Published in final edited form as:
Image Vis Comput. 2010 April 1; 28(4): 605–613. doi:10.1016/j.imavis.2009.09.015.

A Portable Stereo Vision System for Whole Body Surface Imaging

Wurong Yu and **Bugao Xu**
Department of Human Ecology, The University of Texas at Austin, Austin, TX 78712, USA

Abstract

This paper presents a whole body surface imaging system based on stereo vision technology. We have adopted a compact and economical configuration which involves only four stereo units to image the frontal and rear sides of the body. The success of the system depends on a stereo matching process that can effectively segment the body from the background in addition to recovering sufficient geometric details. For this purpose, we have developed a novel sub-pixel, dense stereo matching algorithm which includes two major phases. In the first phase, the foreground is accurately segmented with the help of a predefined virtual interface in the disparity space image, and a coarse disparity map is generated with block matching. In the second phase, local least squares matching is performed in combination with global optimization within a regularization framework, so as to ensure both accuracy and reliability. Our experimental results show that the system can realistically capture smooth and natural whole body shapes with high accuracy.

Keywords

Whole body scanner; 3D surface imaging; stereo vision; stereo matching; disparity

1. INTRODUCTION

In three-dimensional (3D) body surface imaging, the external shape of the human body is measured using non-contact optical techniques. This kind of technology is conventionally called body scanning, although there is no actual mechanical scanning process in some systems. Over the last two decades, a variety of body scanning systems [1,2] have been developed, and a particular interest has been given to whole body scanning, which has covered a wide spectrum of applications, such as virtual clothing try-on and apparel mass customization [3–5], anthropometric survey for new sizing standards [6–8], realistic human animation [9,10], and body fat estimate in human body composition research [11–13]. Most of the existing systems use the laser scanning or structured light technologies. Detailed discussions of the underlying principles are available in some review articles. For example, Daanen and van de Water [14] provided an overview of eight whole body scanning systems. Istook and Hwang [4] reviewed major body scanning systems with focus on applications in the apparel industry. Lately, multiple-camera systems based on the visual hull method have been exploited to construct the human body from its silhouette images (e.g., [15]).

Recently, we have developed a whole body surface imaging system based on stereo vision technology with a goal to improve portability and affordability and reduce data acquisition time. The basic unit of the system is a stereo head that consists of a pair of cameras and a projector. The two cameras are used to capture images from different perspectives, and the projector is used to cast a random speckle pattern onto the body to create an artificial texture. We have found that the number of stereo units can be reduced to four to provide sufficient data for surface reconstruction of the whole body. The compact configuration has made the system much more portable and affordable than current commercially available systems. Furthermore, images in the multiple cameras can be captured simultaneously so that data errors artifacts caused by slight body movement can be reduced dramatically.

However, the computation in stereo vision is complex and intensive. The challenge lies in stereo matching, which is to establish correct correspondence between two images. Stereo matching has been intensively investigated and remarkable progress has been made over the last three decades [16,17], but the research is still far from completion yet due to at least two important reasons. On one hand, there is no general solution for surfaces that differ in geometry and texture appearance in the scene. On the other hand, stereo vision is an application-oriented problem, and therefore objectives and requirements often vary across different systems. Thus, for a specific application, we need to choose or develop algorithms that best suit for its special requirements, such as the smoothness of the surface, the desired level of geometric details, the texture properties of the scene, the required density of data, and the time efficiency. It is worth noting that the complexity of computation can be reduced by illuminating the subject with a sequence of various projecting patterns. The hybrid of stereo vision and structured light would make stereo matching much easier and more reliable as demonstrated in [18]. However, multiple pattern projection would lose the advantage of instantaneous image acquisition and thus increase a chance of unacceptable inaccuracy in body measurement due to involuntary body movement.

This study focused on developing a novel sub-pixel, dense stereo matching algorithm for whole body surface reconstruction that is able to provide accurate and comprehensive body measurements for various applications. The algorithm is able to accurately segment the body from the background in addition to high precision in matching. Since stereo matching is the most critical factor in determining system performances, detailed description of the algorithm will be covered in this paper.

The remainder of the paper is organized as follows. Section 2 presents the system setup after a brief introduction to the principle of stereo vision. Section 3 describes details of the stereo matching algorithm. Experimental results are given in Section 4, which is followed by conclusions in Section 5.

2. STEREO VISION SYSTEM

The fundamental principle underlying a stereo vision system is passive optical triangulation as demonstrated in Figure 1, where a parallel-axis configuration is used to simplify the discussion. In this ideal configuration, the two image planes are parallel and equidistant to the baseline $\overline{O_l O_r}$, where O_l and O_r are the optical centers of the left and right cameras. For an object point P, its space position can be determined by intersecting two rays $\overrightarrow{O_l p_l}$ and $\overrightarrow{O_r p_r}$, where p_l and p_r are the projections of P in the left and right images, and are located in the same horizontal scanline. If the focal length f and the baseline length b are known, then the depth of P relative to the cameras can be obtained by triangulation,

$$Z = -\frac{bf}{d} = \frac{bf}{x_l - x_r}, \tag{1}$$

where $d = x_r - x_l$ is termed disparity, with x_l and x_r relative to the image centers C_l and C_r, respectively.

During the construction of the whole body surface imaging system, the engineering factors of our major concern are cost, portability, and accuracy. To reduce the cost and shorten the duration of development, we have used off-the-shelf components including cameras and projectors. The basic unit of the system is a stereo head that consists of a pair of cameras and a projector. The projector is used to shed artificial texture onto the body, which will facilitate the process of stereo matching. Multiple stereo heads are needed for full body imaging. Our previous work on a rotary laser scanner [19] indicates that full body reconstruction can be made from two scanning units that are placed in the frontal and rear sides of the subject, respectively. A similar construction has been used in this study. However, two stereo heads are needed to cover each side of the body, due to the limited field of view of the cameras and projectors. Therefore, there are totally four stereo heads in the system. The setup is illustrated in Figure 2. The four stereo heads are mounted on two steady stands. Compared to some existing whole body scanners, the unique configuration of our system has greatly improved its affordability and portability.

A more specific description of the prototype system is given here. We have used four pairs of monochromatic CMOS cameras (Videre Design, Menlo Park, CA) with a resolution of 1280 × 960. The focal length of the cameras is 12 mm. The baseline length is set as 9 mm. We have chosen NEC 575VT LCD projectors (NEC Corp., Tokyo, Japan) since they were one of the few types of portable ultra-short throw projector on the market at the initiation of this project. At a projection distance of 2.3 m, the image size is 1.5 m × 1.15 m. Hence, when two such projectors are used together with a slight overlap, the field of view can be as large as 1.5 m × 2.0 m, which is large enough for the majority of population. A personal computer is used to control the cameras and projectors. The cameras communicate with the computer via IEEE 1394 Firewire, and image acquisition can be completed in 200 ms. An NVIDIA GeForce 6500 dual-port graphics card is used to send a texture pattern to the projectors through a VGA hub (Gefen Inc., Woodland Hills, CA). All components are off-the-shelf and readily available.

3. STEREO MATCHING ALGORITHM

3.1. Overview

At this point, we assume the system has been calibrated, i.e., the intrinsic and extrinsic parameters of each camera have been determined. We also suppose the images are perfectly rectified to abide by the parallel-axis geometry as shown in Figure 1. In this case, disparities only exist in the horizontal direction, which simplifies stereo matching to a one-dimensional searching problem.

Based on the system setup described in the last section, a stereo matching algorithm with sub-pixel precision is needed to recover sufficient geometric details of the body. Additionally, because the system has been designed to capture the frontal and rear views of the body only, some portions of the body are invisible to the cameras. To deal with this issue, we have developed a surface reconstruction algorithm [20] that is capable of filling in the gaps in 3D data caused by occlusions. However, if the boundaries of the body in each view cannot be accurately located, it will be difficult to recover the surface from incomplete data.

Figure 3 presents a flowchart for the developed stereo matching algorithm which involves two major phases. In the first phase, foreground objects are accurately segmented from the background of the scene based on the matching costs and a predefined so-called virtual interface, and meanwhile, a disparity map with integer-pixel precision is computed. In the second phase, the disparity map is iteratively refined to reach sub-pixel precision with local least squares matching followed by global optimization. Details of the algorithm are explained below.

3.2. Matching Cost

The first consideration in developing a matching algorithm is to choose a proper matching metric (similarity or dissimilarity measure between two corresponding pixels). Let $I_l\,(x, y)$ and

I_r (x, y) be the left and right intensity images, respectively, and the left image is taken as the reference image. To take into account unbalanced exposure, gain and contrast of the camera pair as observed in our experiments, we have used normalized cross-correlation (NCC) as the similarity measure, which is defined as

$$\rho(x,y,d)=\frac{\sum_{(u,v)\in W(x,y)}\left(I_l(u,v)-\bar{I}_l(x,y)\right)\left(I_r\,(u+d,v)-\bar{I}_r\,(x+d,y)\right)}{N\sigma_l(x,y)\sigma_r(x+d,y)},\tag{2}$$

where W (x, y) is a correlation window around (x, y) with a total pixel number N, and $\bar{I}_l$ (x, y) $(\bar{I}_r(x, y))$ and $\sigma_l(x, y)$ $(\sigma_r(x, y))$ are the local mean and standard deviation of intensity for the left (right) image. The normalization in the local mean and standard deviation makes NCC less sensitive to photometric distortions [18,21]. Based on $\rho(x,y,d)$, the cost function can be defined by

$$C(x,y,d)=1-\rho(x,y,d).\tag{3}$$

Since $-1 \le \rho(x, y, d) \le 1$, we have $0 \le C(x, y, d) \le 2$. $C(x, y, d)$ is defined in the whole image space and at each possible disparity; this trivariate function is usually termed the disparity space image (DSI) [17]. For the sake of conciseness, we will also denote the cost function as C_p (d) with p being a pixel.

3.3. Foreground Segmentation

In this study, foreground segmentation is related to a class of matching algorithms called layered stereo [22–24], which has received attention lately because it is more effective in dealing with occlusions and discontinuities in the scene. Nevertheless, these existing methods almost exclusively rely on color segmentation. For our application, the natural appearance of the scene is eclipsed by the projection of artificial texture, which makes it difficult to perform segmentation from color, contrast, or texture. However, we can take advantage of enhanced stereo cues, since artificial texture would reduce ambiguity in stereo matching.

