stream’s bandwidth. Some of these algorithms use Discrete Fourier or Cosine Transforms to obtain video frequency information. In many of these cases high-frequency coefficients are discarded, since the human visual system does not have the bandwidth to make use of high-frequency information anyway. However, both of our vehicle detection algorithms depend on the presence of sharp contrast edges between a vehicle and the background. Algorithms that discard high-frequency components will degrade our vehicle detector by blurring these sharp edges.

Other algorithms, such as those used in MPEG coding or Sony DVCAM digital cameras, employ a motion segmentation algorithm to separate the changing parts of a video stream from the static background. Since our algorithms are also trying to segment out moving objects, these compression schemes will likely confound our vehicle tracker.

In summary, we wish to avoid image compression if at all possible. If we must accept compression to meet bandwidth requirements, we should use a completely invertible compression scheme.

4.2.3 Color Separation

We wish to determine how well we can match vehicles between cameras using color. The performance of color matching algorithms depends on how well we can distinguish between similarly-sized vehicles that appear in our cameras more or less simultaneously. This, in turn, depends on how precisely we measure vehicle color. Note that in principle it is not necessary to measure color accurately with respect to any accurate scale; it will be sufficient for us to obtain repeatable, distinguishable measurements in a “well-behaved” color space.

Most commercially available cameras measure color in terms of its Red, Green, and Blue (RGB) components. Many other color spaces are used for different purposes; most are derived by either linear (YIQ, CMY) or nonlinear (HSV) invertible transforms of the RGB space. Regardless of which colorspace we ultimately use to match vehicles, we need the highest dynamic range possible in each color channel (Red, Green, or Blue). To this end, it is better that we use cameras with three separate Charge-Control Device (CCD) arrays, tape decks that accept and record separate RGB channels, and three sets of cables for each camera/tape deck pair.

If the separate RGB approach proves too expensive, we can cut costs by incrementally downgrading our cameras and VTRs. As a first downgrade, we could still use three-CCD cameras but downgrade our VTRs and data link to SVHS, which is a two-channel system that sends grayscale luminance (Y) and color chrominance (C) signals separately. As a next downgrade, we could use a single-CCD color camera that produces SVHS output. Finally, we could downgrade both our cameras and VTRs to single-channel, composite NTSC color.

4.2.4 Cabling

Ideally, we would place our Video Tape Recorders (VTRs) close to the cameras to minimize loss in the cables running between them. We also should place our VTRs in a sheltered area where commercial AC power is available. On the roof of the Emeryville tower these two goals are at odds with each other, as the nearest sheltered, powered area is a maintenance shack about 100 feet from our proposed camera mounts. If we use high-quality shielded cables and connectors between cameras and VTRs, we will minimize the effects of noise, crosstalk, and signal loss that inevitably occur when running long lengths of cables together. The cables must be able to handle bandwidths of around 20 MHz with minimal loss.

In addition, for a given camera/VTR pair we must make the three cable lengths (R, G, and B) as close to each other as possible, and the cables and connectors for each RGB triplet should be of the same brand and lot if at all possible. This would minimize color changes caused by having the three color signals encounter different loss factors between camera and VTR.

4.3 Durability

We are mounting these cameras on a structure that stands about 320 feet above the ground. Our cameras will likely be exposed to high winds, extremes in temperature, and precipitation. We will therefore need environmental housing that protects our cameras and makes them usable under most weather conditions. We will also need to regulate the temperatures of our cameras to improve signal consistency, so we will need blowers to cool the cameras in extreme heat and heaters for extreme cold. Some housing units can also provide “windshield wipers” that keep the view reasonably clear in heavy rain, however we expect that such wipers would confuse our motion detection algorithms. Finally, many of cameras require adapters that convert AC to DC and also convert the camera’s raw CCD outputs to RGB, SVHS, or NTSC. Each adapter must be located within a few feet of the camera it supports. Since we have

no sheltering available close to the cameras, we should choose environmental housing with enough space to enclose one camera adapter unit for each camera mount.

4.4 Control

In addition to camera adapters, we need to remotely control the iris, focus, and zoom of each camera. Ideally these parameters will be set once when the cameras are mounted and never changed, however at this juncture we do not know exactly what the optimal tradeoff point is between resolution and coverage (as a matter of fact this is something we hope to learn). The most efficient setup would involve one control unit and one monochrome monitor connected to a switching device. This device would choose a camera as both a destination for control signals and as a source for grayscale video. In an RGB setup we would typically use the Green channel as an approximation to equivalent grayscale, since green has a larger contribution than either Red or Blue. In SVHS or NTSC systems the grayscale luminance (Y) channel is readily available. We need not mount the monochrome monitor permanently, and monitor quality is not of great importance as long as details of the road are reasonably discernable.

4.5 Cameras for Human Surveillance

So far, we have concentrated on camera requirements for an automated Video Surveillance System. Since we are already paying the fixed cost of designing this site, renting the roof space, and mounting cameras, it makes sense that we increase our benefit-to-cost ratio by including cameras more suited to human surveillance. Such cameras should be mounted on remotely controllable Pan-Tilt-Zoom (PTZ) units for maximum flexibility. We do not need high-color, high-resolution cameras, since humans will be using these cameras primarily to supervise the automated system.

We would need a way to broadcast this low-resolution, low frame-rate, grayscale video offsite to the humans performing the surveillance; one way would be to use the local telephone company's ISDN services. As an alternative, we might consider using Emeryville's cable television system to broadcast this information back to the laboratory. A company known as OMTV systems (<http://www.itsonline.com/omtv/>) markets such a system. In addition, the company claims to be able to send control signals from the remote user back to the camera unit, forming a closed-loop control and communications system. If we wish to design such cameras into our system, it may be worth investigating this further.

5 Design Details

We present this section as justification for the numbers we calculate in the next section. It can be safely skipped by readers who are only interested in the final design and cost estimate for the site.

5.1 Camera Parameters

5.1.1 Extrinsic Parameters

Figure 5.1.1 shows a diagram of a single camera setup. The world origin O lies on the ground plane, directly beneath the camera origin C . The camera is mounted at a height Z_0 from the ground plane. The camera's optical axis (a line that is perpendicular to the image plane at its center) intersects the ground plane at point $P_0 = (X_0, Y_0, 0)$. The distance from P_0 to the world origin O is given by $D_g = \sqrt{X_0^2 + Y_0^2}$. The distance from P_0 to the camera origin C is given by $D_0 = \sqrt{Z_0^2 + D_g^2} = \sqrt{X_0^2 + Y_0^2 + Z_0^2}$.

