

# Model-based Interpretation of Stereo Imagery of Textured Surfaces\*

Wenyi Zhao<sup>†</sup>, N. Nandhakumar<sup>‡</sup>, and Philip W. Smith\*

<sup>†</sup> Dept. of Electrical Engineering, University of Maryland, College Park, MD 20742

<sup>‡</sup> Electroglas, Inc., 2901 Coronado Dr., Santa Clara, CA, 95054

\* Dept. of Electrical Engineering, University of Virginia, Charlottesville, VA 22903

\*This work was funded in part by a research contract from Simpson Weather Associates, Inc., and in part by the NSF under grant IRI-91109584.

# Abstract

We present a scheme for reliable and accurate surface reconstruction from stereoscopic images containing only fine texture and no stable high level features. Partial shape information is used to improve surface computation – first by fitting an approximate, global, parametric model, and then by refining this model via local correspondence processes. This scheme eliminates the window size selection problem in existing area based stereo correspondence schemes. These ideas are integrated in a practical vision system that is being used by environmental scientists to study wind erosion of bulk material such as coal ore being transported in open rail cars.

# 1 Introduction

Due to both economic and environmental concerns, scientists have begun to study the process of wind erosion in open rail cars transporting bulk material, such as gravel, iron or coal ore. Research teams have found their research impeded, however, because many of the tools necessary for this work have yet to be developed. In this paper, we address the design of a binocular stereo system called **CCLPS** (Computerized CoaL Profiling System) that provides dense, accurate disparity maps of coal as it is being transported. As can be seen in Fig. 1, three cameras constituting two wide-baseline, parallel axis stereo pairs are placed above the train and image sets are collected as the cars pass underneath. Two-dimensional depth profiles are stereoscopically extracted from these sets and combined to form three-dimensional material surface maps for each car. These maps are then used by the environmental science research team to study the effects of wind erosion on ore loss.

Because the system employs a wide baseline for enhanced depth resolution and operates in an uncontrolled environment, solving the correspondence problem is quite challenging. The large distance between the cameras produces significant photometric variation and foreshortening in the stereo pair since the imaged surface is non-lambertian and is not parallel to the image planes. Also, the material surface contains only fine grain textures without stable high level features. Both of the above issues further complicate the establishment of correspondence.

Establishing correspondence is an important, although difficult, part of computational stereo. The task of correspondence matching, in the case of binocular stereo is to find paired primitives in two different images of the same scene. Although the human vision system routinely performs stereo correspondence for depth recovery, developing a computational scheme to find these pairs automat-

## Three Cameras Operating As Two Stereo Pairs

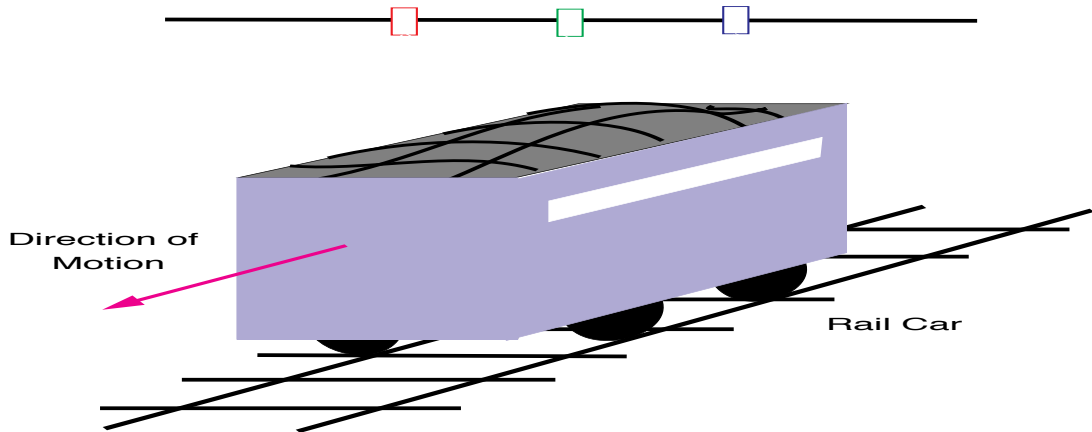


Figure 1: CCLPS Data Acquisition Scheme. Three cameras acting as two stereo pairs are placed above the train and image sets are collected as the train travels underneath.

shape information to improve surface reconstruction.

We describe a new (in the sense that it is different from traditional two-step coarse-to-fine methods and it uses a piece-wise linear model and minimum depth profile information to explicitly model discontinuity ) approach for stereo matching and illustrate the advantages of this scheme in implementing a practical system aimed at a specific application. The application involves the processing of imagery containing fine texture and no stable, high level, local features (Smith and Nandhakumar 1994). Such imagery is typically obtained when viewing scenes containing coal, gravel, iron ore, *etc.* We also discuss other possible applications of this approach.

This paper is organized as follows: The following section briefly presents some related work in stereo matching. Section 3 describes the new stereo correspondence algorithm that uses *a priori*, minimal, approximate surface shape information. Section 4 presents experimental results using both simulated data and real data, and a performance comparison between our algorithm and an alternative stereo algorithm (Kanade and Okutomi 1994). The last section contains conclusions and comments on future work in this area.

## 2 Related Work

Many different stereoscopic camera configurations have been developed, *e.g.*, binocular stereo, omniscopic stereo (Adelson and Wang 1992) and trinocular stereo (Ayache and Lustman 1987). However the binocular stereo configuration is still the most common, especially the parallel axis binocular system. This may be partly due to the fact that the correspondence problem has not been completely solved even for the simplest, binocular, configuration.

The entire process from data acquisition to depth reconstruction typically consists of five

steps (Marapane and Trivedi 1994). Thus, stereoscopic vision systems may be classified based on each of these steps. Of those classifications, important ones are based on different primitive representations and different stereo matching techniques. Popularly used matching primitives include pixel intensities (Barnard and Thompson 1980) or their functions, sign of the Difference of Gaussian or Laplacian of the Gaussian  $\nabla^2 G$  filtered image (Nishihara 1984), zero crossing in  $\nabla^2 G$  filtered image (Marr and Poggio 1979) and edge elements (Medioni and Nevatia 1985; Baker and Binford 1981). These existing correspondence schemes can be classified into two broad categories: (1) *Area-based* methods and (2) *Feature-based* methods. For a comprehensive review of existing stereo technologies, we refer the reader to several surveys, *e.g.*, (Barnard and Fischler 1982; Dhond and Aggarwal 1989; Marapane and Trivedi 1994).

Feature-based correspondence methods use local features such as edge elements for matching. The popularity of this approach is increasing because of the belief that the feature is more abstract than the original image, and thus should be more stable (Medioni and Nevatia 1985). Also this method has been shown to be more accurate than area-based schemes; in some cases the feature can be located to sub-pixel precision. However, feature-based methods suffer three main disadvantages. First, these methods are *not useful* when the local features cannot be reliably extracted. Second, these methods are *computationally more expensive* than some area-based methods because of the cost of the feature extraction procedure, the ‘choose-one-from-many-candidates’ procedure and/or ‘discard-wrong-candidates’ procedure involved. Finally *sparse and irregular* features produce sparse disparity maps, requiring depth interpolation and surface reconstruction (Grimson 1981; Terzopoulos 1988) if a complete depth map is needed. Recently a feature-based stereo algorithm was proposed by Lew et al. (Lew, Huang and Wong 1994) to alleviate these difficulties which integrates learning, feature selection, and surface reconstruction.

An area-based scheme is a natural choice when dealing with textured images. Computing the correlation coefficient and sum of squared differences (SSD) are simple yet effective techniques for obtaining a dense depth map from images (Wood 1983; Kanade and Okutomi 1994). For example, a practical vision system has been developed by Nishihara (Nishihara 1984) using auto-correlation function of signals formed by the sign of DOG filtered image. But the system needs structured lighting, to ensure dense texture of high contrast and low noise levels to perform well. Other methods, *e.g.* (Lee et al. 1993), have employed region invariants to determine correspondence. However, these methods are sensitive to repetitive textures and also are computationally expensive.

Kanade and Okutomi (Kanade and Okutomi 1994) have demonstrated the importance of window size selection for area-based methods. Erroneous matches are generated if the window is too small. On the other hand, if too large a window is chosen, large disparity changes within the window cause displacements between the detected match and the correct match. The use of adaptive window sizes (which are difficult to specify) does not always rectify this limitation of area-based techniques. Consider a pair of images of a steeply slanted surface. The previous methods prescribe small windows for correspondence, followed by surface interpolation. Foreshortening distortion and low SNR will cause false matches and hence a jagged reconstruction of the surface. Cepstral techniques (used in acoustical signal processing) have been adopted to increase tolerance to low SNR in each window (Hassab and Boucher 1975; Oppenheim and Schaffer 1993; Olson 1993; Yeshurun and Schwartz 1989; Smith and Nandhakumar 1993; Bandari and Little 1993). The improvement in performance for low SNR images has been shown via analytical arguments as well as experimental tests. However, repetitive fine texture in the images generate many false matches, and the search space needs to be tightly constrained about the correct location. Unfortunately, this leads to a “chicken-and-egg” problem, i.e, shape information is required to constrain the search for

correspondence (correct shape).

Recently, a hierarchical scheme that integrates different matching techniques at different levels of processing has been proposed (Marapane and Trivedi 1994) for computing improved, denser disparity maps. However, when feature extraction becomes infeasible or unreliable, the scheme relies entirely on area matching, and hence suffers from the limitations discussed above. Another scheme integrates correspondence and surface interpolation (Hoff and Ahuja 1989), but also requires extraction of stable local features and assumes piecewise planar and quadratic surface patches. Other hierarchical schemes have similar problems (Cohen et al. 1989; Kim and Binford 1987).

In the problem that we are addressing, there exist large *photometric variation* (because of the wide-baseline), *discontinuities* in surface, *frequent occlusions*, very few or no *stable, local features*, and large amounts of *noise*. We were unable to find existing techniques or schemes that could be directly applied to our problem without significant modification. The most promising approach was to adopt an area based technique but devise an intelligent algorithm to overcome its limitation.