3.3.1. Definition of the Energy Function—The problem of foreground segmentation can be formalized in the framework of energy minimization. Let P denote the pixel set of the reference image. We define $L = \{F, B\}$ as a label set with F and B representing the foreground and background, respectively. Then the goal is to find a segmentation (or labeling) $f\,(P)\mapsto L$ that minimizes an energy function $E(f)$ defined on a given stereo image pair I_l and I_r. The energy function $E(f)$ usually consists of two terms [25],

$$E(f)=\sum_{i\in P}D_p\left(f_p\right)+\sum_{(p,q)\in N}V_{p,q}\left(f_p,f_q\right),\tag{4}$$

where $N \subset P \times P$ is the set of all neighboring pixel pairs; $D_p(f_p)$ is derived from the input images that measures the cost of assigning the f_p to the pixel p ; and $V_{p,q}(f_p, f_q)$ imposes the spatial coherence of the labeling between the neighboring pixels p and q.

Here we derive D_p (f_p) from the disparity space image C_p (d). First, we assume the disparity space can be divided into two subspaces: the foreground space and the background space that contain the object and the background, respectively, as shown in Figure 4. We assume there exists a virtual interface between the two subspaces, which is denoted by d^* (P). Now we define

$C_p^F=\min_{d_{\min}\le d\le d_p^*} C_p(d),\ C_p^B=\min_{d_p^*<d\le d_{\max}} C_p(d)$, and thus C^F (P) and C^B (P) represent the minimum surfaces in the foreground and background spaces, respectively. If $C_p^F<C_p^B$, then we can expect that there is a good chance that the pixel p belongs to the foreground. The same applies to $C_p^B<C_p^F$ and the background. Therefore, we can define D_p (f_p) by

$$D_p(f_p)=\begin{cases} C_p^F, & f_p=F \\ C_p^B, & f_p=B \end{cases}. \tag{5}$$

However, the above definition is invalid for pixels that cannot be matched. It usually occurs at occlusions, but can also happen in textureless regions that are usually caused by shadows in our system. For the unmatched pixels, we assign constants to the D_p (f_p),

$$D_p(f_p)=\begin{cases} C_o^F, & f_p=F \\ C_o^B, & f_p=B \end{cases}. \tag{6}$$

Here we set $C_o^B<C_o^F$ to favor the background, since we assume that occlusions and shadows exist in the background.

Now the problem becomes to compute the disparity space image C_p (d), to determine the virtual interface d^* (P), and to detect unmatched pixels. The computation of the C_p (d) is straightforward and can be expedited by the box filtering [21] or running sum algorithm [26], both of which has a time complexity that is independent of the size of matching window.

In most cases, the d^* (P) is not available, since we usually lack the prior knowledge about the structure of the scene. But fortunately, for the body imaging system, the virtual interface can be well defined based on the system configuration, which will be described in the next subsection. For the moment, we assume that the d^* (P) has been determined.

To detect unmatched pixels, we use some conventional methods based on block matching. In block matching, the disparity for each pixel is obtained by searching the minimum in the DSI, i.e.,

$$d_p=\arg\min_d C_p(d), \tag{7}$$

which is equivalent to searching the correlation peak according to Equation (3). However, false matches can occur, because disparities are undefined at occlusions, and matching also may fail in other regions due to image noise, geometric distortion, or insufficient texture. We will take the false matches as unmatched. Three criteria are used for deciding a good match. First, the variation of intensity in the matching window should be above a threshold σ_t, otherwise the definition of NCC (and thus the matching cost) is unstable. Secondly, the correlation value should be greater than a threshold ρ_t. Thirdly, the match should pass the left-right check, which means it is also related to a correlation peak if we take the right image as the reference. There is a tradeoff in setting the parameters σ_t and ρ_t: the larger they are, the more confident we are in decided good matches, but the chance of missing good matches will also increase. Ideally, σ_t should be set above the noise level of image, and ρ_t should be determined by such factors as noise level, degree of perspective distortion, size of matching window, and accuracy of image rectification. But in practice, it is hard to optimize these parameters by incorporating the above-mentioned factors, so in our experiments, they are set empirically.

Now we consider the spatial coherence term in Equation (4). Since there are only two states in the label space L, the Potts model [25] can be used, i.e.,

$$V_{p,q}(f_p, f_q)=\begin{cases} \beta_{p,q}, & f_p \neq f_q \\ 0, & f_p=f_q \end{cases}. \tag{8}$$

In the 8-neighborhood system, we set $\beta_{p,q} = \beta_0$ if p and q are horizontal or vertical neighbors, and $\beta_{p,q}=\frac{\beta_0}{\sqrt{2}}$ if they are diagonal neighbors.

3.3.2. Virtual Interface—

The success of the segmentation technique depends on a correct definition of the virtual interface that partitions the disparity space into the foreground and background subspaces. Here we describe how to determine the virtual interface for the developed stereo vision system based on the effective imaging volume, which is defined as the volume in which the body can be fully captured by the system. It is located in between the two imaging stands as shown in Figure 2. According to the optical geometry of the system and the body sizes of the majority of population, the dimensions of the effective imaging volume is set as 1200 mm × 2000 mm × 800 mm (width × height × depth), as illustrated in Figure 5. The origin of the world coordinate system, O_w, is at the center of the floor plane of the volume, and the positive Z_w -axis points to the frontal stereo heads. The space within the volume should be clear except the subject during imaging, and any external object should be ignored by the matching algorithm. Thus, we can use the virtual walls of the volume to divide the 3D space into the foreground and background. In practice, the two side walls are not required because objects beyond them are invisible to the cameras. The necessary floor, roof, front and rear walls are indexed from 0 to 3 in Figure 5. For each stereo head, three of them are applied to segment the foreground from the background. For example, the floor, roof and rear walls are used for the frontal stereo heads.

Nevertheless, we need to convert the interface in the 3D space to the virtual interface in the disparity space. The problem is essentially how to compute the disparity map of a 3D plane. It is well known that a 3D plane in general induces a homography between the two image planes in stereo vision [27,28]. A specific solution for the parallel-axis stereo vision is derived as follows.

In Figure 6, the 3D plane Π is defined in the left camera coordinate system with the normal $\mathbf{n}$ and the perpendicular distance s from the origin O_l. Let $\mathbf{X}_l$ and $\mathbf{X}_r$ be the left and right camera coordinates respectively of an arbitrary point P in Π. For the parallel-axis stereo geometry, the two camera coordinate systems are related by,

$$\mathbf{X}_r=\mathbf{X}_l+\mathbf{t}, \tag{9}$$

where $\mathbf{t} = [-b\ 0\ 0]^{\mathrm{T}}$. Since $\mathbf{n}^{\mathrm{T}}\mathbf{X}_l=s, i.e., \frac{1}{s}\mathbf{n}^{\mathrm{T}}\mathbf{X}_l=1$, it yields

$$\mathbf{X}_r=\mathbf{X}_l+\mathbf{t}\frac{1}{s}\mathbf{n}^{\mathrm{T}}\mathbf{X}_l=\left(\mathbf{I}_{3\times3}+\frac{1}{s}\mathbf{t}\,\mathbf{n}^{\mathrm{T}}\right)\mathbf{X}_l=\mathbf{H}\,\mathbf{X}_l, \tag{10}$$

with

$$\mathbf{H}=\mathbf{I}_{3\times 3}+\frac{1}{s}\mathbf{t}\ \mathbf{n}^{\mathrm{T}}, \tag{11}$$

which is the homograph matrix associated with Π .

Denote $\tilde{\mathbf{x}}_l=\begin{bmatrix} x_l \\ y_l \\ f \end{bmatrix}$ and $\tilde{\mathbf{x}}_r=\begin{bmatrix} x_r \\ y_r \\ f \end{bmatrix}$, which are the homogeneous coordinates of the projections of the point P in the left and right image planes, respectively. According to the perspective projection, we have $\lambda\tilde{\mathbf{x}}_l = \mathbf{X}_l$, and $\lambda\tilde{\mathbf{x}}_r = \mathbf{X}_r$, where λ is a scalar value. Then by replacing $\mathbf{X}_l$ and $\mathbf{X}_r$ in Equation (10), we obtain

$$\tilde{\mathbf{x}}_r=\mathbf{H}\ \tilde{\mathbf{x}}_l. \tag{12}$$

By combining Equation (11) and Equation (12), we have

$$x_r=x_l-\frac{b}{s}\mathbf{n}^T\tilde{\mathbf{x}}_l. \tag{13}$$

As a result, we can compute the disparity by

$$d=x_r-x_l=-\frac{b}{s}\mathbf{n}^T\tilde{\mathbf{x}}_l. \tag{14}$$

In practice, it is more convenient to define the plane Π in the global world coordinate system O_w-$X_\mathrm{w}Y_\mathrm{w}Z_\mathrm{w}$ as described in Figure 5, and then transform it to each individual camera coordinate system according to the camera extrinsic parameters.

3.3.3. Energy Minimization—Belief propagation [29,30] and graph-cuts [25,31,32] are among the state-of-the-art methods to solve labeling problems in computer vision. However, belief propagation can only provide approximate solution when there are loops in the graph (such as a two-dimensional image), even if the label space is binary [33]. In contrast, exact minimum of the energy can be obtained by graph-cuts for a binary segmentation problem [34]. Thus, we use graph-cuts to perform the energy minimization of Equation (4).

Let $G = \langle V, E\rangle$ be a weighted graph. The set V contains the nodes that correspond to the pixel set P and two additional nodes called terminals (the source s and the sink t). The nodes are connected by the edges in the set E .

In construction of the graph for our application, we let s represent the foreground (F), and t be the background (B). As shown in Figure 7, for each node that is associated to a pixel, say p, we connect it to s and t, and denote the edges as e_p^s and e_p^t, respectively. For each pair of neighboring pixels, say $(p, q) \in N$, we connect the corresponding nodes and denote the edge as $e_{p,q}$. The edges are assigned weights (costs) as follows: $c\left(e_p^s\right)=D_p(F)$, $c\left(e_p^t\right)=D_p(B)$, and $c(e_{p,q}) = \beta_{p,q}$. A cut $S \mid T$ is defined as a partition of the nodes in V into two disjoint sets S and T, subject to $s \in S$ and $t \in T$. The cost of $S \mid T$ is the sum of costs of all edges that go from S to T,

$$c\left(S|T\right)=\sum_{p\in T}c\left(e_p^s\right)+\sum_{p\in S}c\left(e_p^t\right)+\sum_{p\in S\ ,q\in T}c\left(e_{p,q}\right). \tag{15}$$

It is easy to see that the sum of the first two terms in *c*(*S* | *T*) corresponds to the first term of the energy function in Equation (4), and the third term in *c*(*S* | *T*) corresponds to the second term of the energy function. Therefore, the cut *S* | *T* is equivalent to a labeling *f*, and *c*(*S* | *T*) = *E*(*f*). As a result, to minimize the energy function is equivalent to searching for a cut with the minimum cost. According to the theorem of Ford and Fulkerson [35], the minimum cut problem can be solved by computing the maximum flow from the source to the sink. Some implementations of the maximum flow algorithms with polynomial complexities are available in the public domain [25, 31].