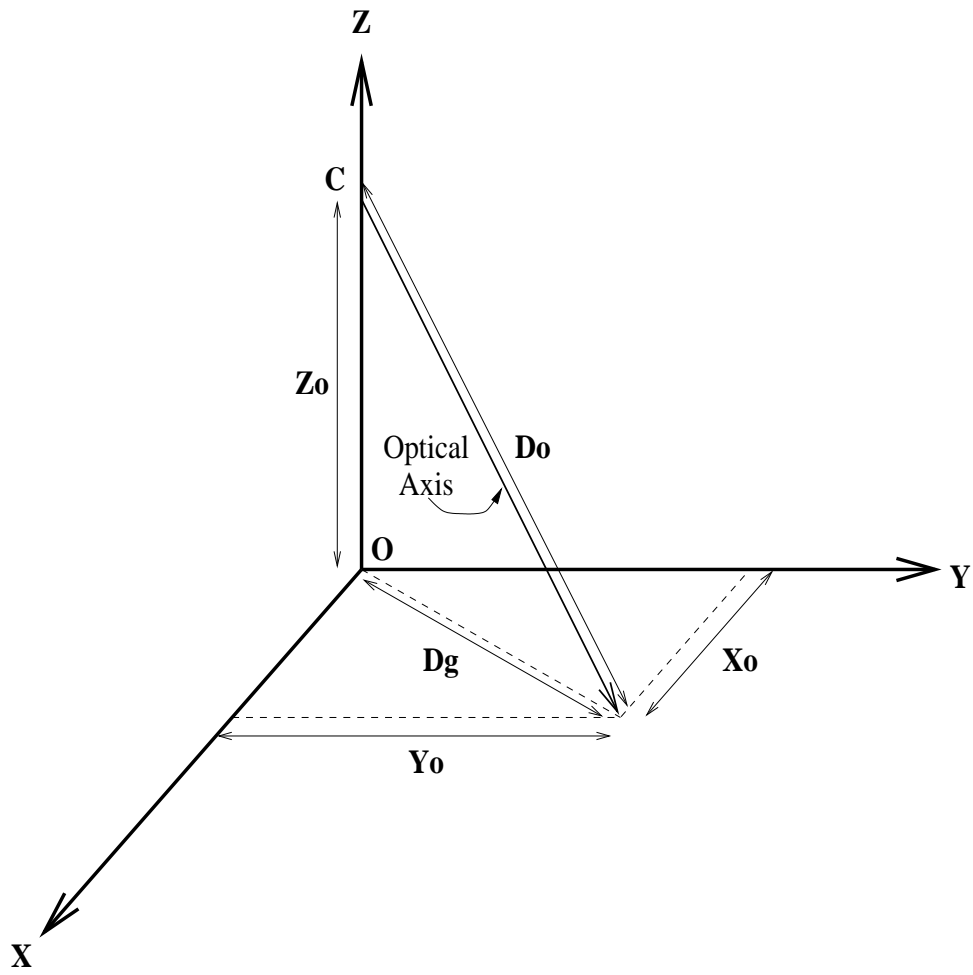


Figure 3: Extrinsic Camera Parameters

5.1.2 Intrinsic Parameters

There are several intrinsic camera parameters that affect both camera coverage and resolution. Figure 5.1.2 shows a diagram of the CCD image plane, which for convenience is located between the camera origin C and the image scene. Note that C in this figure is the same as the camera origin C in Figure 5.1.1.

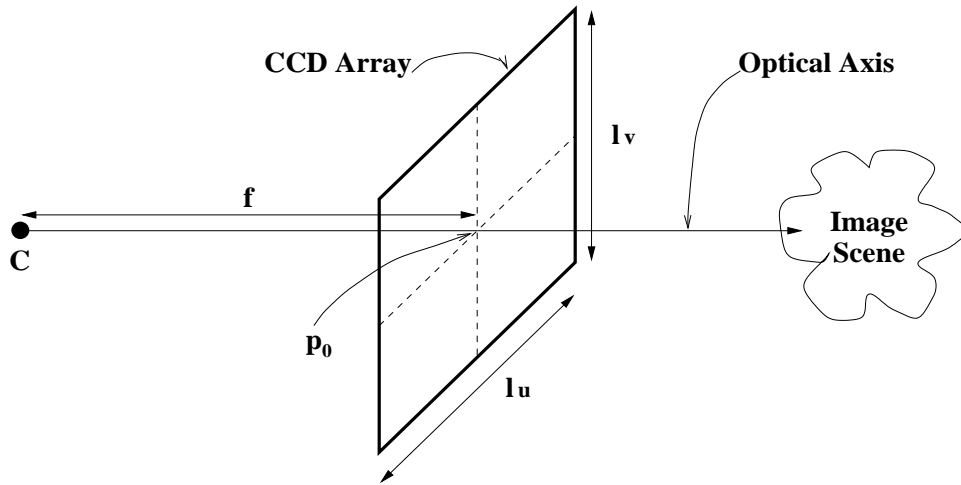


Figure 4: Intrinsic Camera Parameters

In Figure 5.1.2, p_0 is the *principal point*, defined as the intersection between the camera's optical axis and the image plane. For any given camera, it is not true in general that p_0 lies at the center of the image plane. However, in the absence of specific camera calibration information, it is reasonable to assume that a randomly chosen camera will have an expected principal point at or near the image center. The distance between the principal point and the camera center is given by the focal length f . The size of the CCD array in the horizontal direction is given by l_u , and the vertical size is given by l_v . There are also two parameters, k_u and k_v , that reflect the scaling between physical units on the image plane (usually millimeters) and pixel units in the horizontal and vertical directions, respectively.

There are other intrinsic parameter errors that will affect how the image scene is projected on the image plane, including the actual location of p_0 in a given camera, synchronization errors between horizontal scans of the CCD array (which show up as an axis skew parameter), and nonlinearities in the lenses (which translate into radial distortion errors). For the purposes of this analysis, we assume these errors are negligible and ignore them. In our tracking algorithms, we will take steps to "calibrate out" these errors as much as possible.

5.2 Calculation of Coverage and Resolution

5.2.1 The image coordinate system

Figure 5.2.3 shows the image of a typical freeway traffic scene with the image coordinate axes drawn to the side for convenience. By convention we choose the origin of the digitized image in computer frame coordinates to be the top left pixel, with the horizontal axis u going to the right and the vertical axis v going downward. Note that in this case the image origin is not the same as the principal point, which is expected to be somewhere near the center of the image. We choose a third axis w , which is parallel to the optical axis and points away from the reader to make a right-handed coordinate system.