We first select a parametric form for the global shape that engenders computationally simple algorithms, yet adequately represents the limited shape knowledge available, e.g. the surface is higher in the middle than at either end. Given a stereo image pair, we then compute the optimal model parameters - which gives us an approximate (but reliable) global, parametric depth profile. Finally, we use the computed approximate shape to constrain the local correspondence processes that refine the shape. This scheme can be applied even when the physical scene is not a single smooth surface, and the image is noisy. We use a piecewise linear surface model to approximate the scene. We show that with only a very approximate idea of the global scene structure we can obtain a highly accurate local shape estimate. Section 3 will discuss our algorithm in detail.

### 3 Computing Accurate and Reliable Shape

The need for an improved area based correspondence scheme, the required performance characteristics, and the formulation of the new scheme is best explained by describing the new algorithm in terms of the application that warranted this research effort. Our goal is to compute 3D surface shape/height of coal, gravel or other bulk material present in open rail cars. The cameras in Fig. 1 continually acquire images as the rail cars pass beneath them. Rather than use an entire image collected by each camera to reconstruct the 3D surface, we use only a few raster lines near the piercing point<sup>1</sup> of each camera to compute a one dimensional depth profile across the width of the rail car. We perform this task repetitively as the rail car moves under the cameras - thus computing transectional 3D profiles at regularly spaced intervals along the length of the rail car. We stack these profiles to form a 3D surface map for the entire rail car. In terms of stereo vision technology, our application requires values of *relative depth* of the surface given some absolute depth value. We use the depth of the sill of the rail car as the reference depth because sills are distinctive in the images and can be easily extracted and matched. Here by *relative depth* (or *relative disparity*), we mean that after we choose some absolute value of depth as reference (e.g., the depth of the sill), we can investigate the local change around that value.

#### 3.1 Using Approximate Shape Information - An Overview

Previous area-based methods use correlation-related schemes that only consider image intensity information. These approaches failed when adopted for this application because of the repetitive texture and frequent occlusion, which, when compounded by noise, produced a large number of

<sup>1</sup>The piercing point is the intersection of the optical axis and the image plane

false correspondences. Although the number of these false correspondences can be reduced by using larger windows in the area-based methods, the presence of large disparity changes and associated projective and photometric distortions cause mismatches when large windows are used (Kanade and Okutomi 1994). One way of overcoming this problem would be to correct for the projective & photometric differences between the two images in a stereo pair before attempting correspondence. However, this requires that surface shape be known *a priori* - resulting in a variation of the “chicken-and-egg” problem. However, even limited information about the shape of the depth profile (along the epipolar line) can be sufficient to effect the projective corrections necessary for reliable and accurate correspondence when large windows are used. This partial information may be of the form, for example, that the surface is higher in the middle than at either extremity. Indeed, after using such information to correct for projective variation between the image pair the windows used for establishing correspondence may be as wide as the entire image itself! The correction for photometric differences will be discussed in section 3.4.

The key issue in correcting projective distortions is the choice of a surface shape representation that allows us to specify the shape in a very approximate manner. We first select a parametric form for the global shape that engenders computationally simple methods yet adequately represents the very limited shape knowledge available. Once we choose the model, our stereo correspondence scheme transforms to a search for the correct model parameters. Given a stereo image pair, we compute the optimal model parameters - which gives us an approximate (but reliable) global, parametric depth profile. We use the partial shape information to constrain the search for the optimal model parameters. The search consists of a hypothesize and test process in which the hypothesized shape is used to apply a foreshortening (projective) correction to the left image to produce a virtual right image. The *entire* virtual right image and *entire* real right image are

compared to verify the hypothesis. The global, approximate shape computation technique is reliable since it uses the widest possible window – one that is as wide as the image itself, and models disparity changes by applying projective corrections.

Once the best parametric surface fit is found we then refine the surface by performing a local correspondence match at different positions along the epipolar line. At this stage we use the parametric surface fit to tightly constrain the search space during the local correspondence process. An advantage of this two step approach of (1) computing a reliable though approximate, global, parametric surface, and then (2) refining the surface, is that if the local correspondence process in step (2) does not produce a high measure of confidence, one still has the approximate surface computed in step (1).

We first discuss the specification of the parametric shape model, and then describe an algorithm that uses the two step process described above to extract a reliable and accurate estimate of the 3D surface.

## 3.2 Modeling Approximate Shape Information

Consider the variation in height/depth along an epipolar line. We first specify the shape of this depth profile in parametric form and then search for the parameters *in lieu* of establishing local correspondences at many points along the profile. It is important to select a simple parametric form that results in a low dimensional parameter search space. Yet, the model must adequately capture all of the available *a priori*, approximate shape information that can facilitate stereoscopic depth computation. The choice of an appropriate parametric form is application dependent, and depends on the nature of the surfaces reconstructed. A higher order polynomial is one possible choice for this purpose. However, there are many disadvantages in using higher order polynomials here. These

include – (1) high computational cost involved in fitting higher order polynomials, (2) large search space for each coefficient due to the high sensitivity of the shape to changes in the coefficient values, and (3) difficulty in fitting discontinuous profiles.

For many applications (including ours) a simple yet adequate form consists of a piecewise linear approximation consisting of  $N$  segments, where  $N$  is some fixed preselected number for that application. The slopes and endpoints of the segments may also be constrained by *a priori*, partial knowledge. For our application, we divide the depth profile into two halves – the left half is illustrated by Fig. 2, while the right half is a (horizontal) mirror reflection of this plot. The left half is reconstructed from the left image and center image pair, and right half is reconstructed from the center image and right image pair (see Fig. 3). For each half we approximate the profile using three linear segments. A variety of shapes may be modeled by varying the locations of the endpoints and the slope of each segment. The only *a priori* knowledge we use to constrain these parameters is that the depth profile is higher in the middle than at either end, and each end of the profile may consist of a short, horizontal segment.

Let  $D(y)$  denote the height of the imaged surface (with respect to some fixed plane that is parallel to the image plane of the cameras), and along a selected epipolar line, the distance along this line being denoted by  $y$ . We divide  $D(y)$  into  $N$  segments:  $D_i(y^i), i = 1, 2, \dots, N$ , where  $y^i$  is the local coordinate for the  $i^{\text{th}}$  segment (Fig. 2). We use the following linear model for each segment.

$$D_i(y^i) = D_{i-1}(Y'_{i-1}) + S_i + K_i y^i, \quad y^i \in [0, Y'_i) \quad (1)$$

where  $Y'_i$  is the delimiter (or *breakpoint*) for this segment,  $K_i$  is the slope of this segment, and  $S_i$  is the parameter (henceforth called the *shift*) that represents the discontinuity occurring at the

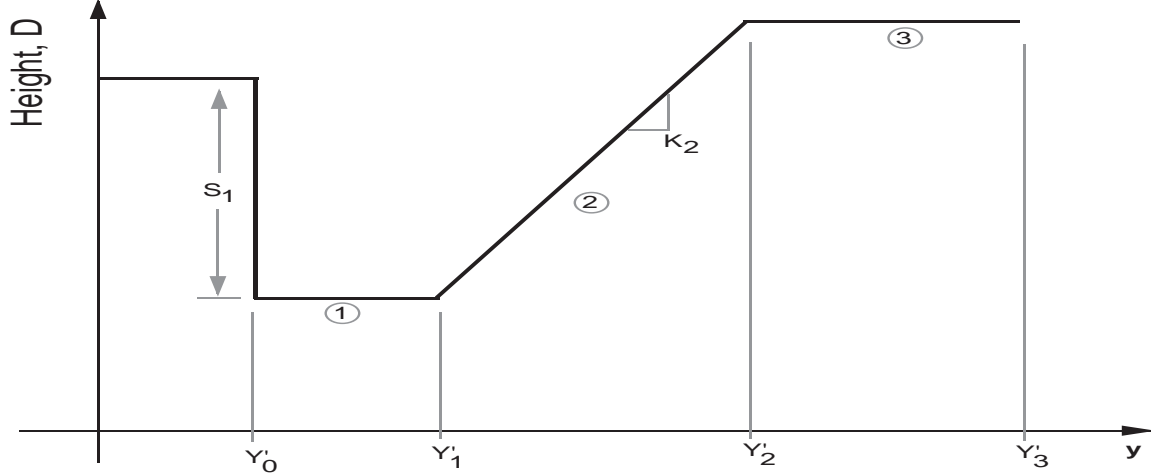


Figure 2: Piecewise linear depth/height model: The depth profile is divided into several segments where  $S_i$  is called the *shift* which represents the discontinuity occurring at the beginning of  $i^{\text{th}}$  segment,  $K_i$  is the slope, and  $Y_i'$  is the delimiter (or *breakpoint*) for this segment.

beginning of this segment. When  $S_i \neq 0$ , it indicates the presence of occlusion, and the image in which this occurs depends on the sign of  $S_i$ .

As mentioned earlier, the search for the optimal parameters of the model consists of hypothesizing the values for each within a search space, then performing a projective and photometric correction (refer to section 3.4) to one of the images, say the left image in the stereo pair, based on the hypothesized shape and known calibration parameters for the cameras. The right image and the corrected left image are then compared to determine the validity of the hypothesized shape. During this process, a significant computational advantage is gained if the variation in *disparity*, rather than the height, along the epipolar line is assumed to be piecewise linear. This assumption is justified if the variation in depth is much less than the distance between the cameras and the scene, as discussed below.