Once the foreground is segmented, its pixels are assigned a disparity based on Equation (7). However, the obtained disparity map can be noisy. A median filter [36] is used to quench the impulse noise. Furthermore, morphological close and open operators [37] are used to smooth the contour.

3.4. Disparity Refinement

So far, the disparity map takes discrete values, from which the reconstructed surface will appear as staircases. A disparity refinement process is needed to achieve sub-pixel precision. One of the standard methods is fitting a curve (e.g., parabolic [21] or Gaussian curve [38]) to the matching costs defined at discrete values. However, the curve fitting technique suffers from systematic error called “pixel-locking” effect in which disparity values are pulled towards integers [38]. Some research efforts have been made to address this problem. For example, Shimizu and Okutomi [39] attempted to reduce the bias by performing additional curve fitting on matching costs defined at half-pixel locations. Nehab et al. [40] suggested symmetric refinement by fitting a parametric surface over a 2D neighborhood of the matching cost function. Stein et al. [41] proposed an iterative refinement method that is essentially based on Lucas-Kanade algorithm [42].

It should be noted that the aforementioned improvements are all focused on reducing the “pixel-locking” effect and make disparity refinement on each individual pixel independently. However, in practice, like all other local methods, the refined disparity map is prone to be noisy. Thus, it is necessary to take into account spatial coherence during disparity updating.

Here we have developed a method that iteratively performs disparity refinement at a global level within a regularization framework [43,44]. There are two steps in each iteration: local estimation and global optimization. For the first step, the amount of update is estimated locally for each pixel. The estimation can be made by minimizing the matching cost function defined in Equation (3),

$$\delta d=\arg\min_{\delta d}C(x,y,d+\delta d)=\arg\max_{\delta d}\rho(x,y,d+\delta d), \tag{16}$$

where *d* is the current disparity value, and δ*d* is the amount to be updated. However, the process is difficult since the correlation function ρ is highly nonlinear. Although it is possible to perform linearization of ρ with first-order approximation, the computation is still extensive. So instead, we will apply the sum of squared differences (SSD) as the matching cost as in Lucas-Kanade algorithm [42]. If the SSD takes into account the gain and bias factors between cameras, it is essentially equivalent to normalized cross-correlation. Now the matching cost is defined as

$$C_{SSD}\,(x,y,d)=\sum_{(u,v)\in W(x,y)}\left(I_r\,(u+d,v)-(aI_l\,(u,v)+b)\right)^2, \tag{17}$$

where *a* and *b* are the gain and bias factors, respectively. Here we assume the disparity is constant within the matching window *W*. But this assumption is generally not true except for frontal-parallel surfaces. To allow the disparity to vary within the window, we first warp the right image based on the current disparity map,

$$\widehat{I_r}\,(x,y)=I_r\,(x+d\,(x,y)\,,y)\,. \tag{18}$$

To estimate δ*d*, *a* and *b*, we define an error function with $\hat{I}_r$ based on the SSD,

$$e^2(\delta d,a,b;x,y)=\sum_{(u,v)\in W(x,y)}\left(\widehat{I_r}(u+\delta d,v)-(aI_l(u,v)+b)\right)^2. \tag{19}$$

With a first-order approximation, we get

$$e^2(\delta d,a,b;x,y)=\sum_{(u,v)\in W(x,y)}\left(\widehat{I_r}(u,v)+\widehat{I_{rx}}(u,v)\delta d-(aI_l(u,v)+b)\right)^2, \tag{20}$$

where $\widehat{I_{rx}}=\frac{\partial\widehat{I_r}}{\partial x}$ is the intensity gradient of the warped right image.

Let $\mathbf{p}=[\delta d\ a\ b]^{\mathrm{T}}$, $\mathbf{a}=[I_{rx}\ -I_1\ -1]^{\mathrm{T}}$, then a concise form of Equation (20) is

$$e^2(\mathbf{p})=\sum\left(\mathbf{a}^T\mathbf{p}+I_r\right)^2. \tag{21}$$

This is a classic least squares problem. To minimize e^2 (**p**)is equivalent to solve the normal equations,

$$\mathbf{A}\mathbf{p}=\mathbf{b}, \tag{22}$$

where $\mathbf{A}=\sum\mathbf{a}^T\mathbf{a}$, and $\mathbf{b}=-\sum I_r\mathbf{a}$.

We have described how to estimate δ*d* at each pixel. Now we show how to update the disparity map at a global level. First, an energy function is defined by

$$E(d)=\iint\left(d(x,y)-\tilde{d}(x,y)\right)^2dxdy+\lambda\iint\left(d_x^2+d_y^2\right)dxdy, \tag{23}$$

where $\tilde{d}$ is the local estimate of the disparity, and d_x, d_y are the disparity gradients. The first term in the equation measures the consistency with the local estimation, and the second term imposes smoothness constraints on the solution. λ is called the regularization parameter that weighs the smoothness term.

For the *n*-th iteration, we set $\tilde{d}^n = d^{\,n-1} + \delta d^{\,n}$. Then the discrete form of $E(d)$ can be expressed as

$$E\left(d\right)=\sum_{(i,j)\in I}\left(\left(d^{n}\left(i,j\right)-\left(d^{n-1}\left(i,j\right)+\delta d^{n}\left(i,j\right)\right)\right)^{2}+\lambda\left(\left(d^{n}\left(i+1,j\right)-d^{n}\left(i,j\right)\right)^{2}+\left(d^{n}\left(i,j+1\right)-d^{n}\left(i,j\right)\right)^{2}\right)\right), \tag{24}$$

where (i, j) is the discrete coordinates of a pixel in the image plane I, and the discrete gradients are computed using the forward difference. Minimizing the energy function yields

$$\left(1+k_p\lambda\right)d_p^n-\lambda\sum_{q\in N(p)}d_q^n=d_p^{n-1}+\delta d_p^n \tag{25}$$

for each pixel p whose number of neighboring pixels is $k_p = |N(p)|$. Then we can establish a linear system

$$\mathbf{P}\,\mathbf{d}=\mathbf{h}, \tag{26}$$

where $[\mathbf{P}]_{p,p}=1+k_p\lambda$, $[\mathbf{P}]_{\substack{p,q\\ p\neq q}}=\begin{cases}-\lambda, & q\in N(p)\\ 0, & \text{otherwise}\end{cases}$, $[\mathbf{d}]_p=d_p^n$, and $[\mathbf{h}]_p=d_p^{n-1}+\delta d_p^n$. Since **P** is a sparse, positive, symmetric matrix, the solution can be searched efficiently using the conjugate gradient method [45].

4. EXPERIMENTAL RESULTS AND DISCUSSION

4.1. Qualitative Evaluation

The stereo vision system has been used to capture both static objects and human subjects for performance evaluation. The cameras were carefully calibrated using the techniques described in [46], and the images were rectified prior to stereo matching. The same set of parameters was used throughout the test, as listed in Table 1. The virtual interface in the disparity space was created for each stereo head according to its calibration parameters. The results on the four views of a human subject are shown in Figure 8, where the rectified image pair is shown in (a) and (b); the coarse and refined disparity maps are shown in (c) and (d), respectively; and to better evaluate the performance of object segmentation, the refine disparity map has been overlaid onto the reference (left) image as shown in (e). The disparities are coded with the standard cold-to-hot color mapping that corresponds to “far-to-close” to the cameras. Note that the images and disparity maps have been rotated 90° clockwise for better display. The results show that the algorithm is capable of recovering geometric details and is effective in foreground segmentation even if the background is non-static and relatively complex.

A disparity map can be easily converted into a 3D point cloud by Equation (1), and then a complete 3D surface model can be reconstructed by merging point clouds from all four stereo heads. A deviation from a perfect alignment between the upper and lower stereo heads often exists, and a slight separation in the overlapping regions of the point clouds may occur as shown in Figure 10. Thus, a smooth weighting function [47] is applied to fuse the upper and lower data sets prior to converting them to a single mesh.

The surface reconstruction algorithm [20] developed in our previous study for a two-view body scanner was slightly modified to be used in this project. In this method, a 3D model is reconstructed by fitting a subdivision surface to the raw data. It has been verified that the algorithm is effective in creating a smooth surface with automatic gap closing. To demonstrate the performance of the whole body imaging system, some reconstructed body models with various shapes and sizes are shown in Figure 9, where the first model corresponds to the subject in Figure 8. The results indicate the system can realistically capture smooth and natural whole body shapes.

4.2. Quantitative Analysis

To quantitatively estimate the matching precision, a planar surface target was placed at the center of the effective imaging volume and imaged by one of the stereo heads. A region with a size of around 250 mm × 200 mm was used for surface fitting, and the residual errors were used as a measure of matching precision. The results are displayed in Figure 11. The standard deviation of the residual errors is about 0.7 mm, which provides an estimate of depth resolution of the system.

To evaluate the overall accuracy of the system, it was tested on a mannequin, which is a size-12 Wolf body form widely used for apparel design. A MyoTape body tape measure (AccuFitness, LLC, Greenwood Village, CO) was used to measure the waist and hip circumferences, and an anthropometer (Lafayette Instrument Company, Lafayette, IN) was used to measure the depth and breadth of the waist. The waist and hip were located according to the criteria as indicated in Figure 12. The mannequin was imaged 10 times with repositioning in a given hour period. The waist and hip circumferences, and waist breadth and depth were measured on 3D data automatically, and the results were compared to those obtained with manual methods. As shown in Table 2, although there were significant differences in three of the four measures, the differences were very small. The difference in hip circumference was relatively large because it was difficult to determine the location consistently.

Lastly, the time complexity of the stereo matching algorithm was considered. The test was carried out on a personal computer with an AMD Athlon™ 2.0 GHz dual-core CPU and 1.0 G RAM. Because, in principle, the computation for the four stereo heads can been undertaken in parallel, we can take advantage of the multithreading function of a multi-core CPU to improve the time efficiency by simply separating the problem into a thread per stereo pair for computing the disparity maps. In our system, two pairs of images can be matched simultaneously, and it takes about 80 s to complete the computation for a full set (four pairs) of images.

5. CONCLUSIONS

In this paper, we have presented a stereo vision system for whole body surface imaging. The system consists of four stereo units to cover the frontal and rear views of the body. The off-the-shelf components, the fast image-acquisition method, and the robust two-phase stereo matching algorithm have made the system highly portable, reliable and economical. The experimental results have shown that the system can realistically capture whole body shapes with high accuracy. The portability and low cost of the system make it promising for applications such as a large-scale anthropometric survey to collect body dimensions for apparel design and human engineering, and routine use in clinical settings and health clubs for body composition assessment.