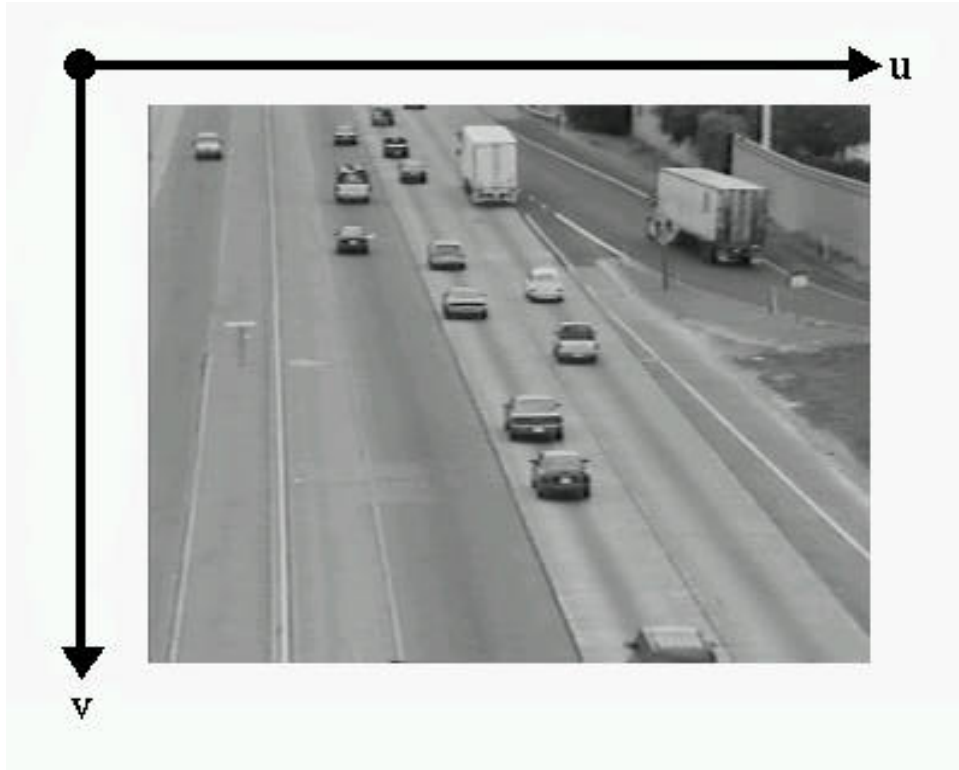


Figure 5: Image with Coordinate Axes

We can think of the conversion from world to image coordinates as a two-step process: the first process uses the camera's extrinsic parameters and its focal length f to determine the projection of the world point onto a camera coordinate system on the CCD array; the second process uses the location of p_0 and the scale factors k_u and k_v (along with axis skew and radial distortion, which we are ignoring) to determine the mapping between camera coordinates and image coordinates. To determine the camera's coverage area we need only consider the first process; to determine resolution we must consider both processes.

5.2.2 The Projection Matrix

Both transformations described in Section 5.2.1 can be combined into a single 3x4 matrix P known as the *Projection Matrix*. The mapping between points in world coordinates and points on the image plane involves rotation, translation, and scaling, therefore this mapping is not linear but affine. To express transformations as single matrices, we use a projective geometry trick: express our points in *homogeneous coordinates* rather than in Euclidean coordinates and write all our transformations as proportionalities rather than as equalities. Expressing a point in homogeneous coordinates is simple: just copy the first n elements of vector $x \in \mathbb{R}^n$ into an n by 1 column vector y , and set y_{n+1} equal to unity. With this in mind our projective relationship becomes:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \propto P \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (1)$$

To use Relation 1, we first introduce an intermediate coordinate system and turn the proportionality into an equality:

$$\begin{bmatrix} U \\ V \\ S \end{bmatrix} = P \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (2)$$

Next, we divide all values in this intermediate system by S (assuming $S \neq 0$) to obtain our image coordinates:

$$u = \frac{U}{S}, \quad v = \frac{V}{S} \quad (3)$$

5.2.3 Obtaining the Projection Matrix from Camera Parameters

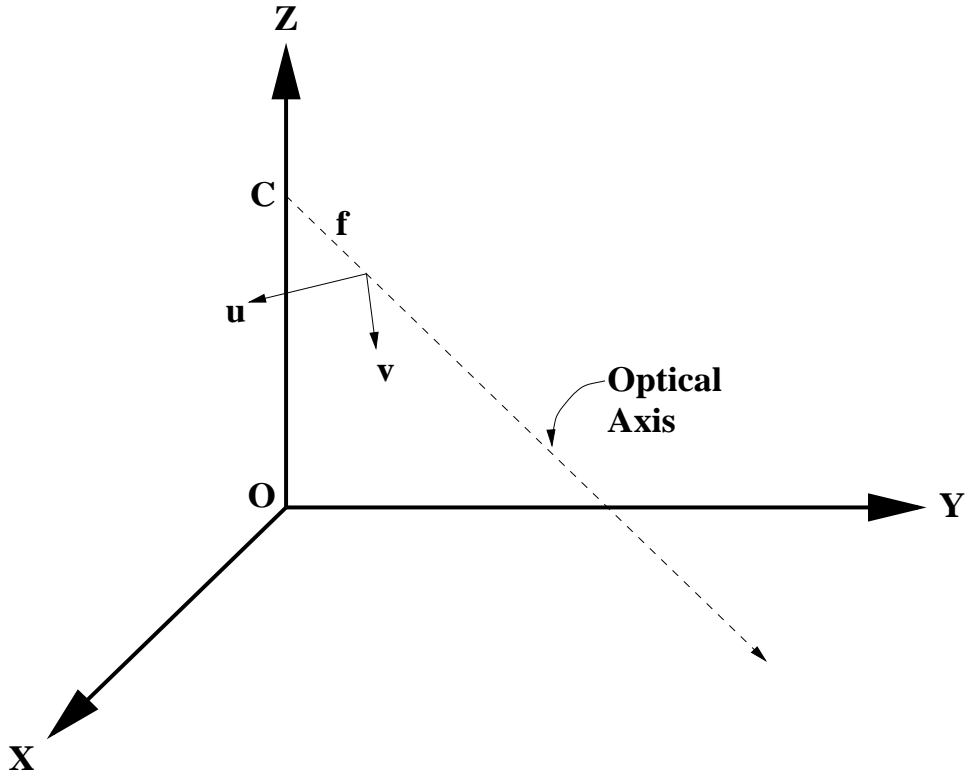


Figure 6: Orientation of Camera Axes in World Coordinate System

Now that we know how to use the Projection Matrix P , we discuss how to obtain P from camera parameters. To make this process simpler we make a few assumptions about the physical setup of the camera. Our first assumption is also our simplest: that the camera height Z_0 is strictly positive. Next, for simplicity we assume that we only have two degrees of freedom in our camera's motion: pan (rotation about the camera's v -axis) and tilt (rotation about the camera's u -axis). Roll (rotation about the optical or w -axis) is not allowed. We make the related assumption that

after the camera is rotated with respect to the world coordinate system, the camera's u -axis will lie in a plane that is parallel to the world XY -plane, or equivalently, that the projection of the camera's u -axis onto the world Z -axis is zero. We make the further restriction that an observer located at the camera origin looking down the positive w -axis sees a picture similar to that shown in Figure ; that is, the camera's u -axis points to the observer's right and the camera's v -axis is pointing down. Another way of stating this restriction is to say that the projection of the camera's v -axis onto the world Z -axis is nonpositive. When we combine these assumptions and restrictions we can obtain a unique description of the camera's orientation, except for one case: when the camera is pointing straight down. In this case we must specify the roll angle (the angle between the camera's u -axis and the world X -axis). The general case is shown in Figure 5.2.3.