Let  $D_{ref}$  denote the depth of some reference point in the scene. The disparity produced by this

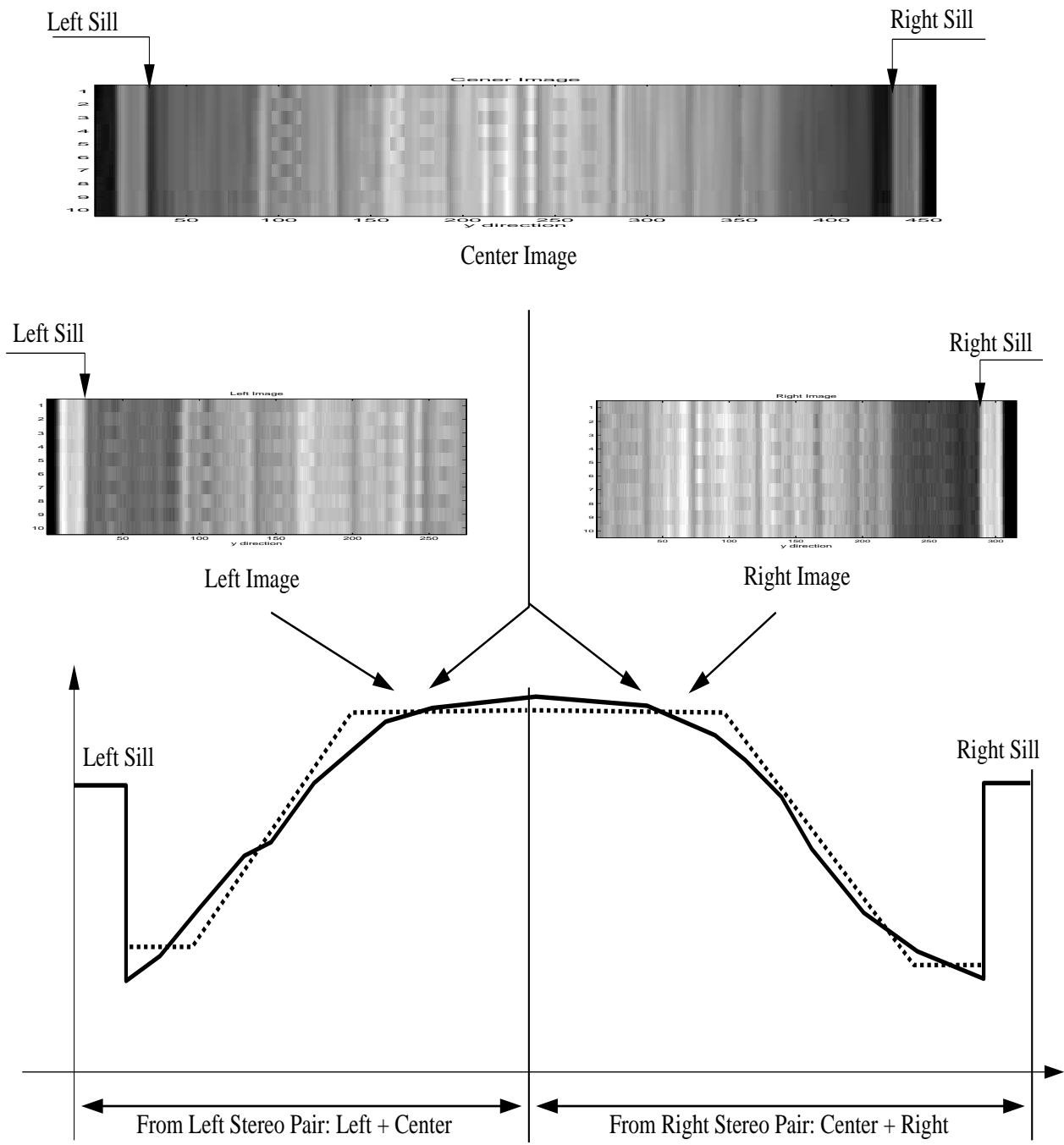


Figure 3: Computing height profile using two pairs of images obtained when viewing coal in a rail car. Ten horizontal scan lines near the epipolar line are used. The bottom figure shows the piece-wise linear fit (dashed line) and the final computed profile (solid line).

point is given by

$$d_{ref} = \frac{fB}{D_{ref} + f} \quad (2)$$

where  $B$  is the baseline and  $f$  is the focal length.

Consider a small increase,  $D_{rel}$ , in the depth relative to  $D_{ref}$ . The new disparity value is given by

$$d = \frac{fB}{D_{ref} + D_{rel} + f} \quad (3)$$

and the corresponding change in disparity is

$$d_{rel} = d_{ref} - d = \frac{fB}{D_{ref} + f} - \frac{fB}{D_{ref} + D_{rel} + f} \simeq \frac{fB}{D_{ref}^2} D_{rel} \quad . \quad (4)$$

The last equality is obtained by simple approximation with assumptions of  $D_{ref} \gg D_{rel}$  and  $D_{ref} \gg f$ . In our application, the maximum value of  $\frac{D_{rel}}{D_{ref}}$  is only 0.1. Due to the above linear relationship between the relative disparity and the relative depth we can use a piecewise linear approximation for the former:

$$d_{rel,i}(y^i) = d_{rel,i-1}(Y'_{i-1}) + s_i + k_i y^i, \quad y^i \in [0, Y'_i) \quad (5)$$

where  $s$  and  $k$  are analogous to  $S$  and  $K$  above.

Partial shape knowledge imposes constraints on the values of the model parameters. Such knowledge may also introduce dependencies between parameters which can be exploited during the search for values of the model parameters. We describe below the use of this model along with limited shape information to compute an approximate, yet reliable, global surface fit, which is then locally refined.

### 3.3 Model-Based Shape Computation

For the sake of simplicity, consider the computation of a 3D height/depth/distance profile given the image intensity profiles from the two images of a stereo pair, and from the same epipolar line. As mentioned above, we first find the best fit of the assumed parametric surface model. This global, approximate surface computation involves a search for the optimal, or sufficiently accurate values of the model parameters. We then refine this approximate, reliably computed surface by performing local correspondence between small areas of the left and right images. During the latter phase, the search space is tightly constrained by the approximate surface computed in the first phase. We describe each phase of this process.

#### *A. Computing Approximate, Global, Parametric Shape*

The computation of the approximate shape uses the shape model described above along with constraints that are imposed by partial knowledge of the expected shape. In our application many of the parameters are constrained by the knowledge that the pile of coal in a rail car is higher in the middle than at either side. This limited knowledge imposes the following specific constraints. In our application, we analyze each half of the depth profile separately - since each pair of the three cameras has a field of view of only one half of the entire scene (Fig. 3). We set  $N = 3$  for each half of the depth profile. Furthermore, the sill of the coal car can be easily extracted in the image pairs. Correspondence between the sill image regions provides the reference distance; thus  $d_{rel,0}(Y'_0)$  for the sill is 0. Since the surface of the coal at the wall of the rail car must be at or below the sill height, we have the constraint  $s_1 \leq 0$ . The profile is assumed to be continuous at the junction of the first and second linear segments, and also at the junction of the second and third linear segments, which translates into the constraints  $s_2 = s_3 = 0$ . Moreover, the first and third segments

are assumed to be horizontal, which imposes the constraints  $k_1 = k_3 = 0$ . The left camera is located either above or outside the left wall of the rail car. The left wall occludes the coal surface that is immediately adjacent to the wall and makes it impossible to reconstruct this surface. Hence, we use the heuristic constraint  $s_1 = -Y'_1$ . The remaining parameters  $k_2$ ,  $Y'_1$ , and  $Y'_2$  are the unknowns that are determined from the stereo image pairs – analyzing one epipolar line at a time. The exhaustive search algorithm is illustrated in Fig. 4. For our application, the final parameters are selected as follows:  $Y_{max}$  values is chosen to be between 20 and 60 (about 1/6 to 1/4 of the whole effective image width). Within this range the algorithm is not sensitive to the choice of  $Y_{max}$ . Since a larger value increases computation time - a value of 20 was chosen for our experiments. Value of  $k_{max}$  is chosen large enough to handle surfaces ranging from flat to maximally tilted – at the expense of high computational cost. Choice of values of  $r1$  and  $r2$  are based on the same criterion as that of  $k_{max}$ . Values of  $q_y$  and  $q_k$  are chosen for high precision. Since an exhaustive search is performed over all combinations of parameter values, a larger parameter range increases the range of surfaces that can be handled, but at the expense of increased computational cost. The above range of parameter values is used for both the center–left image pair and the center–right image pair.

In the search algorithm, for each hypothesized set of parameter values, the intensity profile sensed by the left camera is back-projected onto the hypothesized surface and then projected onto the image plane of the right camera to produce a “corrected” left image. Photometric correction (Section 3.4) is also applied during this process. Ideally, if the hypothesized shape is accurate, then the “corrected” left image and the sensed right image should be identical. Hence, these two are compared to evaluate the validity of the hypothesized shape, viz. model parameters. For computational efficiency we use the sum of absolute differences between pixel intensities to compare the two image profiles, instead of computing the more commonly used correlation coefficient. Any other error measure may also be

1. **Initialization:**  $N = 3, k_1 = s_2 = s_3 = k_3 = 0$ ;
2. **Searching optimal parameters:**
  - for  $Y'_1 = 0, q_y, 2q_y, \dots, Y_{max}$ 
    - Assume  $s_1 = -Y'_1$ ;
    - for  $k_2 = 1, q_k, 2q_k, \dots, k_{max}$ 
      - for  $Y'_2 = r_1 * W, r_1 * W + 1, \dots, r_2 * W$ 
        - Choose intensity profile  $g_r(y)$  along epipolar line of right image.
        - Use model parameters  $(s_i, k_i, Y'_i)$  to specify  $d_{rel,i}$  using equation 5. Hence find  $d_{rel}(y)$  by combining the N segments.
        - Apply projective correction to the left image:
          - $g'_l(y) = g_l(y + d_{rel}(y))$
        - Measure validity of hypothesized parameters using sum of absolute error:
          - $\epsilon(Y'_1, k_2, Y'_2) = \sum_y | \hat{g}_r(y) - \hat{g}'_l(y) |$
          - where  $\hat{g}_r$  and  $\hat{g}'_l$  are normalized version of  $g_r$  and  $g'_l$  respectively.
3. **Select  $s_1, k_2, Y'_2$  that produce the smallest  $\epsilon$  to construct the piece-wise linear disparity profile.**

Figure 4: Search algorithm to find the best model parameters.  $W$  is the width of the entire image along the epipolar line. The values for  $q_y, q_k, r_1, r_2, Y_{max}$ , and  $k_{max}$  are chosen based on the application. For our application we use 1, 0.1, 0.2, 0.8, 20, and 2 respectively.

used in this comparison. It should be pointed out that the search scheme described above is quite naive, and further optimization could improve computational characteristics.