REFERENCES

1. Robinette KM, Vannier MW, Jones PRM. 3-D Surface Anthropometry: Review of Technologies. AGARD, Neuilly-sur-Seine. 1997
2. D'Apuzzo, N. State of the art of the methods for static 3D scanning of partial or full human body. Proceedings of Conference on 3D Modeling; Paris, France. 2006 Jun 13–14.

3. Paquette S. 3D Scanning in apparel design and human engineering. IEEE Computer Graphics and Applications 1996;16(5):11–15.

4. Istook CL, Hwang S-J. 3D body scanning systems with application to the apparel industry. Journal of Fashion Marketing and Management 2001;5(2):120–132.

5. Cordier F, Hyewon S, Magnenat-Thalmann N. Made-to-measure technologies for an online clothing store. IEEE Computer Graphics and Applications 2003;23(1):38–48.

6. Robinette, KM.; Daanen, H.; Paquet, E. The CAESAR project: a 3-D surface anthropometry survey. Proceedings of the Second International Conference on 3-D Digital Imaging and Modeling; 1999. p. 380-386.

7. Wells JCK, Treleaven P, Cole TJ. BMI compared with 3-dimensional body shape: the UK national sizing survey. American Journal of Clinical Nutrition 2007;85(2):419–425. [PubMed: 17284738]

8. Wells JCK, Cole TJ, Bruner D, Treleaven P. Body shape in American and British adults: between country and inter-ethnic comparisons. International Journal of Obesity 2008;32:152–159. [PubMed: 17667912]

9. Thalmann, D.; Shen, J.; Chauvineau, E. Computer Graphics International, Pohang, Korean. 1996 Jun. Fast realistic human body deformations for animation and VR applications.

10. Magnenat-Thalmann, N.; Seo, H.; Cordier, F. Fourth International Conference on 3-D Digital Imaging and Modeling (3DIM 2003). 2003. Automatic modeling of animatable virtual humans—a survey; p. 2-10.

11. Wells JCK, Douros I, Fuller NJ, Elia M, Dekker L. Assessment of body volume using three-dimensional photonic scanning. Annals of the New York Academy of Sciences 2000;904:247–254. [PubMed: 10865749]

12. Wang J, Gallagher D, Thornton JC, Yu W, Horlick M, Pi-Sunver FX. Validation of a 3-dimensional photonic scanner for the measurement of body volumes, dimensions, and percentage body fat. American Journal of Clinical Nutrition 2006;83(4):809–816. [PubMed: 16600932]

13. Wells JCK, Ruto A, Treleaven P. Whole-body three-dimensional photonic scanning: a new technique for obesity research and clinical practice. International Journal of Obesity 2008;32:232–238. [PubMed: 17923860]

14. Daanen HAM, van de Water GJ. Whole body scanners. Displays 1998;19:111–120.

15. Corazza S, Mundermann L, Chaudhari A, Demattio T, Cobelli C, Andriacchi T. A markerless motion capture system to study musculoskeleta biomechanics: visual hull and simulated annealing approach. Annals of Biomedical Engineering 2006;34(6):1019–1029. [PubMed: 16783657]

16. Scharstein D, Szeliski R. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. International Journal of Computer Vision 2002;47(123):7–42.

17. Brown MZ, Burschka D, Hager GD. Advances in computational stereo. IEEE Transactions on Pattern Analysis and Machine Intelligence 2003;25(8):993–1008.

18. Scharstein, D.; Szeliski, R. High-accuracy stereo depth maps using structured light. Proceedings of IEEE Conference on Computer Vision and Pattern Recognition; 2003. p. 195-202.

19. Xu B, Huang Y. 3D technology for apparel mass customization, Part I: Rotary body scanning. Journal of Textile Institute 2003;94(1):72–80.

20. Yu W, Xu B. Surface reconstruction from two-view body scanner data. Textile Research Journal 2008;78(5):457–466.

21. Sun C. Fast stereo matching using rectangular subregioning 3D maximum-surface techniques. International Journal of Computer Vision 2002;47(123):99–117.

22. Lin M, Tomasi C. Surface with occlusions from layered stereo. Conference on Computer Vision and Pattern Recognition 2003:710–717.

23. Kolmogorov, V.; Criminisi, A.; Blake, A.; Cross, G.; Rother, C. Proc. IEEE Computer Vision and Pattern Recognition (CVPR). San Diego, CA; 2005. Bi-layer segmentation of binocular stereo video.

24. Bleyer M, Gelautz M. A layered stereo matching algorithm using image segmentation and global visibility constraints. ISPRS Journal of Photogrammetry and Remote Sensing 2005;59(3):128–150.

25. Boykov Y, Veksler O, Zabih R. Fast approximation energy minimization via graph cuts. IEEE Transactions on Pattern Analysis and Machine Intelligence 2001;23(11):1222–1239.

26. Lewis JP. Fast Template Matching. Vision Interface 1995:120–123.

27. Hartley, R.; Zisserman, A. Multiple View Geometry in Computer Vision. Cambridge, UK: Cambridge University Press; 2000.

28. Faugeras, O.; Luong, Q-T. The Geometry of Multiple Images. Cambridge, MA: MIT Press; 2001.

29. Sun J, Zheng N-N, Shum H-Y. Stereo matching using belief propagation. IEEE Transactions on Pattern Analysis and Machine Intelligence 2003;25(7):787–800.

30. Felzenszwalb PF, Huttenlocher DR. Efficient belief propagation for early vision. Proc. IEEE CVPR 2004 2004:I-261–I-268.

31. Kolmogorov, V.; Zabih, R. Computing visual correspondence with occlusions using graph cuts. Proceedings of Eighth IEEE International Conference on Computer Vision (ICCV 2001); 2001. p. 508-515.

32. Roy S. Stereo without epipolar lines: a maximum-flow formulation. International Journal of Computer Vision 1999;34(23):147–161.

33. Yedidia JS, Freeman WT, Weiss Y. Bethe free energy, Kikuchi approximations, and belief propagation algorithms. Technical Report TR-2001-16, Mitsubishi Electric Reseach. 2001

34. Greig D, Porteous B, Seheult A. Exact maximum a posteriori estimation for binary images. Journal of the Royal Statistical Society, Series B 1989;51(2):271–279.

35. Ford L, Fulkerson D. Flows in Networks. Princeton University Press. 1962

36. Huang, TS. Two-Dimensional Signal Processing II: Transforms and Median Filters. Berlin: Springer-Verlag; 1981.

37. Dougherty, ER. Bellinghan. Washington: SPIE Press; 2003. Hands-on Morphological Imaging Processing.

38. Westerweel J. Digital Particle Image Velocimetry: Theory and Application. Delft University Press. 1993

39. Shimizu M, Okutomi M. Sub-pixel estimation error cancellation on area-based matching. International Journal of Computer Vision 2005;63(3):207–224.

40. Nehab D, Rusinkiewicz S, Davis J. Improved sub-pixel stereo correspondences through symmetric refinement. International Conference on Computer Vision (ICCV). 2005 October;

41. Stein A, Huertas A, Matthies L. Attenuating stereo pixel-locking via affine window adaptation. IEEE International Conference on Robotics and Automation 2006 May;:914–921.

42. Lucas BD, Kanade T. An iterative image registration technique with an application to stereo vision. Proceedings of Imaging Understanding Workshop 1981:121–130.

43. Poggio T, Torre V, Koch C. Computational vision and regularization theory. Nature 1985;317(26): 314–319. [PubMed: 2413361]

44. Terzopoulos D. Regularization of inverse visual problems involving discontinuities. IEEE Transactions on Pattern Analysis and Machine Intelligence 1986;8(4):413–424.

45. Shewchuk, JR. An Introduction to the Conjugate Gradient Method without the Agonizing Pain. Carnegie Mellon University: Pittsburgh, PA; 1994.

46. Zhang Z. A flexible new technique for camera calibration. IEEE Transactions on Pattern Analysis and Machine Intelligence 2000;22(11):1330–1334.

47. Yu, W. Dissertation of University of Texas at Austin. 2008 Aug. Development of a Three-Dimensional Anthropometry System for Human Body Composition Assessment.

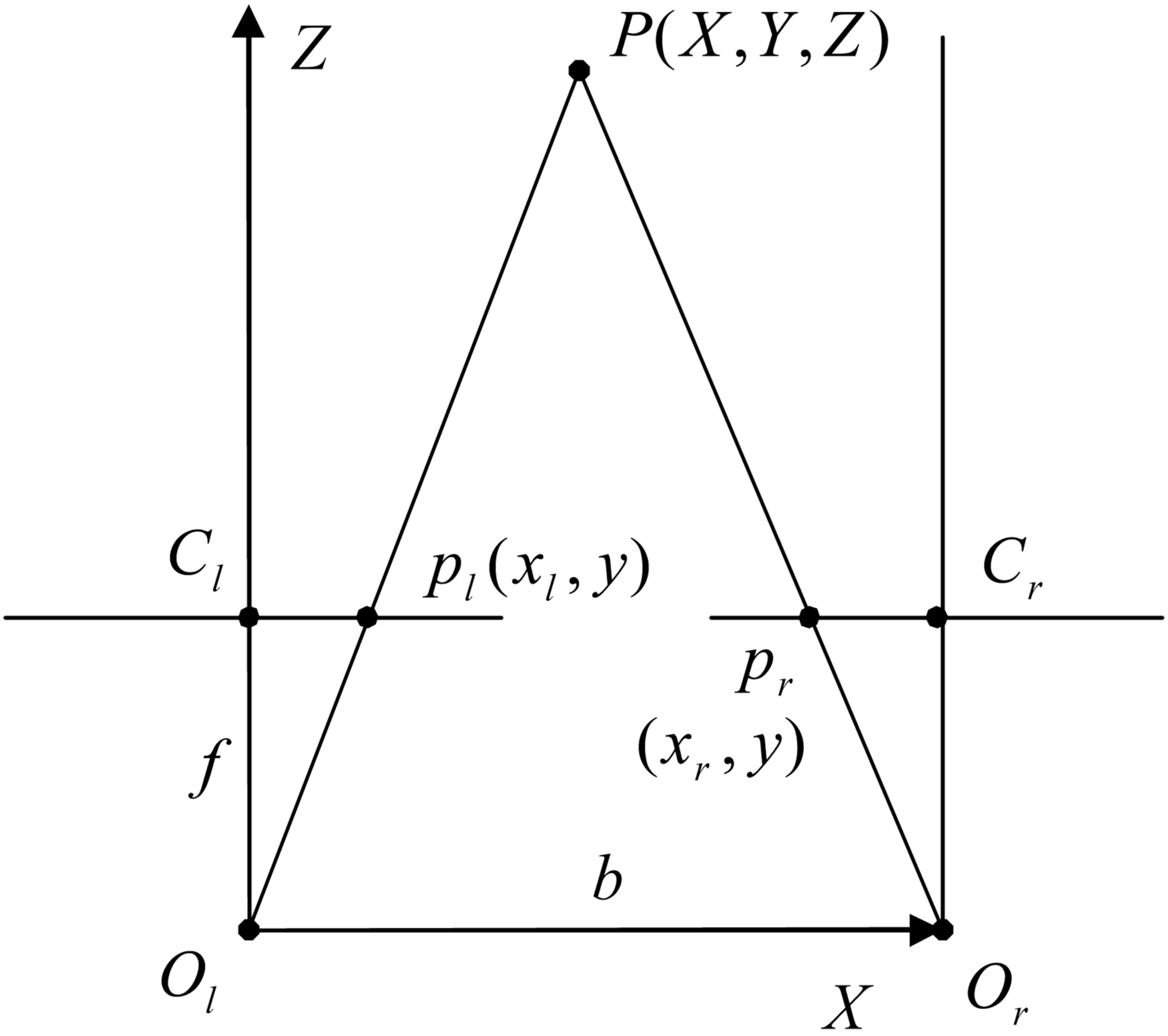

Figure 1.
Triangulation geometry in parallel-axis stereo vision.