With these assumptions in mind, we now construct the Projection Matrix. As we mentioned in Section 5.2.1, we can think of projection as a two-step process; the first maps a point in world (homogeneous) coordinates to a (homogeneous coordinate) measurement on the camera's CCD array, the second scales and translates this measurement from camera coordinates to computer image coordinates. Each of these processes can in turn be separated into two operations. The mapping from world to camera coordinates can be expressed as the matrix product FQ , where Q rotates and translates a point from world view to camera view and F scales this point by the camera's focal length. The mapping from camera to computer coordinates can be expressed as the matrix product KS , where S applies the effects of axis skew and K scales from physical to pixel coordinates and shifts the origin to the image top-left corner. Again, we are assuming that our CCD scan is perfect so that our angle θ in S below is $\pi/2$ (image axes are at perfect right angles). In this case S becomes identity, but we express it below in general form for completeness:

$$K = \begin{bmatrix} k_u & 0 & u_0 \\ 0 & k_v & v_0 \\ 0 & 0 & 1 \end{bmatrix} \quad (4)$$

$$S = \begin{bmatrix} 1 & -\cot(\theta) & 0 \\ 0 & \csc(\theta) & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (5)$$

$$F = \begin{bmatrix} f & 0 & 0 \\ 0 & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \quad (6)$$

$$Q = \frac{1}{D_g D_0} \begin{bmatrix} D_0 Y_0 & -D_0 X_0 & 0 & 0 \\ -X_0 Z_0 & -Y_0 Z_0 & -D_g^2 & D_g^2 Z_0 \\ D_g X_0 & D_g Y_0 & -D_g Z_0 & D_g Z_0^2 \end{bmatrix} \quad (7)$$

To form P , we simply multiply these four matrices together:

$$P = KSFQ \quad (8)$$

5.2.4 Resolution Calculations

Now that we have calculated the projection matrix P , we can use it to determine pixel resolution at any visible point in 3-space. We define two resolutions: one in the horizontal image direction and one in the vertical. For the horizontal direction, our resolution vector is the incremental change in pixel value differentiated by the incremental change in world coordinate value:

$$\vec{\rho}_u = \frac{\partial u}{\partial \vec{x}} \quad (9)$$

and for the vertical direction:

$$\vec{\rho}_v = \frac{\partial v}{\partial \vec{x}} \quad (10)$$

Note that both of these are vector values; we seek a scalar measure of resolution. One measure we might use is the Euclidean length of these resolution vectors. This length will be a function of world position as well as of P . As we stated in Section 4.2.1, we seek to normalize the image size of a vehicle from camera to camera. While it is not in general possible to do so for all sizes of vehicles at all points in space, it is possible to normalize a standard-sized object at a standard location with respect to each camera. We choose a sphere of incremental radius as our standard object and point $P_0 = (X_0, Y_0, 0)^T$ as our standard location (note that this is the point on the ground plane corresponding to image center).

With these choices we can form a horizontal resolution metric:

$$d_u = \|\vec{\rho}_u\|_{\vec{x}=P_0} \quad (11)$$

and a vertical resolution metric:

$$d_v = \|\vec{\rho}_v\|_{\vec{x}=P_0} \quad (12)$$

Using Equations 8, 2, and 3 we can write u as a function of X , Y , and Z (with skew angle θ set to $\pi/2$):

$$u = u_0 + \left(\frac{fk_u}{D_g}\right) \left[\frac{D_0(Y_0X - X_0Y)}{(X_0X + Y_0Y - Z_0Z + Z_0^2)} \right] \quad (13)$$

and similarly for v :

$$v = v_0 - \left(\frac{fk_v}{D_g}\right) \left[\frac{(X_0Z_0X + Y_0Z_0Y + D_g^2Z - D_g^2Z_0)}{(X_0X + Y_0Y - Z_0Z + Z_0^2)} \right] \quad (14)$$

Now, we apply Equation 9 to Equation 13 to produce the following:

$$\vec{\rho}_u|_{\vec{x}=P_0} = \frac{fk_u}{D_g D_o} \begin{bmatrix} Y_0 \\ -X_0 \\ 0 \end{bmatrix} \quad (15)$$

We do the same with Equations 10 and 14 to obtain:

$$\vec{\rho}_v|_{\vec{x}=P_0} = -\frac{fk_v}{D_g D_o} \begin{bmatrix} X_0 Z_0 \\ Y_0 Z_0 \\ D_g^2 \end{bmatrix} \quad (16)$$

Finally, we apply equations 11 and 12 respectively to Equations 15 and 16, use the equalities $D_g = \sqrt{X_0^2 + Y_0^2}$ and $D_0 = \sqrt{Z_0^2 + D_g^2} = \sqrt{X_0^2 + Y_0^2 + Z_0^2}$, and do some tedious algebra to obtain:

$$d_u = \frac{fk_u}{D_0}, \quad d_v = \frac{fk_v}{D_0} \quad (17)$$

Note that if the CCD pixels are square, then $k_u = k_v = k$ and $d_u = d_v = d = \frac{fk}{D_0}$. We have proven two facts that may be obvious to the casual observer: first, that an incremental sphere at a given distance from and at the center of view of a given camera has the same image size regardless of the camera's orientation with respect to that sphere; and second, that it is possible to adjust for changing distance from object to camera by simply changing the camera's focal length. While it may seem that we have done a lot of work to come to these simple conclusions; it is also true that we have derived equations along the way that may help us in the more general case (object not located in the center of view, etc.).

5.2.5 Calculation of Coverage Area

As we mentioned in Section 5.2.1, we need only consider the projection from world to camera coordinates (or more accurately, its inverse) to calculate the camera's coverage area. We define *Coverage Area* as the set of points on the ground plane that are visible from the camera given its orientation with respect to the world coordinate system. When the horizon is not visible in the camera's field of view this coverage area is a trapezoid that can be cut from an isosceles triangle; we will use our projection matrix to describe both the triangle and the trapezoid.

We begin with the projective relationship, Equation 1 (see page 12). Now, let us combine this with the decomposition of P (Equation 8) to get:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \propto KSFQ \begin{bmatrix} X \\ Y \\ Z \\ 1 \end{bmatrix} \quad (18)$$

Since we are only considering world points on the ground plane, we can set $Z = 0$, which in Equation 18 is the equivalent of striking out the third column of Q and the third row of the world vector. We define a new matrix M , equal to this reduced Q , and rewrite Equation 18:

$$\begin{bmatrix} u \\ v \\ 1 \end{bmatrix} \propto KSFM \begin{bmatrix} X \\ Y \\ 1 \end{bmatrix}; \quad (19)$$

where:

$$M = \frac{1}{D_g D_0} \begin{bmatrix} D_0 Y_0 & -D_0 X_0 & 0 \\ -X_0 Z_0 & -Y_0 Z_0 & D_g^2 Z_0 \\ D_g X_0 & D_g Y_0 & D_g Z_0^2 \end{bmatrix} \quad (20)$$

Incidentally, the composition $KSFM$ in Equation 19 represents the projective mapping between points on the ground plane and points in the image plane; it is called the *inverse homography matrix* and is denoted by G . Under all but a few degenerate conditions G has an inverse; this inverse is of course the *homography matrix* and is denoted by H .