### *B. Refining the Shape*

The above process produces a reliable estimate of global, approximate shape. Reliability is achieved by using a window/area that is the largest possible (the entire image width). The reliability is obtained at the cost of accuracy in the recovered shape. However, the approximate shape forms a good initial condition in a second step of searching for the accurate shape. This search is implemented by a conventional local correspondence search process using an established area based

method, *e.g.* (Smith and Nandhakumar 1993) or (Kanade and Okutomi 1994). During this second step, the search space is tightly constrained to be within a few pixels of the disparity computed by the approximate shape produced by the first step. Also, the search is performed on the projectively corrected images,  $g_r(y)$  and  $g'_l(y)$  specified by the first step. This approach results in a significantly reduced number of false matches.

Thus the two step process – of (1) using partial shape knowledge to produce a reliable but approximate parametric fit for the global shape, and then (2) using the approximate model to limit the search space for the local correspondence process that refines the shape – produces both accurate and reliable estimates of the scene structure.

### 3.4 Preprocessing

The process of establishing correspondence involves the estimation of the degree of similarity between two candidate segments. We have described above a method for overcoming mismatches due to projective distortions between the two segments being compared. Another cause of wrong matches is photometric differences between the two correct candidate segments. We compensate for these photometric differences by preprocessing the images as described below.

#### *A. Image Intensity Variation*

Real surfaces, especially coal, gravel, *etc.*, deviate significantly from Lambertian behavior. In fact, for a flat coal surface the average intensity appears to be roughly proportional to  $\sin(\gamma)$ , where  $\gamma$  is the angle between the surface normal and the reflected ray. When non-frontal surfaces are imaged, especially those exhibiting steep surface slopes, the image intensities in the stereo pair differ significantly. It is, in theory, possible to use an appropriate model for the bidirectional reflectance function of the surface and compensate for this effect. However, the choice of a sufficiently general

model, and the selection of the optimal parameter values for the surface being imaged results in an unnecessarily complex scheme. Even if the surface were a frontal one, the amount of light collected from an elemental patch near the optic axis is higher than that for a patch at the extremities of the image (Haralick and Shapiro 1993). For a wide baseline stereo configuration, an elemental patch may exhibit markedly different intensities in the two images.

We assume that the intensity variation  $I(x, y)$  in a local neighborhood comprises two components -  $I^l(x, y)$  due to local texture, and  $I^g(x, y)$  due to smooth, global variation. If there is no occlusion or foreshortening between the two corresponding regions in a stereo image pair, the intensity variation may be expressed as

$$I_{left}(x, y) = I_{left}^g(x, y) + I_{left}^l(x, y), \quad I_{right}(x, y) = I_{right}^g(x, y) + I_{right}^l(x, y) \quad (6)$$

where the globally varying intensities are related by

$$I_{right}^g(x, y) = f[I_{left}^g(x, y)] \quad (7)$$

Here,  $f(\cdot)$  represents the unknown viewing angle dependency of the sensed light. The local intensity variation is given by

$$I_{right}^l(x, y) = I_{left}^l[x + d_x(x, y), y + d_y(x, y)] + \mathcal{N} \quad (8)$$

$\mathcal{N}$  is assumed to be *white Gaussian* noise that is *uncorrelated* with the intensities produced by the surface texture,  $d(x, y)$  is the disparity map that needs to be computed via stereopsis, and  $d_x$  and  $d_y$  are the  $x$ -direction (*vertical*) and  $y$ -direction (*horizontal*) elements of vector  $d(x, y)$ . Constraining

the y-axis to lie along the *epipolar* line, we may simplify this relationship to be

$$I_{right}^l(x, y) = I_{left}^l[x, y + d_y(x, y)] + \mathcal{N} \quad (9)$$

It is now obvious that it is essentially the local intensity variation that directly contributes to the computing of disparity based on equations 7 and 9.

### B. Texture Preserving Filter

The purpose of the preprocessing step is to minimize the global intensity variation and preserve the local textural variation which is useful in computing correspondences. This facilitates the reliable comparison of large segments of the stereo images intensity profiles. We will call the class of filters having the above property *texture preserving filter*. For different applications the filters might be different. For our case, a simple and intuitively appealing filter is one that subtracts from the original signal a local average value resulting in a computationally efficient high pass filter. In the time domain, the impulse response of the filter is given by

$$h(n) = \delta[n] - \frac{1}{2M} \sum_{k=-M}^M \delta[n - k] . \quad (10)$$

The frequency domain characterization is given by

$$H(e^{jw}) = 1 - \frac{1}{2M} \frac{\sin[w(2M + 1)/2]}{\sin(w/2)} \quad (11)$$

which describes an all-pass filter minus a moving average filter (Fig. 5a).

The performance of this filter is illustrated in Figs. 5b and 5c. The left image intensity profile shown in Fig. 5b contains a slowly varying component. Applying the above filter removes this global

variation as shown in Fig. 5c. To increase robustness, we map the resulting gray level image to a 3 level image utilizing histogram information. Experiments show that this improves the reliability of the algorithm. Thus *complete preprocessing* phase consists of spatial filtering and histogram based requantization. The algorithm could be further improved by using more elegant filters.

### 3.5 Generalizing the scheme for other applications

The new scheme described above can be divided into three steps (Fig. 6): Step I consists of preprocessing the image, by which we hope to eliminate global intensity variation and preserve local textural information. This step improves the performance of the next step which involves fitting an approximate, global, parametric surface. Step II is the key element in our algorithm which consists of choosing an appropriate surface model, surface parameter initialization, and parameter search based on correlation of the reference image<sup>1</sup> and the other stereo image after it has been projectively transformed. Finally, Step III refines the approximate global surface via tightly constrained local correspondence processes. It is important to notice that here we only present a general three step framework and a specific application with naive techniques employed in each step. More appropriate techniques should be used in each step for different applications. A specific application may dictate the choice of a specific surface model and parameters that efficiently model the type of surface most commonly found in that application. For step III, there are many techniques available such as the basic correlation method, cepstral filtering techniques, *etc.* The simplest technique that demonstrates acceptable performance may be chosen.

<sup>1</sup>By reference image we mean the image on which we will build up our disparity map. The other image is the one to which we will apply projective transformation. For convenience, we will assume the right image is the reference image, and the left image is the other image.

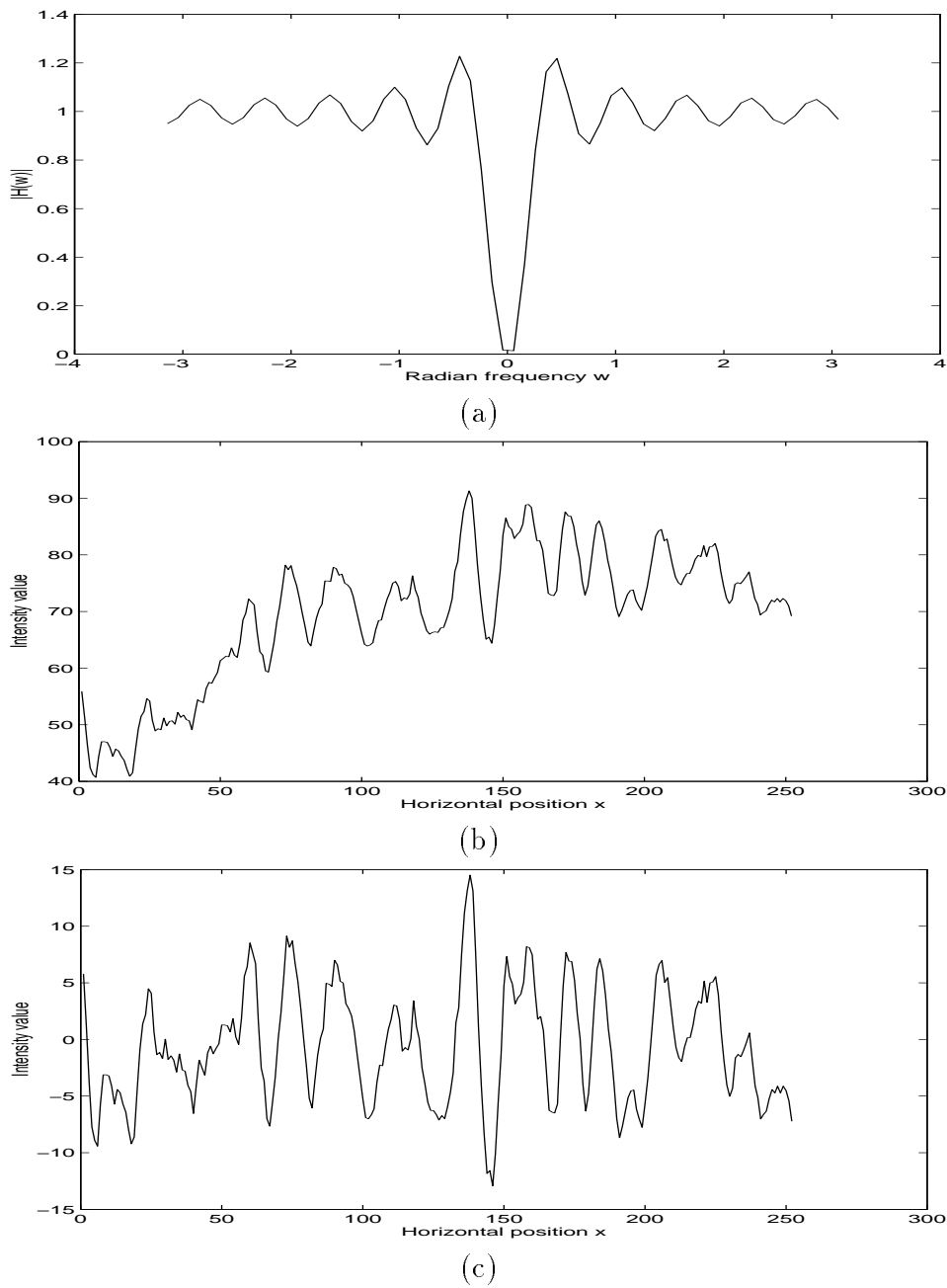


Figure 5: Image Preprocessing: (a) Frequency Response for Texture Preserving Filter, (b) The left image intensity profile along the epipolar line, (c) The intensity profile after preprocessing.