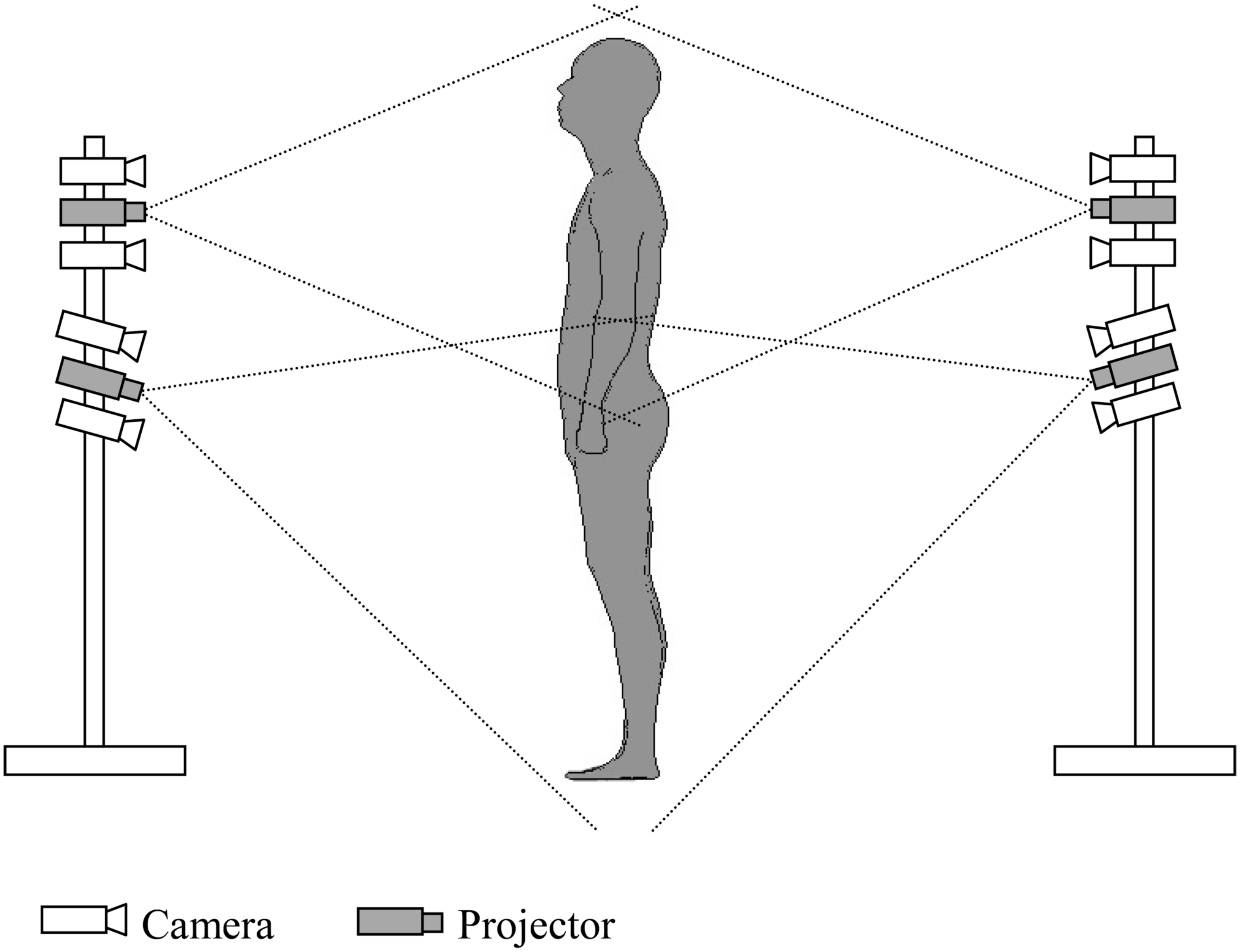

Figure 2.
Schematic of the stereo vision system.

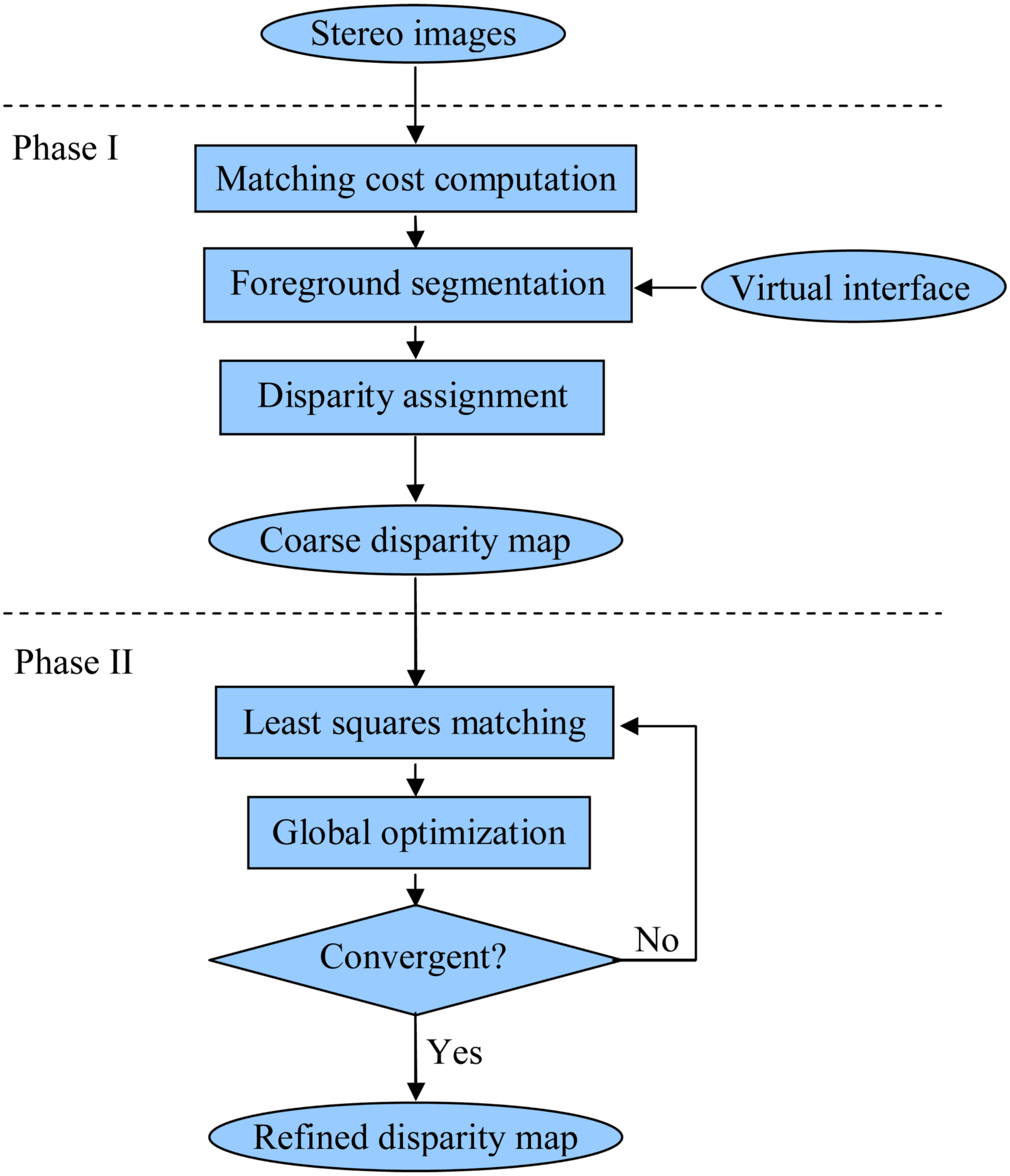

Figure 3.
Flowchart of the stereo matching algorithm.