Returning to the problem at hand, we stated in Section 5.2.1 that we only needed to know the mapping from image to camera coordinates to determine the camera coverage area. In terms of our matrices this means we only need F and Q in three-dimensional space; if we restrict our interest to coverage on the ground plane we only need F and M . Let us form the matrix $C = FM$ and show that C has an inverse. Since the determinant of F is f^2 and the determinant of M is $-Z_0$, the determinant of C is $-f^2 Z_0$. This determinant is nonzero when the focal length is nonzero and the camera is off the ground, both of which are always true in our system, therefore C has an inverse. Let us call this inverse matrix W . Recall that since C governs the mapping from world to camera coordinates, W governs the mapping from camera to world coordinates.

Now, let us label our camera coordinates u' and v' . Since we are still dealing with homogeneous coordinate systems, we can use our definition of W above to write:

$$\begin{bmatrix} X \\ Y \\ 1 \end{bmatrix} \propto W \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} \quad (21)$$

Following a path similar to that presented in Section 5.2.2, we use this proportionality by first calculating intermediate values:

$$\begin{bmatrix} X' \\ Y' \\ T' \end{bmatrix} = W \begin{bmatrix} u' \\ v' \\ 1 \end{bmatrix} \quad (22)$$

then by dividing by the third element (assuming $T \neq 0$):

$$X = \frac{X'}{T'}, \quad Y = \frac{Y'}{T'} \quad (23)$$

We apply this algorithm, expand, and simplify terms to obtain:

$$X = \frac{Z_0}{D_g (fZ_0 + D_g v')} [D_0 Y_0 u' - X_0 Z_0 v' + f D_g X_0] \quad (24)$$

for the X -coordinate on the ground plane, and:

$$Y = \frac{Z_0}{D_g (fZ_0 + D_g v')} [-D_0 X_0 u' - Y_0 Z_0 v' + f D_g Y_0] \quad (25)$$

for the Y -coordinate. Recalling Figure 5.1.2 and assuming that P_0 corresponds to the center of the image plane, we now find the four endpoints of the ground plane trapezoid covered by our camera:

First Point:

$$P_1 = \frac{Z_0}{D_g (2fZ_0 - l_v D_g)} \begin{bmatrix} 2f D_g X_0 - l_u D_0 Y_0 + l_v X_0 Z_0 \\ 2f D_g Y_0 + l_u D_0 X_0 + l_v Y_0 Z_0 \end{bmatrix} \quad (26)$$

Second Point:

$$P_2 = \frac{Z_0}{D_g(2fZ_0 - l_v D_g)} \begin{bmatrix} 2fD_g X_0 + l_u D_0 Y_0 + l_v X_0 Z_0 \\ 2fD_g Y_0 - l_u D_0 X_0 + l_v Y_0 Z_0 \end{bmatrix} \quad (27)$$

Third Point:

$$P_3 = \frac{Z_0}{D_g(2fZ_0 + l_v D_g)} \begin{bmatrix} 2fD_g X_0 + l_u D_0 Y_0 - l_v X_0 Z_0 \\ 2fD_g Y_0 - l_u D_0 X_0 - l_v Y_0 Z_0 \end{bmatrix} \quad (28)$$

Fourth Point:

$$P_4 = \frac{Z_0}{D_g(2fZ_0 + l_v D_g)} \begin{bmatrix} 2fD_g X_0 - l_u D_0 Y_0 - l_v X_0 Z_0 \\ 2fD_g Y_0 + l_u D_0 X_0 - l_v Y_0 Z_0 \end{bmatrix} \quad (29)$$

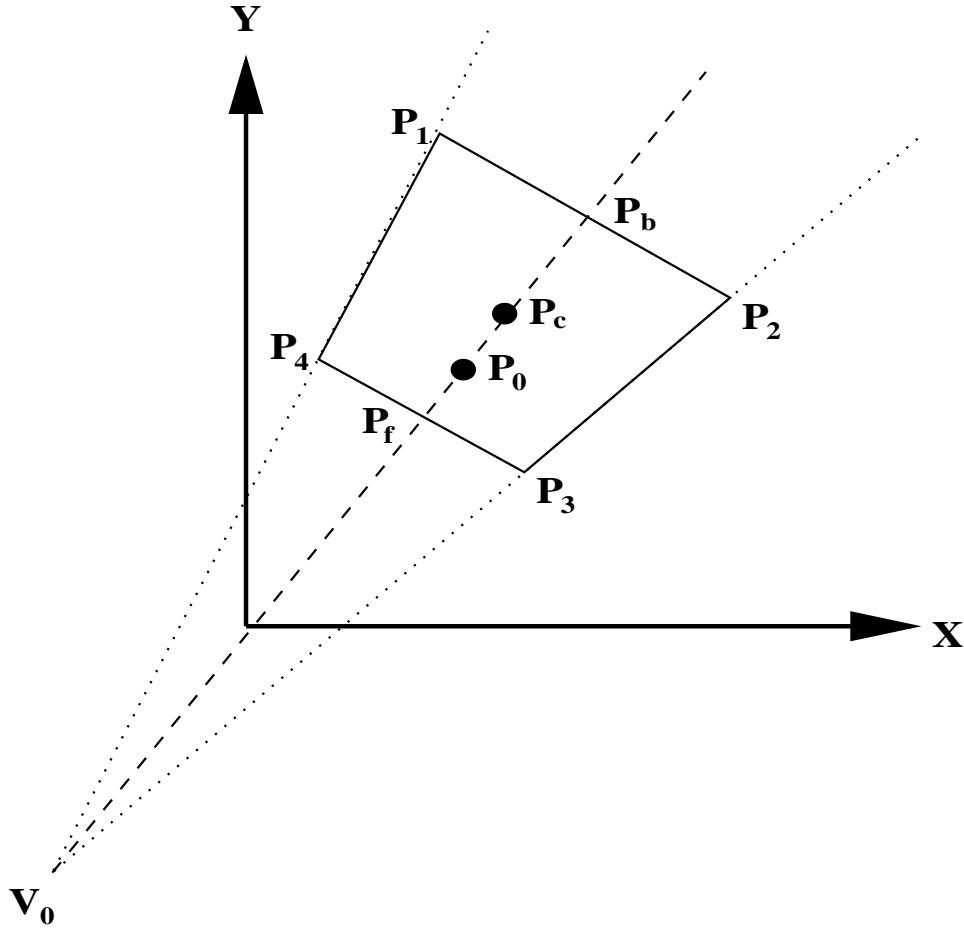


Figure 7: Ground Plane Coverage Area

The locations of these points for positive X_0 , Y_0 , and Z_0 are shown in Figure 5.2.5. Note that the two points furthest from the origin (P_1 and P_2) have the term $(2fZ_0 - l_v D_g)$ in the denominator. You may have already wondered what happens when D_g becomes large enough (all other terms are fixed by camera placement and design) to drive this term to zero. We solve for the D_g that makes this term zero and obtain:

$$D_g^* = \frac{2fZ_0}{l_v} \quad (30)$$

The value D_g^* represents a *critical radius*; when $D_g = D_g^*$ the points P_1 and P_2 go to infinity. In physical terms this means that if we aim our camera directly at a point on a circle of radius D_g^* the horizon just becomes visible at the top edge of the camera's image. As we would expect, this radius increases with camera height, and also increases with zoom (which is proportional to focal length). The critical radius decreases with the size of our CCD sensor as a taller CCD array gives a longer viewing angle.