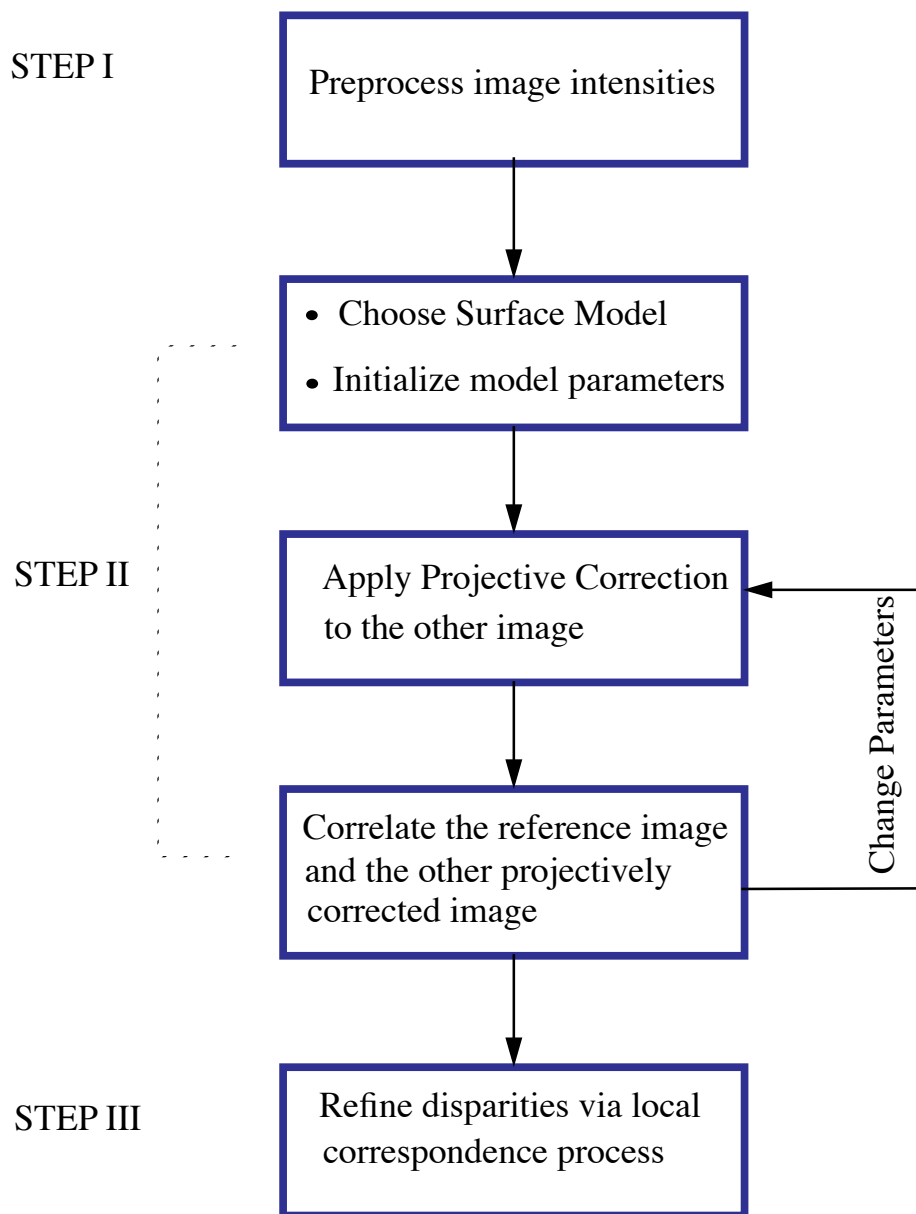


Figure 6: Flow chart of general algorithm

## 4 Experimental Results

We have developed a practical stereoscopic vision system that utilizes the approach described above. The data collection system has been deployed on a “boom-truck” that travels along train routes and collection images from different rail yards through which the train passes. The data are processed off-line on high performance workstations to determine surface height profiles. The system is being used by environmental scientists to study the process of wind erosion of bulk material, such as gravel, iron or coal ore being transported in open rail cars. Since our goal was to develop a fully functional, practical stereo system, we performed extensive experiments to validate the performance. We tested it using data obtained during the day and at night. We also tested it with different types of coal in actual rail cars. Images were acquired from more than 100 rail cars! Since it is very difficult to obtain ground truth regarding the height of the surface of coal in a moving rail car, we developed a simple surface visualization tool for qualitative verification. Also, we tested the system on simulated data as well as data collected from full-size mockups of a transection of the rail car to check the accuracy of the algorithm using the available ground truth information. In the following, we also compare the results obtained by our algorithm with those obtained by the Kanade-Okutomi algorithm.

### 4.1 System Configuration

The new surface extraction scheme has been implemented in the CCLPS system. A downward-looking three-camera parallel-axis system views the top surface of these rail cars to stereoscopically compute a series of 2D depth profiles as the train passes beneath (Fig. 1). These profiles are combined to form a 3D surface height map. Currently CCLPS is a parallel-axis, 1370mm baseline

stereo system which employs three Pulnix TM-745E cameras, which have pixel sensors of size 0.01523mm, with Cosmicar 8.5mm lenses, an Oculus-TCX real time RGB imaging board, and a PC host computer. A diagram of the system is shown in Fig. 7. Prior to use, the cameras are calibrated and aligned in the laboratory (Zhao and Nandhakumar).

The data acquisition is as follows: we freeze the image every  $t_1$  seconds, where  $t_1$  depends on the speed of the moving rail cars and the data transfer rate to hard disk. We then transfer only the central (relative to the piercing point) 10 raster lines of each of the 512x512 images to the hard disk. Since the frame consists of two interlaced fields, each acquired 17 ms apart, only the even field consisting of 5 image raster lines is used for stereo analysis. Image blur is also minimal since the images are acquired when the train is moving at speeds of less than 5mph – the image acquisition is conducted in urban areas near rail yards where the train is forced to be proceed very slowly. The camera rig also carries two high intensity halogen flood lamps which are used to illuminate the surface during night-time imaging. This ensures shutter speeds of 1/500 s or faster and apertures of f 8 or smaller, and hence large depth-of-focus.

The data acquisition system also locates the position of the sill in the image pairs. The raw image contains regions corresponding to the coal, the sill and whatever lies outside of the coal car. We have observed that the sill is always brighter than coal and the inner edge of the sill is easily discernible by the high local contrast that it exhibits. We use the zero-crossings in the Laplacian of Gaussian convolved image to extract the sill, i.e. the edge. This segmentation process extracts individual rail cars from the data set, and also extract only those regions that correspond to coal surface for further analysis. Fig. 8 shows some experimental results of the segmentation process.