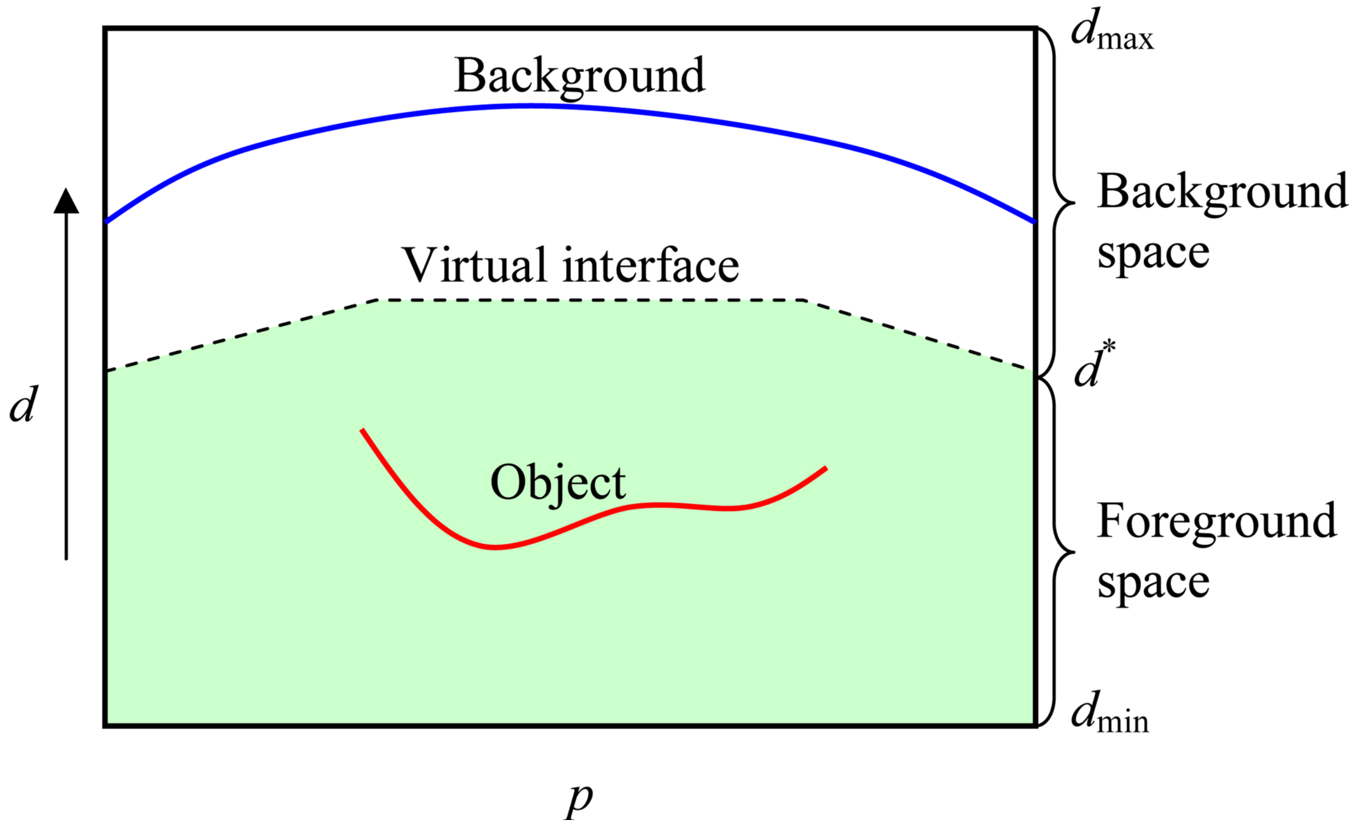

Figure 4.
Partition of the disparity space by an assumed virtual interface.

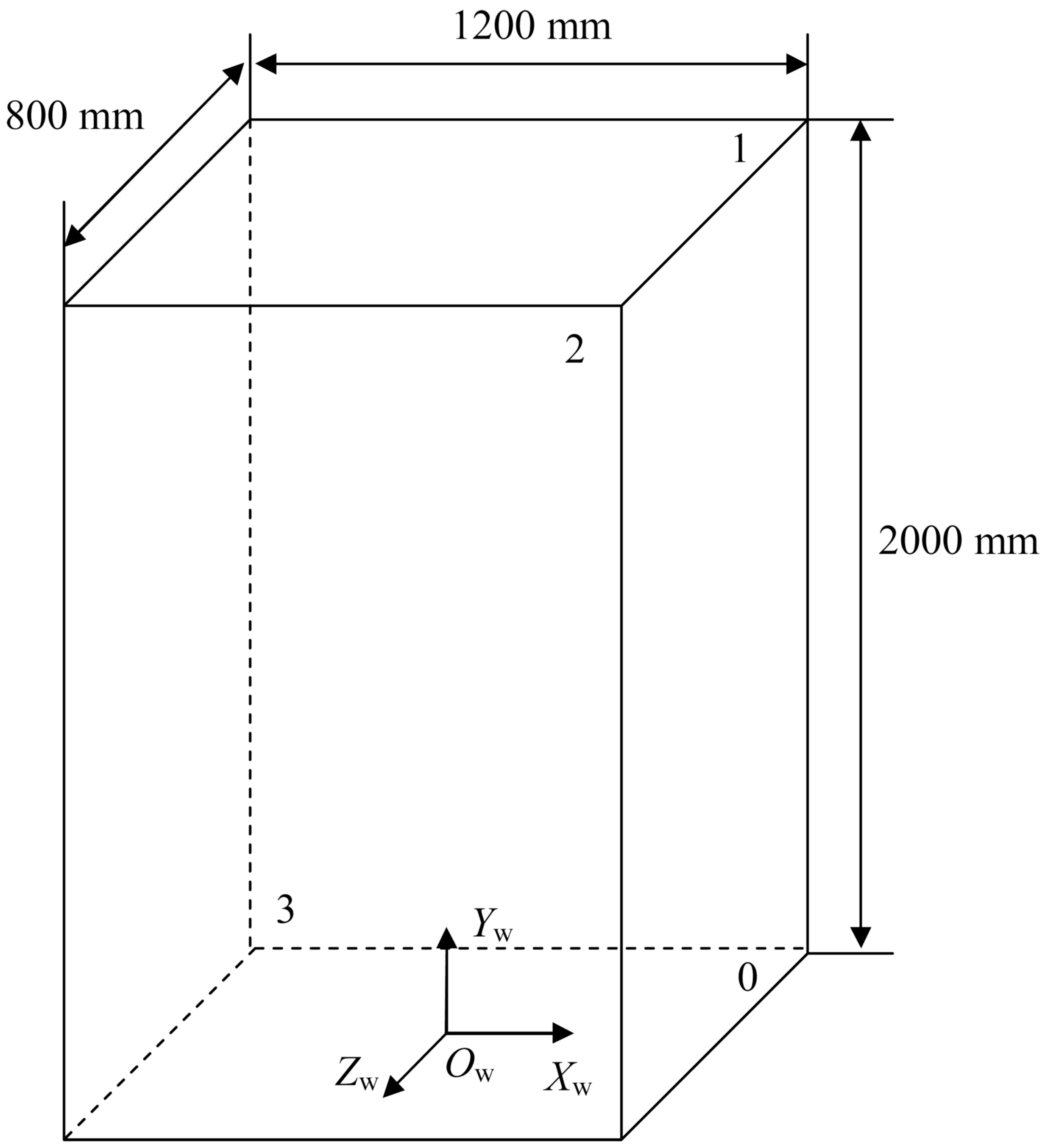

Figure 5.
The effective imaging volume of the stereo vision system.

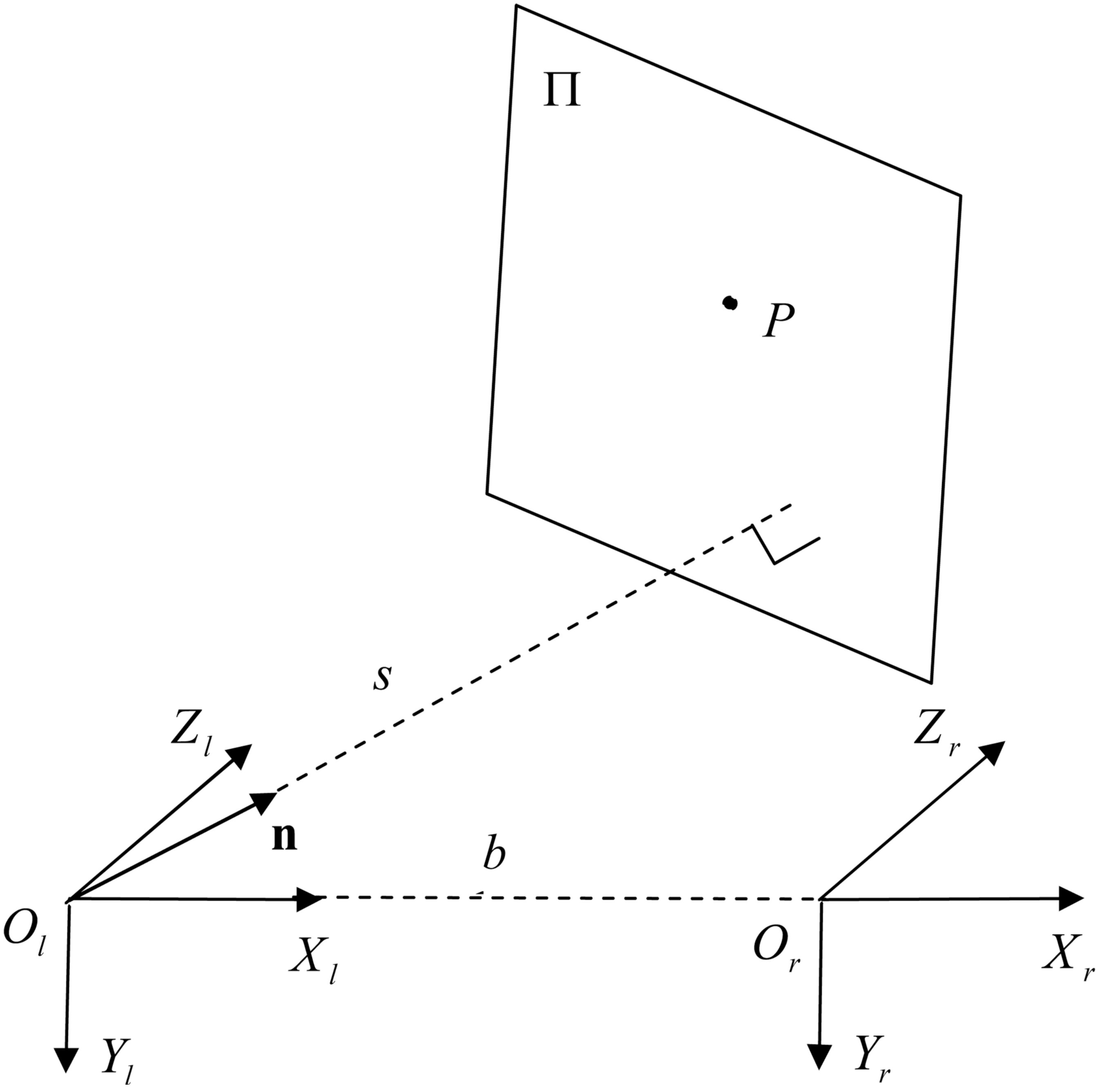

Figure 6.
A 3D plane induces a homography between the image planes in stereo vision.