In Figure 5.2.5 we see four other points of interest, labelled P_f , P_b , P_c , and V_0 . P_f is the center of the “front” of the trapezoid (from the camera's viewpoint) and is obtained by taking the vector average of P_1 and P_2 :

$$P_f = \frac{1}{2} (P_1 + P_2) = \frac{Z_0 (2fD_g - l_v Z_0)}{D_g (2fZ_0 + l_v D_g)} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \quad (31)$$

Similarly, P_b is the center of the “back” of the trapezoid and is given by:

$$P_b = \frac{1}{2} (P_3 + P_4) = \frac{Z_0 (2fD_g + l_v Z_0)}{D_g (2fZ_0 - l_v D_g)} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \quad (32)$$

The trapezoid center P_c is obtained by taking the average of P_f and P_b :

$$P_c = \frac{1}{2} (P_f + P_b) = Z_0^2 \frac{(4f^2 + l_v^2)}{(4f^2 Z_0^2 - l_v^2 D_g^2)} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \quad (33)$$

Note that in general, $P_c \neq P_0$, except for the special case $X_0 = Y_0 = 0$, which only occurs when camera is pointing straight down.

This can be shown by comparing the scalar multiple $Z_0^2 \frac{(4f^2 + l_v^2)}{(4f^2 Z_0^2 - l_v^2 D_g^2)}$ in Equation 33 with unity when f , l_v , and Z_0 are all nonzero. The last point of interest in Figure 5.2.5 is the *virtual vertex* V_0 , defined as the vertex of the isosceles triangle from which the trapezoidal coverage area is cut. In geometric terms, V_0 is found by extending two lines: one between P_1 and P_4 and the other between P_2 and P_3 . It can be shown that these two lines intersect at a point that is collinear with P_0 and the origin, and this point is given by:

$$V_0 = -\frac{Z_0^2}{D_g^2} \begin{bmatrix} X_0 \\ Y_0 \end{bmatrix} \quad (34)$$

For nonzero D_g , as Z_0 goes to zero (camera height decreases), V_0 approaches the origin. For nonzero Z_0 , as D_g goes to zero (camera pointing straight down), V_0 approaches a point at infinite distance in the direction of $-P_0$. This corresponds to having the two extended lines become parallel, at which point our trapezoid becomes a rectangle.

5.3 Design Strategy

In Section 5.2 we described in excruciating detail how to determine the coverage area and resolution metric for each camera. In this section we will briefly discuss how we apply these calculations to our particular design.

5.3.1 Keeping Resolution Constant

Recall that in Section 4.2.1 we stated that we wished to keep the resolutions more or less constant from camera center to camera center. Using our calculation for resolution metric (Equation 17) we can be more precise by saying we require the quantity $\frac{f}{D_0}$ to be constant over all cameras. Note that this approach will work whether our camera pixels are square or not.

5.3.2 Determining Coverage Area

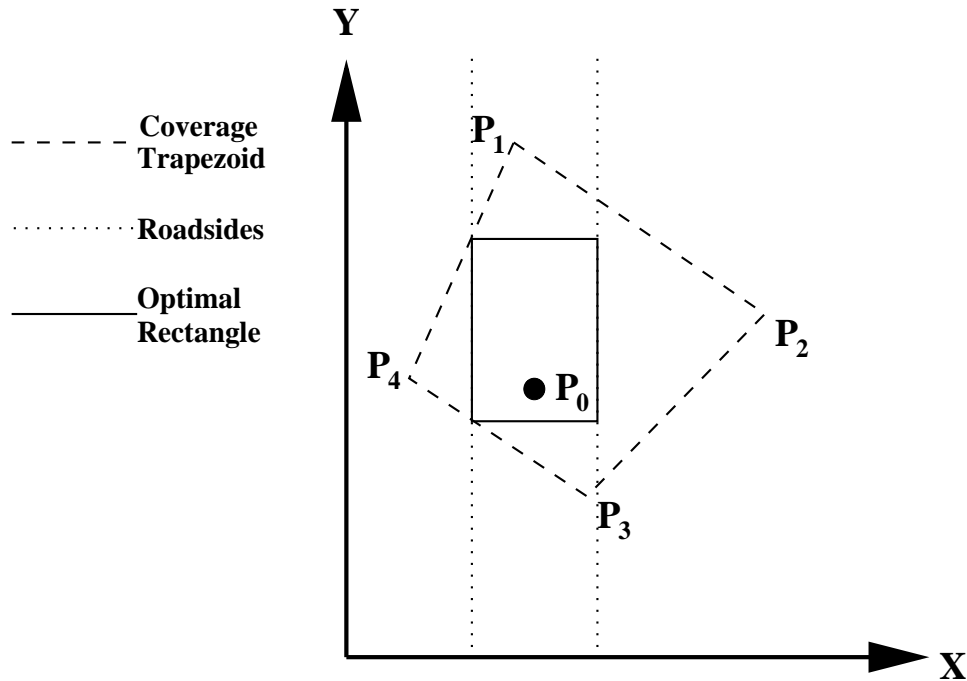


Figure 8: Determination of the Optimal Coverage Rectangle

We spent a great deal of time trying to determine the optimal usable coverage area for our tracking task. In the beginning, we wished to find the largest axis-aligned rectangle that would fit inside our trapezoid. After some work and a little thought, we realized that this was not the best approach, as it ignored an important practical truth: despite the fact we have ignored nonlinear lens distortion in our derivations, such distortion does in fact occur, but it is usually least severe around the image center. Furthermore, because we are dealing with roadways with set widths, we need only concern ourselves with the set of rectangles that are bounded by the roadsides. We therefore decided

to apply an admittedly more heuristic approach: assume that the optimal X_0 lies at the center of the road, then adjust f and Y_0 to determine coverage and resolution. Our coverage area will be defined as the largest rectangle bounded by the roadsides and the coverage trapezoid. We see a diagram of a typical case in Figure 5.3.2. Note from this figure that the far end of the rectangle lies in an area of decreased resolution where tracking may or may not be possible; to compensate for this we make sure that our sequence of cameras includes plenty of overlap between rectangles.

There is but one complicating factor remaining: from Figure 3.2 on Page 6 we see that our road curves away from the axes as we look south from the tower. To allow for this we approximate the road as a set of piecewise rectangular segments centered at the respective P_0 points for each camera. When we choose an approximate Y_0 for each camera, we estimate the center of the road at that value of Y_0 and use this road center estimate as our X_0 for that particular camera and road segment.