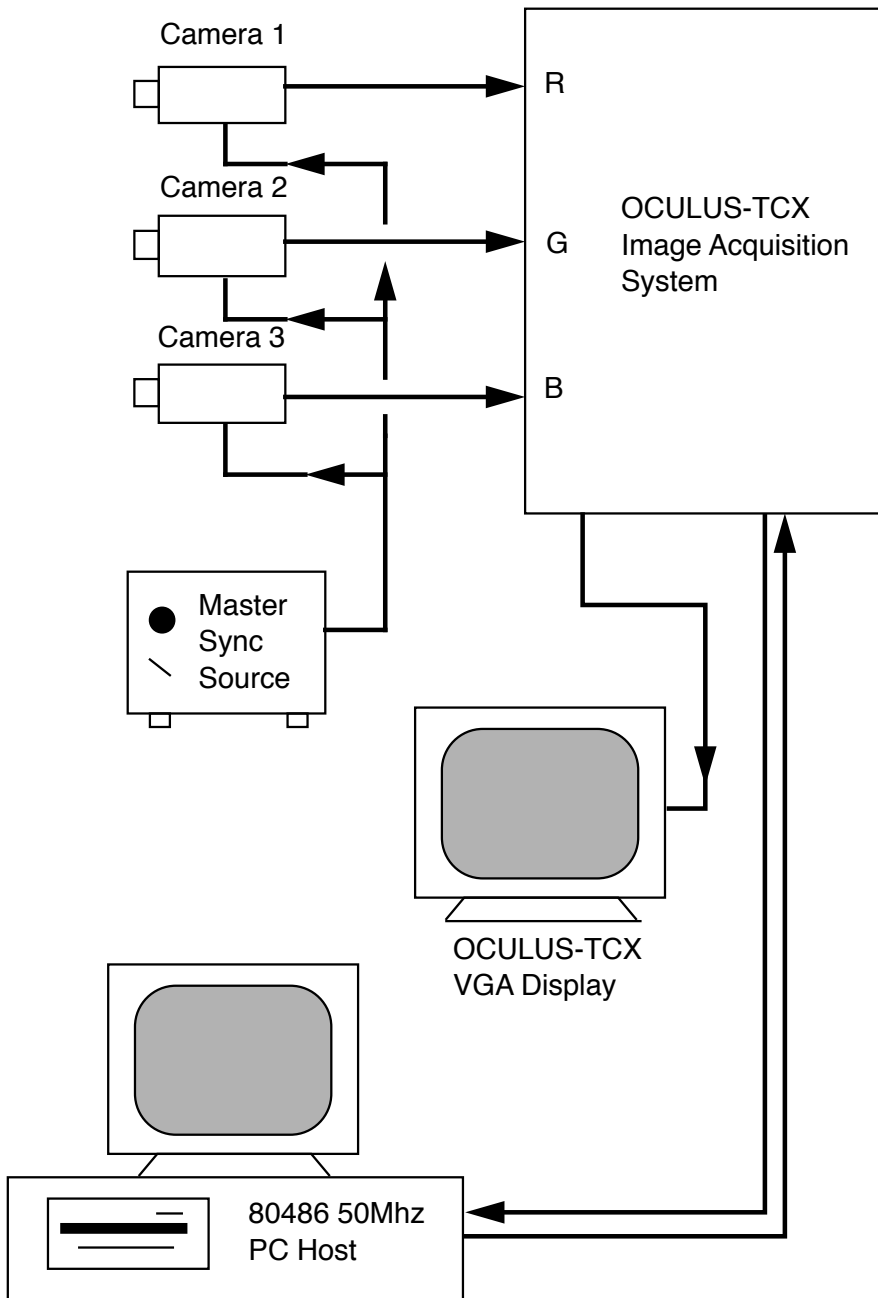


Figure 7: CCLPS System Schematic

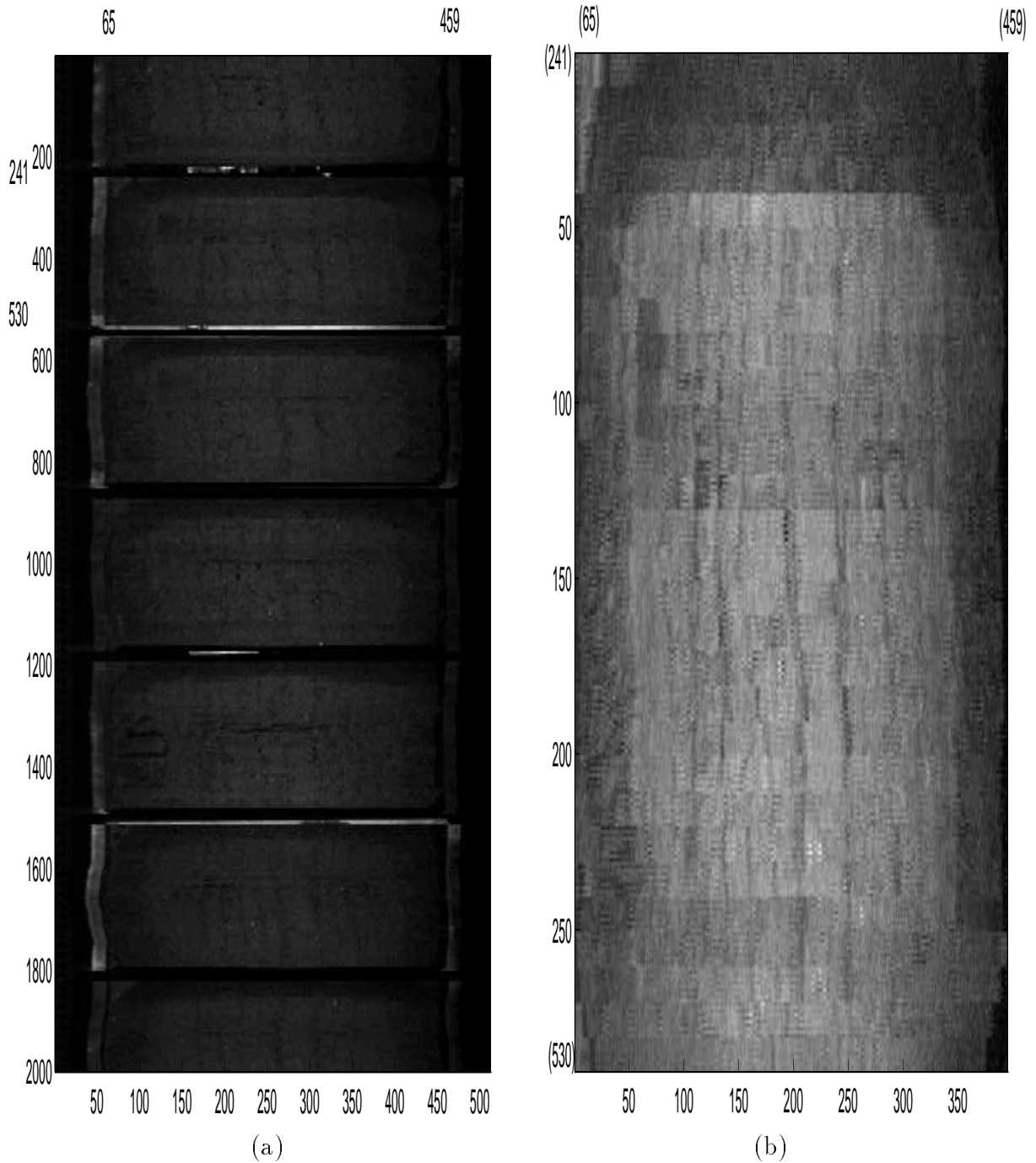


Figure 8: Automatical sill segmentation: (a) the raw image data corresponding to a sequence of 6 coal cars. (b) image of the first coal car after segmentation. This image corresponds to scan lines 241 through 530 of the image in (a), and columns 65 through 459. The horizontal axis is perpendicular to the direction of the rail tracks. The x- and y- axes are labeled in pixel units.

## 4.2 Computing 3D profile

### *A. Results on simulated data*

Since obtaining ground truth of surface height from moving cars is very difficult, we developed a simple test to verify the performance of our algorithm. We use one real image of a coal surface, which is assumed to be the left image of the stereoscopic image pair. The stereoscopic camera calibration parameters are assumed to be the same as for the real system. A virtual surface is assumed to be of some polynomial form (thick curve in Fig. 9a). Using equation 5, the real image is spatially scaled according to this surface to create the simulated right image. The two images are used as inputs to our system. Fig. 9 shows the results of applying our algorithm and Kanade-Okutomi (KO) algorithm to such data.

This comparison indicates that for simulated data the KO-stereo algorithm produces somewhat better results. However, it should be pointed out that: (1) the simulation does not contain any photometric variation, and (2) our naive algorithm could be improved if we utilized interpolation during geometric correction. As expected, both algorithms have more error for occluded and slanted areas.

### *B. Results on real data*

This system has also undergone extensive field tests. Images of more than a hundred rail cars were acquired, during the daytime as well as at night. Rail cars carrying different types of coal were imaged - providing images that varied in texture. Fig. 9 shows the images sensed by the three-camera rig suspended above one rail car carrying coal and the surface reconstructed by the system discussed above.

Although it was not possible to obtain the ground truth of the coal height profile from the

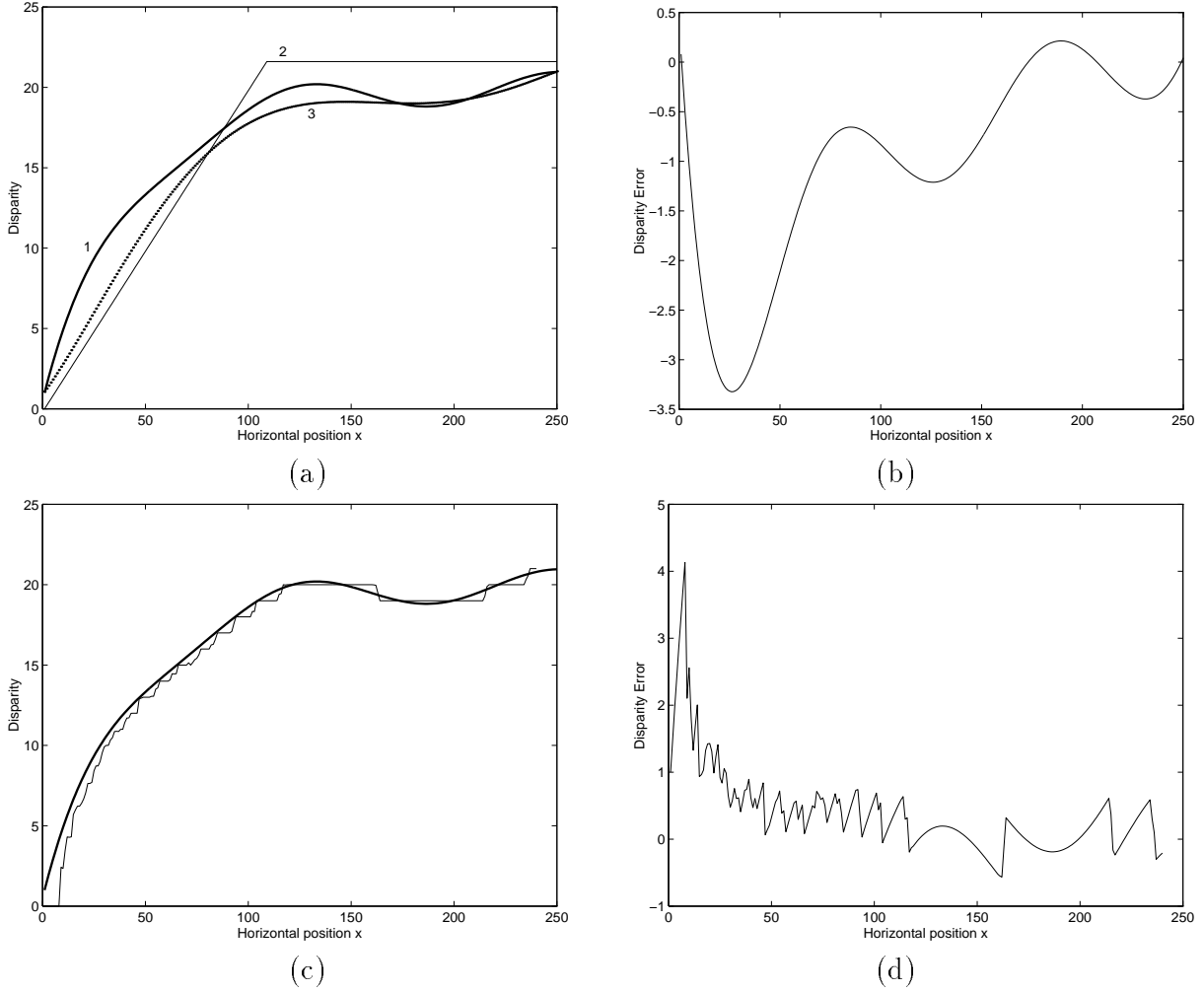


Figure 9: Results using simulated data: (a) and (b) results from our algorithm, (c) and (d) results from the KO-stereo algorithm: **(a)** Thick curve (1) represents the true disparity, thin line (2) shows the piece-wise linear model found (after the first step), dashed curve (3) is the final result after refinement (second step). **(b)** Error between true value and final result produced by our algorithm. **(c)** Thick curve represents the true disparity, thin curve represents the reconstructed curve from KO-stereo algorithm. **(d)** Error between true curve and reconstructed curve for the KO-stereo algorithm. Disparity and horizontal position are in pixel units. Horizontal-axis indicates the distance along the epipolar line.