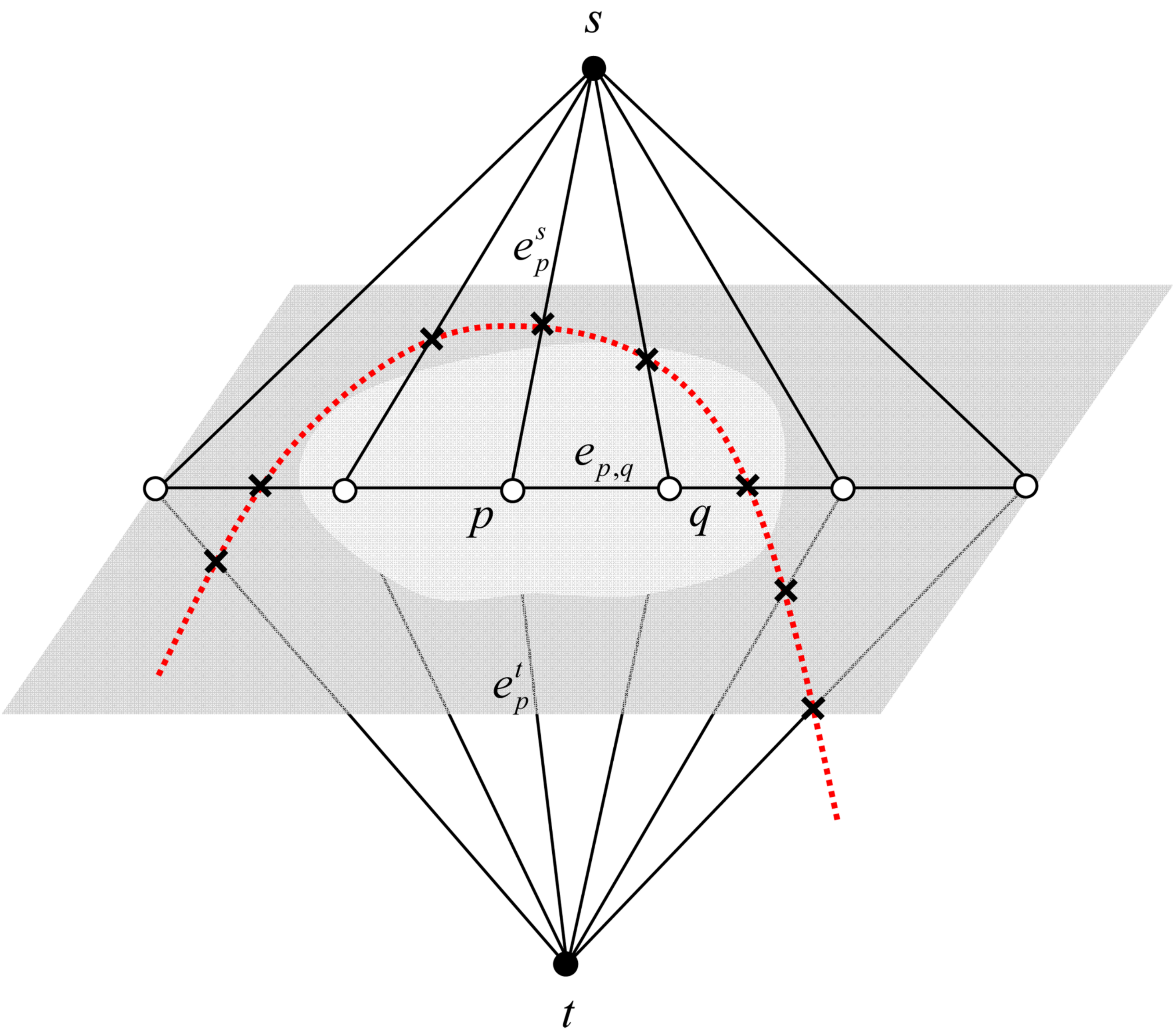

Figure 7.

Graph construction for energy minimization.

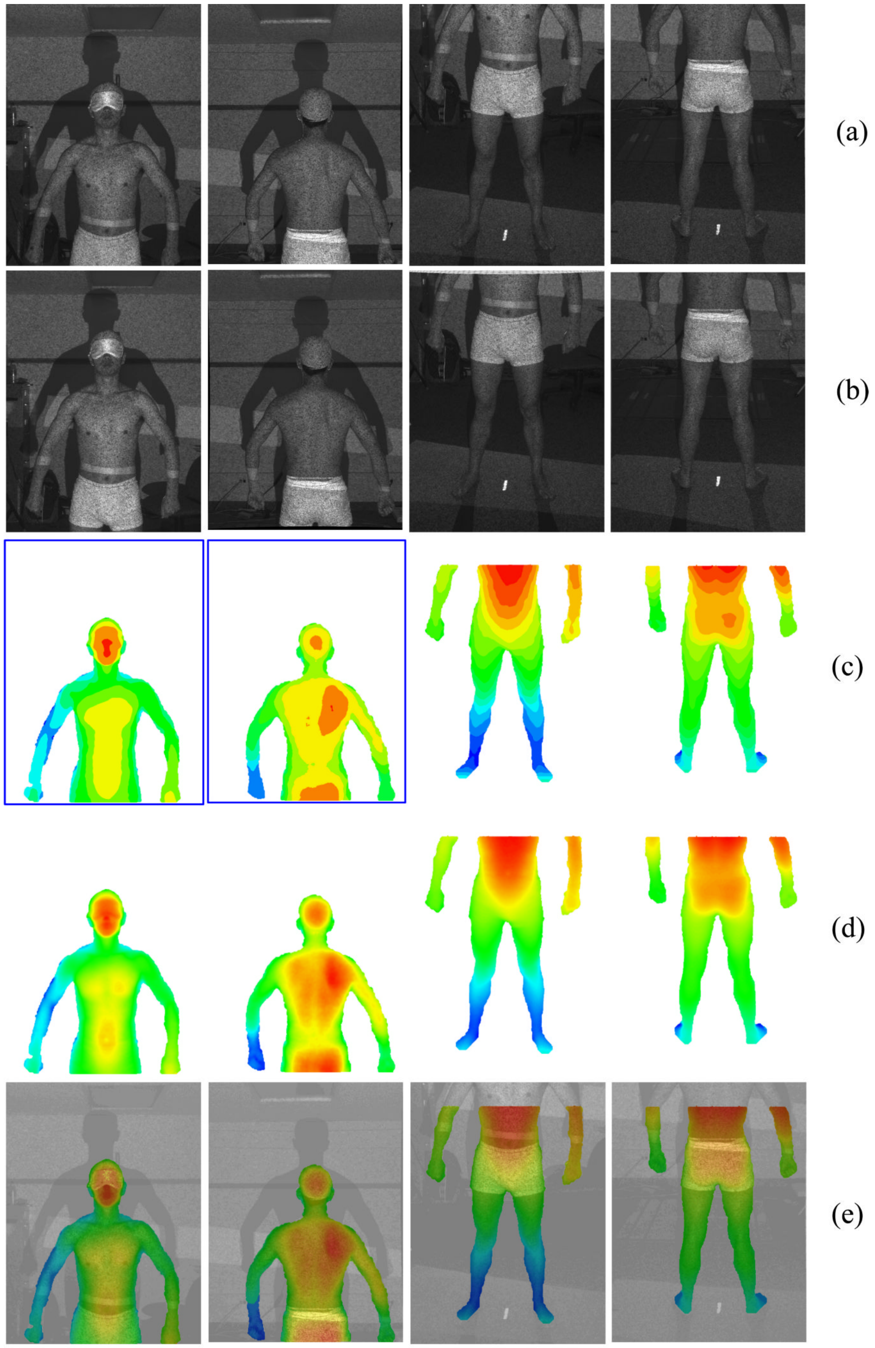

Figure 8.

Results on a human subject. For each column: (a) rectified left image; (b) rectified right image; (c) foreground segmentation and coarse disparity map; (d) refined disparity map; and (d) the refined disparity map is overlaid onto the left image. The images and disparity maps have been rotated 90° clockwise for better display.

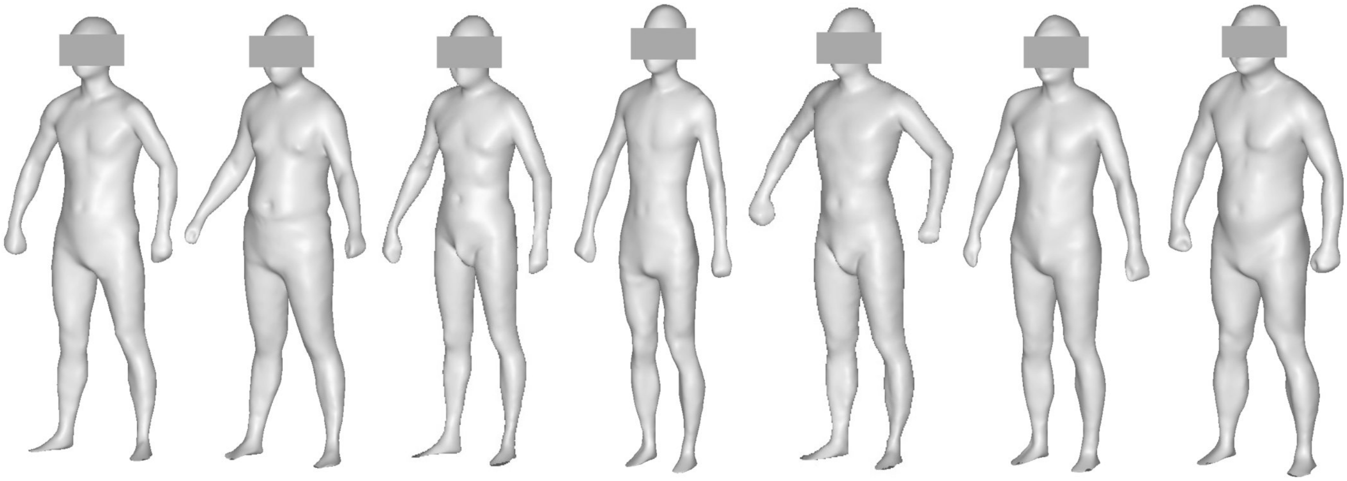

Figure 9.
Examples of body models captured by the system.