6 Final Design

6.1 Camera Setup

6.1.1 Camera CCD sizes

In Section 5 we showed how to calculate camera coverage and resolution, and described our strategy for using these calculations in our particular design. In this Section we present some hard numbers for both directions, using two commonly available camera types: those with 1/2-inch CCD arrays and those with 2/3-inch CCD arrays. Note that CCD array sizes are usually given as diagonal measures, and that the width to height aspect ratio is nominally 4:3 for television broadcast video. The diagonal and sides of a CCD array therefore form a 3-4-5 right triangle, so if the diagonal measure is denoted by l_d , then the CCD width l_u is nominally equal to $0.8l_d$ and the CCD height l_v is nominally $0.6l_d$. In addition, the “Sensing Area” of a CCD array is usually less than would be calculated this way, apparently for the same reason that the “Viewable Area” of a computer monitor is less than would be calculated using its diagonal measure. For this reason, we use manufacturer-supplied sensing dimensions in our calculations: for the 1/2-inch CCD, these dimensions are 6.4 by 4.8 mm, for the 2/3-inch CCD, the dimensions are 8.8 by 6.6 mm. In cases where we will get a larger optimal coverage rectangle (see Figure 5.3.2) by rotating our lens by $\frac{\pi}{2}$ (effectively interchanging the roles of l_u and l_v), we do so.

6.1.2 Design Calculations

We present our calculations for cameras with 1/2-inch CCD arrays looking north in Table 1, and looking south in Table 2. Similar calculations using 2/3-inch CCD cameras are presented in Table 3 for northbound and Table 4 for southbound. In all cases, we used $Z_0 = 320$ feet as our camera height, $d = 2.0$ pixels/ft as our resolution metric, and a pixel resolution of 640 by 480 regardless of CCD size. In all tables we use the following headings:

| | |
|-------|---|
| # | Camera Number |
| X_0 | Center of image in the X-direction, in feet |
| Y_0 | Center of image in the Y-direction, in feet |
| Y_f | Y-location of rectangle edge closest to camera, in feet |
| Y_b | Y-location of rectangle edge furthest from camera, in feet |
| f | Focal length of camera, in mm |
| R | Rotation indicator: Y means camera is rotated 90 degrees, N means no rotation |

| # | X_0 | Y_0 | Y_f | Y_b | f | R |
|---|-------|-------|-------|-------|------|-----|
| 1 | 100 | 50 | -100 | 210 | 7.0 | N |
| 2 | 100 | 300 | 150 | 650 | 9.0 | Y |
| 3 | 100 | 750 | 500 | 1500 | 16.4 | Y |
| 4 | 100 | 1300 | 900 | 2500 | 26.9 | Y |
| 5 | 100 | 2050 | 1500 | 4000 | 41.6 | Y |
| 6 | 100 | 3500 | 2500 | 7000 | 70.3 | Y |

Table 1: Design Calculations for Northbound Direction with 1/2-inch CCDs

| # | X_0 | Y_0 | Y_f | Y_b | f | R |
|---|-------|-------|-------|-------|------|-----|
| 1 | 180 | 50 | -100 | 200 | 7.4 | N |
| 2 | 200 | 300 | 140 | 500 | 9.6 | Y |
| 3 | 220 | 600 | 440 | 1000 | 14.3 | Y |
| 4 | 230 | 1000 | 760 | 1640 | 21.5 | Y |
| 5 | 280 | 1800 | 1450 | 2700 | 37.0 | Y |
| 6 | 300 | 2500 | 2100 | 3500 | 50.8 | Y |

Table 2: Design Calculations for Southbound Direction with 1/2-inch CCDs

| # | X_0 | Y_0 | Y_f | Y_b | f | R |
|---|-------|-------|-------|-------|------|-----|
| 1 | 100 | 50 | -80 | 200 | 8.8 | N |
| 2 | 100 | 300 | 150 | 610 | 13.5 | Y |
| 3 | 100 | 700 | 440 | 1300 | 23.3 | Y |
| 4 | 100 | 1300 | 880 | 2400 | 40.3 | Y |
| 5 | 100 | 2100 | 1500 | 3700 | 63.8 | Y |
| 6 | 100 | 3200 | 2200 | 5900 | 96.5 | Y |

Table 3: Design Calculations for Northbound Direction with 2/3-inch CCDs

| # | X_0 | Y_0 | Y_f | Y_b | f | R |
|---|-------|-------|-------|-------|------|-----|
| 1 | 180 | 50 | -80 | 160 | 11.1 | N |
| 2 | 200 | 250 | 140 | 400 | 13.6 | Y |
| 3 | 220 | 500 | 360 | 760 | 19.0 | Y |
| 4 | 240 | 800 | 620 | 1200 | 26.8 | Y |
| 5 | 260 | 1400 | 1050 | 1900 | 40.9 | Y |
| 6 | 280 | 2100 | 1700 | 2900 | 64.3 | Y |

Table 4: Design Calculations for Southbound Direction with 2/3-inch CCDs

Quick examination of the two sets of tables reveals that the 2/3-inch CCDs yields worse coverage than the 1/2-inch CCDs, which seems counterintuitive. We explain this paradox by pointing out that in most cases the larger CCD cameras had the same sensing elements as the smaller ones; therefore to make the resolution the same we had to increase f for the larger cameras. This in turn increased the zoom, which made the coverage area somewhat smaller. We do expect, however, that in practice we might get sharper pictures with the 2/3-inch cameras because the CCD separation is greater, which should lead to less pixel bleedover.

6.2 Cost Estimates and Specifications

As we discussed in Section 4.2 (Page 7), we have several gradations of quality available in both cameras and VTRs. Therefore, we present four plans, in order of decreasing cost and final image quality. These costs include cameras with power supplies, camera environmental housing, cabling, controllers, and VTRs. They do not include a monitor for camera setup (which likely could be borrowed), an image digitization workstation for the Vision Lab (another topic entirely), camera support structures, video tapes, extra cameras for human surveillance, a rack or enclosure for the VTRs, any security system required to protect the equipment from theft or vandalism, or remote telephone control facilities for the VTRs (which, according to one Sony representative, would likely have to be specially designed by a third party). All prices are manufacturer's list; we will most likely be eligible for volume and/or educational discounts. The total price of each plan reflects the cost using 12 camera/VTR pairs. We also include estimates of weight, dimension, and power consumption, when they are available from the Manufacturer's Specifications. A "?" means such specifications were not available at time of writing; a "-" means the specifications are not applicable.

6.2.1 The Platinum Plan

This plan is not itemized. It would include Digital Cameras connected to Digital Video Recorders using Sony's QSDI uncompressed digital video format.

Total Cost: **Over \$100,000 per Camera/DVR pair.**

Advantages: Best possible image quality available. No digitizing workstation required.

Disadvantages: High cost of media. High cost of insurance. High cost of everything else associated with this plan.