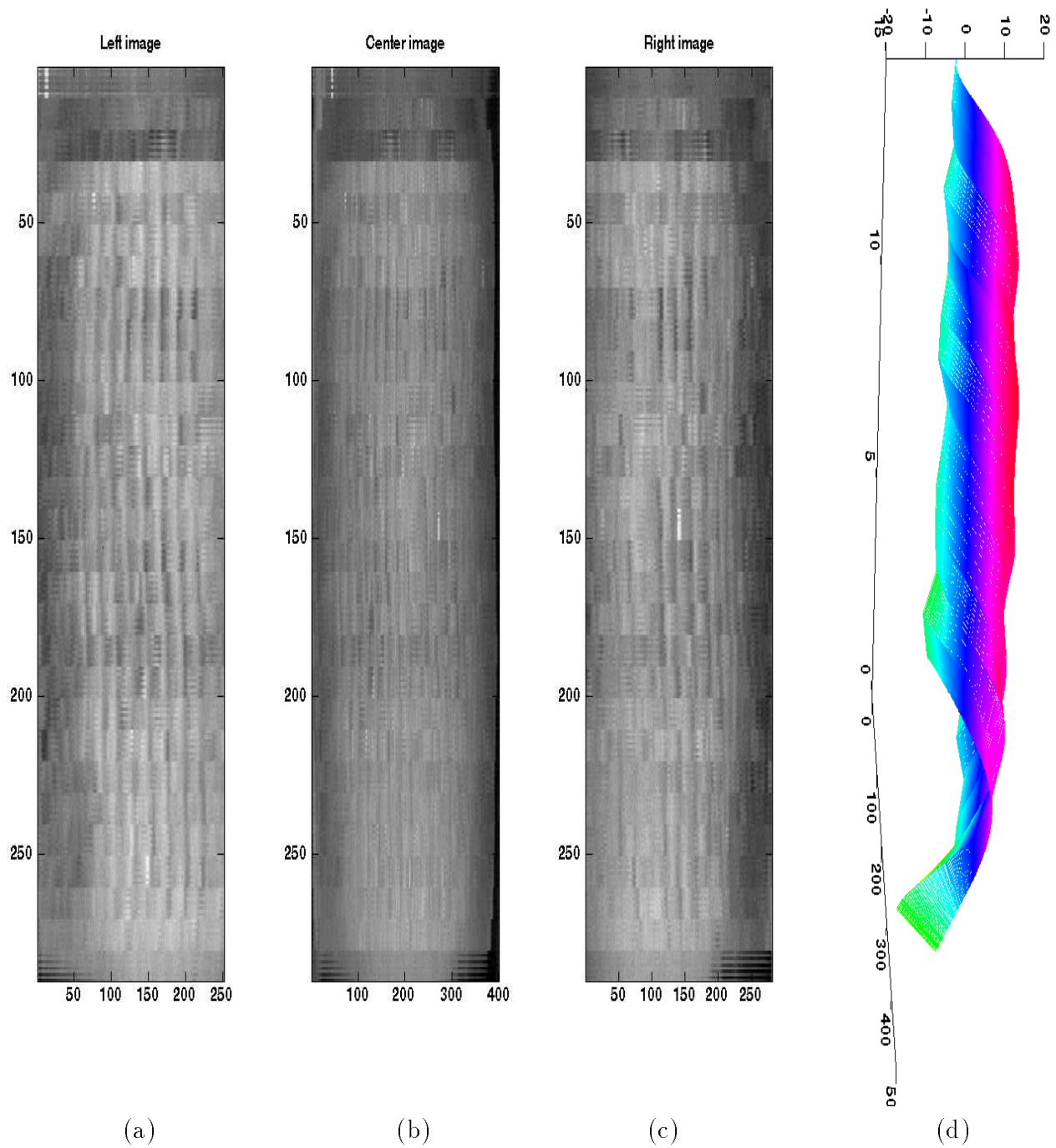


Figure 10: Result from real images of a coal car: (a) Left image (b) Center image (c) Right Image (d) Resulting 3D profiles from both image pairs.

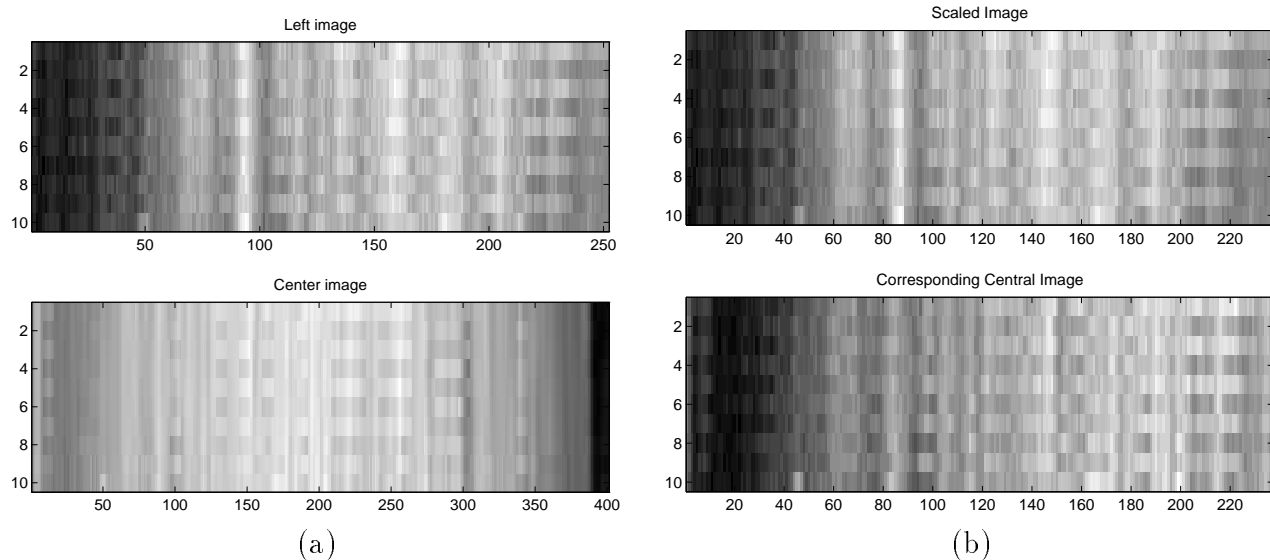


Figure 11: Performance of our algorithm: (a) shows the original image pair (b) shows the image pair after scaling the left image according to the obtained disparity profile from our algorithm and simply cropping the other image.

moving rail cars, it is important to devise a method of evaluating how well our algorithm performs. A qualitative verification is achieved by spatially scaling the image pair based on the disparity profile generated by the algorithm. The two images should appear identical if the generated disparities were accurate. Fig. 11 shows the usefulness of this scheme.

A comparison in Figure 12 shows the results from our algorithm and from the KO-stereo algorithm. As can be seen even without ground truth, our algorithm performs well, while KO-stereo algorithm fails. Possible reasons for the failure of KO-stereo algorithm for this application are: (1) significant portions of the images are occluded, and (2) large photometric variations are present across the images. Hence, the local matching errors accumulate and lead the search astray, while our algorithm overcomes these difficulties by (1) using a global shape model which incorporates a small amount of prior shape information, and (2) preprocessing of the image pair.

### C. Preprocessing

Fig. 13 shows the effect of preprocessing applied to image pairs. For this illustration we used a

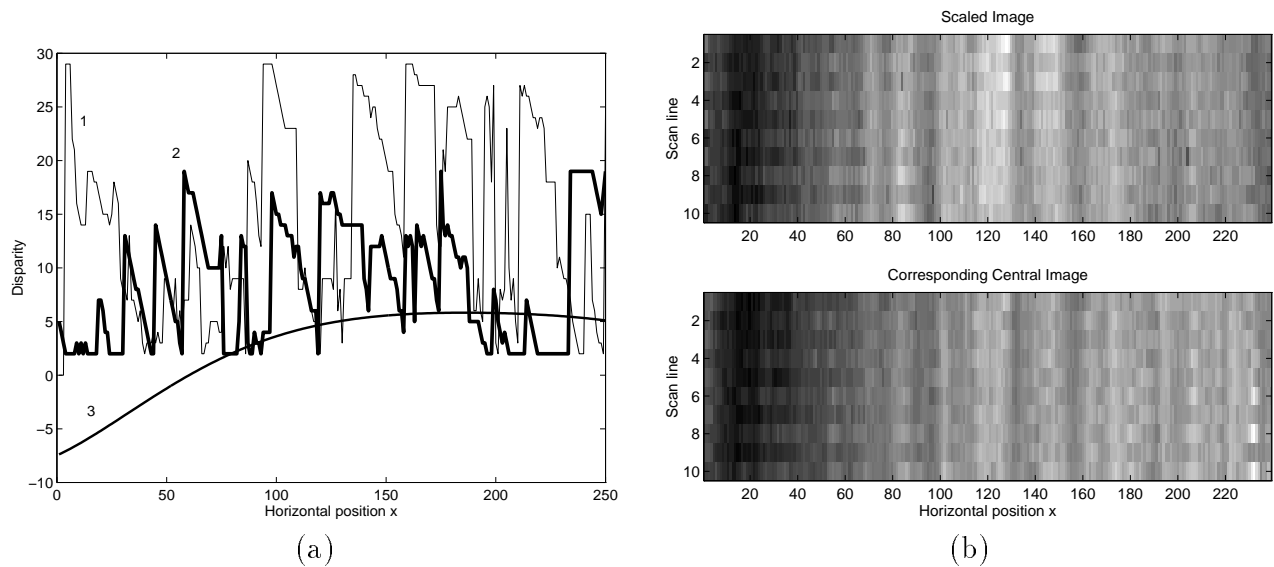


Figure 12: Comparison of our algorithm with KO-stereo algorithm: (a) Curve-1 is a result of the KO-stereo algorithm without manual initialization; curve-2 is a better result obtained from KO-stereo algorithm with manual initialization of the search window to the non-occluded portion of the center image; curve-3, is the disparity profile from our algorithm. Note that the first two plots are very different from the expected profile - unlike curve-3. (b) The image pair after scaling the left image according to the disparity profile obtained from our algorithm - illustrating that the disparity profile produced by our algorithm does indeed produce the correct transformation between the stereo image pair.