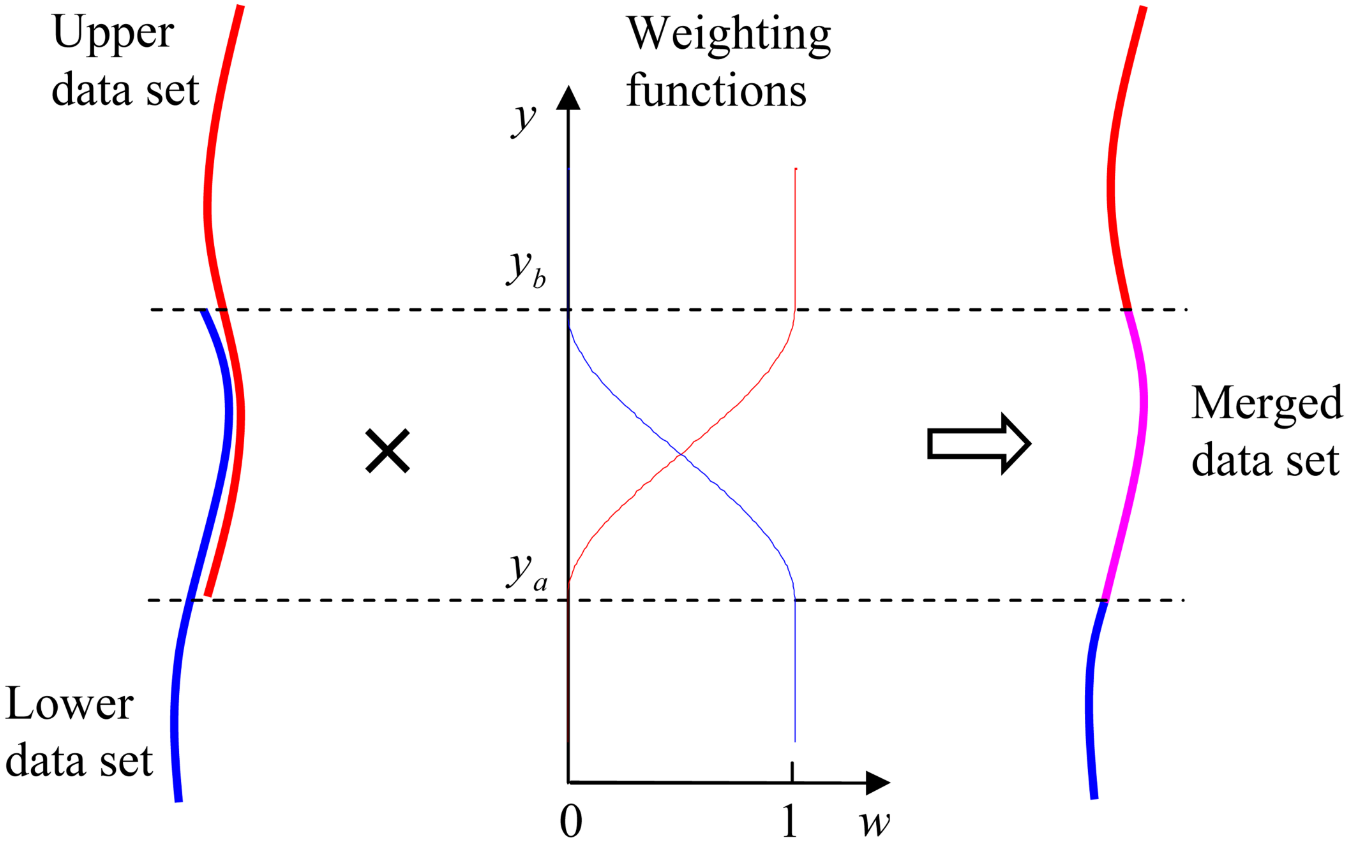

Figure 10.
Smooth merge of the upper and lower data sets in each view.

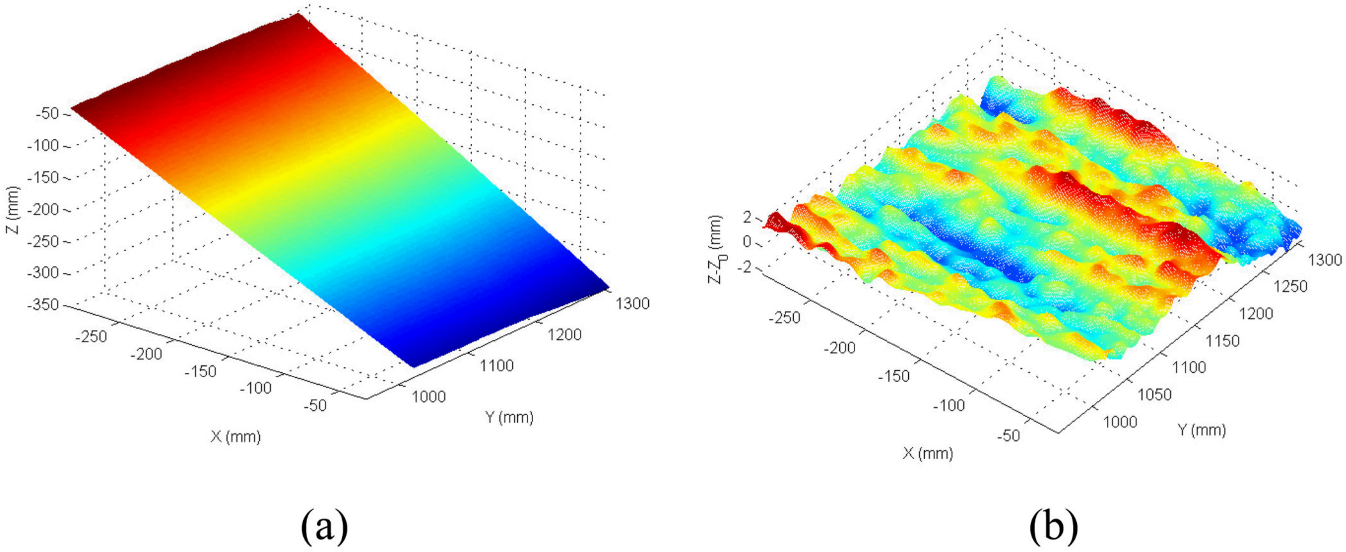

Figure 11.
Results of 3D surface imaging of a planar target. (a) Measured data; and (b) plane-fitting residual errors.

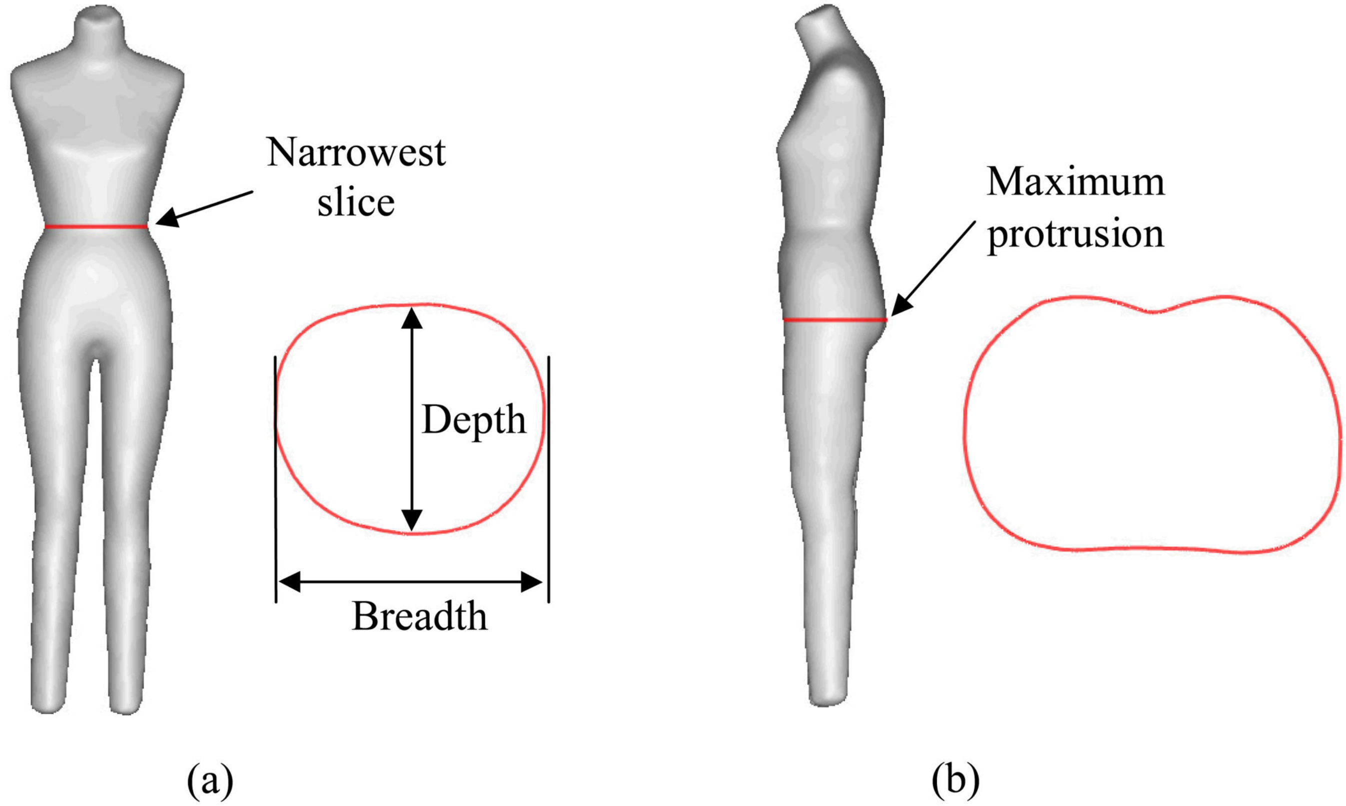

Figure 12.
Definitions of the waist and hip on a mannequin. (a) The waist and its breadth and depth; and (b) the hip.

Table 1

Parameters for the test of stereo matching.

Parameter	Description
$W_{NCC} = 5 \times 5$	Window size of NCC
$\sigma_t = 1.0$	Threshold of the variation of intensity for detecting unmatched pixels
$\rho_t = 0.6$	Threshold of NCC for detecting unmatched pixels
$C_O^F = 1.0$	Cost of assigning an unmatched pixel to the foreground
$C_O^B = 0.2$	Cost of assigning an unmatched pixel to the background
$\beta_0 = 1.0$	Parameter in the Potts model
$W_{Median} = 21 \times 21$	Window size of the median filter for reducing noise in the coarse disparity map
$r_{STE} = 3.3$	Radius of the circular structural element in the morphological close operator for smoothing the contours of foreground objects
$N_{Iter} = 15$	Number of iterations in disparity refinement
$W_{SSD} = 11 \times 11$	Window size of SSD in disparity refinement
$\lambda = 10.0$	Regularization parameter in disparity refinement

Table 2

Dimensions of the mannequin measured by manual methods and the 3D system.

	Tape	3D	Difference	*P*
Waist (mm)	704.6 ± 0.8	705.8 ± 1.3	1.2 ± 1.6	0.027
Hip (mm)	966.5 ± 1.2	958.9 ± 1.7	−6.6 ± 2.1	< 0.001
Breadth (mm)	238.9 ± 0.1	239.9 ± 1.6	1.0 ± 1.6	0.086
Depth (mm)	201.1 ± 0.0	202.2 ± 0.8	1.1 ± 0.8	0.002

Note: The *P*-values were from *t* tests.