6.2.2 The Gold Plan

| Item | Vendor | Model | W(mm) | L(mm) | H(mm) | M(kg) | P(Watts) | Price | Quan | Total |
|---------|--------|-------------|-------|-------|-------|-------|----------|--------|------|----------|
| Camera | Sony | DXC-9000 | 79 | 168.2 | 72 | 0.8 | 11.5 | \$5800 | 12 | \$69,600 |
| Lens | Sony | VCL-716BXEA | 128 | 168.9 | 97.5 | 1.7 | - | \$4695 | 12 | \$56,340 |
| Adapter | Sony | CMA-D2 | 210 | 200 | 44 | 1.1 | 24.5 | \$175 | 12 | \$2,100 |
| Control | Sony | RM-C950 | 212 | 132 | 41 | 0.4 | ? | \$650 | 12 | \$7,800 |
| Housing | Pelco | EH-66XHB | 240 | 370 | 220 | 6.4 | 171 | \$692 | 12 | \$8,304 |
| VTRs | Sony | UVW-1700 | ? | ? | ? | 19 | 85 | \$6500 | 12 | \$78,000 |
| Cables | - | - | - | - | - | - | - | \$200 | 12 | \$2,400 |

Total Cost: Approximately **\$224,500** for listed items.

Advantages: Cameras have 3 CCD arrays, one for each color channel. VTRs accept 3-component video, so color channels are kept separate from sensors to tape. Cameras use Progressive Scan technology to minimize blur caused by motion.

Disadvantages: This setup would require a separate enclosure be build on the roof near the cameras, with AC power available, for the camera adapters, which would not fit in the enclosures. VTRs can only tape 90 minutes at a time.

6.2.3 The Silver Plan

| Item | Vendor | Model | W(mm) | L(mm) | H(mm) | M(kg) | P(Watts) | Price | Quan | Total |
|---------|--------|-----------|-------|-------|-------|-------|----------|--------|------|----------|
| Camera | Sony | DXC-151A | 79 | 168.2 | 72 | 0.8 | 11.5 | \$1395 | 12 | \$16,740 |
| Lens | Canon | J10X10R-1 | ? | ? | ? | ? | - | \$1595 | 12 | \$19,140 |
| Adapter | Sony | CMA-D2 | 210 | 200 | 44 | 1.1 | 24.5 | \$175 | 12 | \$2,100 |
| Control | Sony | RM-C950 | 212 | 132 | 41 | 0.4 | ? | \$650 | 12 | \$7,800 |
| Housing | Pelco | EH-66XHB | 240 | 370 | 220 | 6.4 | 171 | \$692 | 12 | \$8,304 |
| VTRs | Sony | UVW-1700 | ? | ? | ? | 19 | 85 | \$6500 | 12 | \$78,000 |
| Cables | - | - | - | - | - | - | - | \$200 | 12 | \$2,400 |

Total Cost: Approximately **\$137,500** for listed items.

Advantages: Separate RGB channels from Camera output to tape.

Disadvantages: Separate adapter enclosure required. Single-Chip CCD. No Progressive Scan.

6.2.4 The Bronze Plan I

| Item | Vendor | Model | W(mm) | L(mm) | H(mm) | M(kg) | P(Watts) | Price | Quan | Total |
|---------|--------|----------|-------|-------|-------|-------|----------|----------|------|----------|
| Camera | Cohu | 4860 | ? | ? | ? | 20 | 50 | \$2430 | 12 | \$29,160 |
| Lens | Cohu | - | - | - | - | - | - | \$1485 | 12 | \$17,820 |
| Adapter | Cohu | - | - | - | - | - | - | - | - | - |
| Control | Cohu | MPC | - | - | - | - | ? | \$12,000 | - | \$12,000 |
| Housing | Cohu | - | - | - | - | - | - | - | - | - |
| VTRs | Sony | SVO-2100 | ? | ? | ? | ? | 28 | \$1500 | 12 | \$18,000 |
| Cables | - | - | - | - | - | - | - | \$300 | 12 | \$3,600 |

Total Cost: Approximately **\$80,600** for listed items.

Advantages: Camera, adapter, housing, lens, control all built by vendor with Traffic Surveillance Experience. Possibly greater discount if cameras are purchased through CalTrans Channels. 2/3-inch CCD arrays should yield greater image clarity. Less “homegrown” assembly required, as Camera/adapter/lens provided as single unit.

Disadvantages: Video degraded to SVHS (2-channel) rather than RGB (3-channel). Single-Chip CCD. No Progressive Scan. Control Unit would need “Engineering,” (according to vendor) to accommodate 12 cameras, which is out of the ordinary. This might lead to higher costs.

6.2.5 The Bronze Plan II

| Item | Vendor | Model | W(mm) | L(mm) | H(mm) | M(kg) | P(Watts) | Price | Quan | Total |
|---------|--------|-----------|-------|-------|-------|-------|----------|--------|------|----------|
| Camera | Sony | DXC-151A | 79 | 168.2 | 72 | 0.8 | 11.5 | \$1395 | 12 | \$16,740 |
| Lens | Canon | J10X10R-1 | ? | ? | ? | ? | - | \$1595 | 12 | \$19,140 |
| Adapter | Sony | CMA-D2 | 210 | 200 | 44 | 1.1 | 24.5 | \$175 | 12 | \$2,100 |
| Control | Sony | RM-C950 | 212 | 132 | 41 | 0.4 | ? | \$650 | 12 | \$7,800 |
| Housing | Pelco | EH-66XHB | 240 | 370 | 220 | 6.4 | 171 | \$692 | 12 | \$8,304 |
| VTRs | Sony | SVO-2100 | ? | ? | ? | ? | 28 | \$1500 | 12 | \$18,000 |
| Cables | - | - | - | - | - | - | - | \$200 | 12 | \$2,400 |

Total Cost: Approximately **\$77,500** for listed items.

Advantages: Cheapest Solution Presented.

Disadvantages: Separate adapter enclosure required. Single-Chip CCD. No Progressive Scan. Video degraded to SVHS.

6.3 Summary of Options

We have presented several plans of varying cost and quality, four of which could feasibly fit within our budget. All plans require that we pay the fixed cost (in one form or another) of cabling, housing, and camera adapters. The gradations in price, therefore, are reflected in the quality of cameras and VTRs. While the casual human observer might not think that the difference in image quality between the Gold Plan and either of the Bronze Plans would be worth the difference in price, we believe that our video tracking and matching algorithms would benefit greatly from this difference.

7 Conclusion

We have described a design for a multiple-camera video traffic surveillance system, have justified that design with a discussion of projective geometry, and have provided cost estimates implementing the design with several variations of equipment quality. We have stated and met our design goals with varying degrees of success, again in proportion to cost of equipment. We have run out of things to write.

8 Acknowledgements

Cost estimates for Sony and Canon products provided by Mark Flores of ProMax Systems, Inc., <http://www.scsidisk.com/>.

Cost estimates for Pelco products provided by Cindy Maloney of Pelco, Inc.

Cost estimates for Cohu products provided by David Lane of Cohu Video Systems.

References

- [1] Faugeras, Olivier, *Three Dimensional Computer Vision: A Geometric Viewpoint* (Cambridge, MA: The MIT Press, 1993), Chapter 3.
- [2] Sony Corporation, World-Wide-Web Page, <http://www.sony.com/>
- [3] Cohu Corporation, World-Wide-Web Page, <http://www.cohu.com/>
- [4] Pelco Corporation, World-Wide-Web Page, <http://www.pelco.com/>