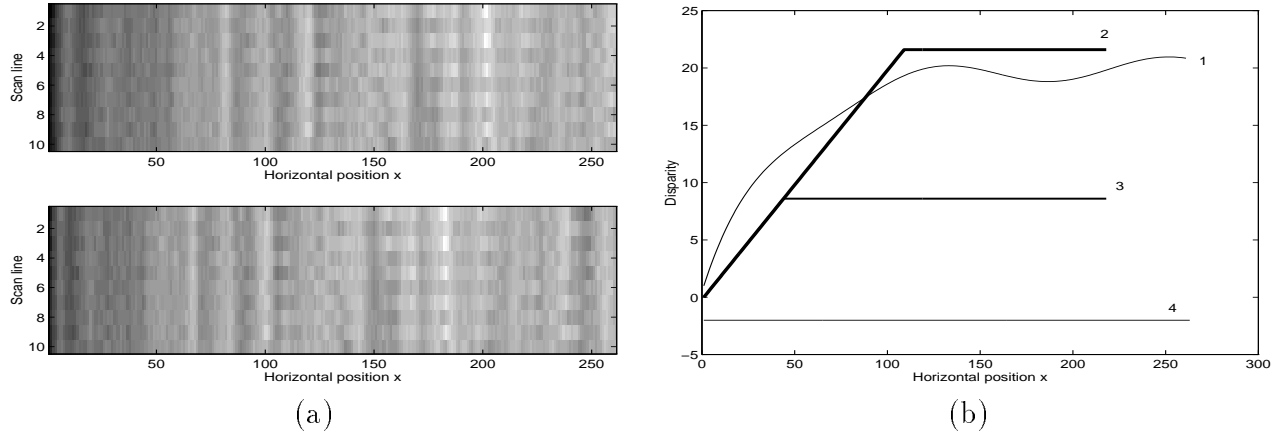


Figure 13: Effect of preprocessing: (a) Image pair before preprocessing (the data are different from that used in 4.2 A). (b) Curve-1 represents the true disparity profile. Curve-2 shows the disparity profile produced by the piecewise-linear model fitting step applied to the spatially filtered and requantized image, Curve-3 shows the profile when only spatial filtering is used (requantization to 3 gray levels is not done). Curve-4 shows the disparity profile produced by the piecewise-linear model fitting step operating on non-preprocessed data.

set of data that were different from that used in section 4.2 A, and ran our algorithm on the filtered gray level image pair, preprocessed 3-gray-level image pair, and on the non-preprocessed data. The improvement due to preprocessing is obvious.

## 5 Conclusion

We have developed a scheme for reliable and accurate surface reconstruction when the stereoscopic images contain only fine texture and no stable high level features. The approach employs a global, parametric model, which is application dependent, to produce an initial approximation to the surface. The surface model explicitly incorporates foreshortening & surface discontinuities. We then refine this model by performing local correspondence. We show, through our application, that with very limited information regarding the surface shape (e.g., the surface is higher in the middle than on either side) and very simple techniques a significant improvement is achieved in the

accuracy and reliability of the reconstruction.

As previously mentioned, our method can be easily adapted for different applications. Specific constraints are imposed by specific applications on the parameters of the shape model. Having chosen the appropriate constraints one can adopt the procedure described above to get reliable and accurate results.

Currently, the surface extraction is performed off-line on stored data - and computational performance has been satisfactory for profiling a number of rail cars each week - as required by the users of this system. The current implementation of the algorithm using MATLAB on a UNIX workstation requires a total 35 seconds of CPU time for processing one transection. The preprocessing step consumes 5.7 seconds, the piece-linear model fitting process consumes 2.5 seconds, and the shape refining process consumes 26.8 seconds. The possible improvement of the current algorithm include:

1. Choose appropriate smaller parameters to speed up processing.
2. Utilize interpolation, more sophisticated matching measure and shape refining scheme to improve accuracy.
3. Use a smarter search strategy.
4. Employ more elegant filters for preprocessing.

## References

- [1] Adelson EH, Wang JYA (1992) Single lens stereo with a plenoptic camera. *IEEE Trans Pattern Analysis Machine Intelligence* 14: 99-106
- [2] Ayache N, Lustman F (1987) Fast and reliable tinocular stereovision. *Proc 1st International Conference on Computer Vision*, pp 422-427

- [3] Baker HH, Binford TO (1981) Depth from edge and intensity based stereo. Proc 7th International Joint Conference on Artificial Intelligence, Vancouver, Canada, pp 631-636
- [4] Bandari E, Little JJ (1993) Multi-evidential Correlation & Visual Echo Analysis. Technical Report 93-1, Laboratory for Computational Vision, Department of Computational Science, University of British Columbia
- [5] Barnard ST, Thompson WB (1980) Disparity analysis of images. IEEE Trans Pattern Analysis Machine Intelligence 10:333-340
- [6] Barnarda ST, Fischler MA (1982) Computational Stereo. ACM Computing Surveys 14:553-572
- [7] Cohen L, Vinet L, Sander PT, Gagalowicz A (1989) Hierarchical Regional Based Stereo Matching. Proc IEEE Conference on Computer Vision and Pattern Recognition , San Diego, CA, pp 416-421
- [8] Dhond UR, Aggarwal JK (1989) Structure from Stereo - A Review. IEEE Trans Systems, Man, and Cybernetics, 19:1489-1560
- [9] Grimson WEL (1981) From Images to Surfaces: A Computational Study of the Human Early Visual System. MIT Press, Cambridge, MA
- [10] Haralick RM, Shapiro LG (1993) Computer and Robot Vision, volume II. Addison-Wesley Publishing Company
- [11] Hassab JC, Boucher R (1975) Analysis of Signal Extraction, Echo Detection, and Removal by Complex Cepstrum in the Presence of Distortion and Noise. Journal of Sound and Vibration 40:321-335
- [12] Hoff W, Ahuja N (1989) Surfaces from Stereo: Integrating Feature Matching, Disparity Estimation, and Detection. IEEE Trans Pattern Analysis Machine Intelligence 11:121-136
- [13] Kanade T, Okutomi M (1994) A Stereo Matching Algorithm with an Adaptive Window: Theory and Experiment. IEEE Trans Pattern Analysis Machine Intelligence 16:920-932
- [14] Kass M, Witkin A, Terzopoulos D (1987) Snakes: Active Contour Models. International Journal of Computer Vision, 1(4):321-331
- [15] Kim HS, Binford TO (1987) Stereo correspondence: A hierarchical approach. Proc DARPA Image Understanding Workshop, Los Angeles, CA, pp 234-241
- [16] Lee CY, Cooper DB, Keren D (1993) Computing Correspondence Based on Regions and Invariants without Feature Extraction and Segmentation. Proc IEEE Conference on Computer Vision and Pattern Recognition, pp 655-656
- [17] Marapane SB, Trivedi MM (1994) Multi-Primitive Hierarchical (MPH) Stereo Analysis. IEEE Trans Pattern Analysis Machine Intelligence 16:227-240
- [18] Marr D, Nishihara HK (1978) Visual Information Processing: Artificial Intelligence and the Sensorium of Sight. Technology Review 81

- [19] Marr D, Poggio T (1979) A computational theory of human stereo vision. Proc Royal Society, London, vol B 204, pp 301-328
- [20] Medioni G, Nevatia R (1985) Segment-based matching. Computer Vision, Graphics, Image Processing, 31:2-18
- [21] Nishihara HK (1984) Practical Real-Time Imaging Stereo Matcher. Optical Engineering, 23:536-545
- [22] Olson TJ (1993) Stereopsis for Verging Systems. Proc IEEE Conference on Computer Vision and Pattern Recognition, pp 55-60
- [23] Oppenheim AV, Schafer R (1993) Discrete-time signal processing. Prentice Hall
- [24] Smith PW, Nandhakumar N (1993) An Accurate Stereo Correspondence Method for Textured Scenes Using Improved Power Cepstrum Techniques.  
Proc IEEE Conference on Computer Vision and Pattern Recognition, pp 651-652
- [25] Smith PW, Nandhakumar N (1994) An Automated Stereoscopic Coal Profiling System - CCLPS. Proc IEEE Workshop on Applications of Computer Vision, pp 10-17
- [26] Terzopoulos D (1988) The computation of visible-surface representations. IEEE Trans Pattern Analysis Machine Intelligence 10:417-437
- [27] Wood GA (1983) Realities of automatic correlation problem. Photogrammetric Engineering and Remote Sensing 49:537-538
- [28] Yeshurun Y, Schwartz (1989) Cepstral Filtering on a Columnar Image Architecture: A Fast Algorithm for Binocular Stereo Segmentation. IEEE Trans Pattern Analysis Machine Intelligence 11:759-767
- [29] M.S. Lew, T.S. Huang and K. Wong (1994) Learning and feature selection in stereo matching. IEEE Trans Pattern Analysis Machine Intelligence 16:869-881
- [30] W. Zhao and N. Nandhakumar (to appear) Effects of Camera Alignment Errors on Stereoscopic Depth Estimates. Pattern Recognition.