

ARTICULATED 3-D MODELLING IN A WIDE-BASELINE DISPARITY SPACE

S. Ivekovic, E. Trucco

School of Computing,
University of Dundee,
Dundee DD1 4HN,
Scotland,

e-mail: {spelaivekovic,manueltrucco}@computing.dundee.ac.uk

Keywords: articulated 3-D modelling, dense disparity map completion, disparity space, 3-D space, novel view synthesis.

Abstract

Image-based novel-view synthesis requires dense correspondences between the original views to produce a high quality synthetic view. In a wide-baseline stereo setup, dense correspondences are difficult to achieve due to the significant change in viewpoint giving rise to a number of problems. To improve their quality, the original, incomplete disparity maps are usually interpolated to fill in the missing regions. When the data is very sparse, such as in the case of the wide-baseline stereo, interpolation alone is not enough. Instead, a 3-D model of the scene can be used to fill in the missing regions more reliably, using *a-priori* knowledge. However, the 3-D model can be used more efficiently and accurately in disparity space, where the disparity data originates from. In this paper we present and compare the two techniques. We show that, in comparison with the 3-D approach, the disparity space approach offers a computationally less expensive and potentially more accurate solution.

1 Introduction

In this paper we address the problem of dense wide-baseline disparity data and its use for novel view synthesis. Dense stereo disparity data is typically generated with a correspondence search algorithm using an image-intensity based similarity measure to determine the matching regions [23]. In order to constrain the search and reduce its complexity, additional information such as epipolar geometry and anticipated depth of the scene is used. As the main source of information available is the image intensity, in practice, a number of factors exist which make the search for image correspondences a difficult problem.

Different cameras have a different colour response, untextured regions are a source of ambiguity and the number of occlusions in the scene (regions only visible to one camera) increases with the scene complexity. In a wide-baseline stereo configuration these problems become even more pronounced. Although the current state-of-the-art stereo correspondence algorithms achieve a higher match density and accuracy

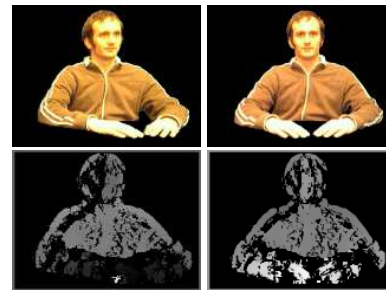


Figure 1: The left and right wide-baseline stereo view and the corresponding left-right and right-left disparity map acquired by a pyramidal search algorithm and checked for consistency. The background and the areas with missing disparity information are coloured in black.

than the previously known methods [5, 23, 24, 25], they nevertheless still fall prey to occluded areas, very common in a wide-baseline setup.

To demonstrate the effect of incomplete wide-baseline disparity data on the quality of the novel-view synthesis, we generated a synthetic data set with known ground truth, using a 3-D kitten mesh model provided courtesy of UU, INRIA, by the AIM@SHAPE Shape Repository [2]. We acquired 3 views of the kitten model under a wide baseline (Figure 2(a)) and used the left and centre view as the reference stereo pair, and the right view as the ground truth for novel view synthesis.

For the purpose of the stereo correspondence search, we textured the model with a synthetic random texture and used a simple pyramidal window-based stereo correspondence algorithm. The resulting left-centre consistency-checked disparity map is shown in Figure 2(b). We deliberately chose a simple and straightforward correlation-based correspondence search algorithm, involving no interpolation, in order to demonstrate the problems arising in the wide-baseline stereo. Figure 3(a) shows the novel view synthesised by using the original, patchy disparity map. The synthesized view is equally patchy. View-synthesis using a simple linearly interpolated disparity map is shown in Figure 3(b). As our model's shape consists of curvy patches, rather than planes, we also interpolated the disparity map with cubic splines. The corresponding view synthesis result is shown in Figure 3(c).

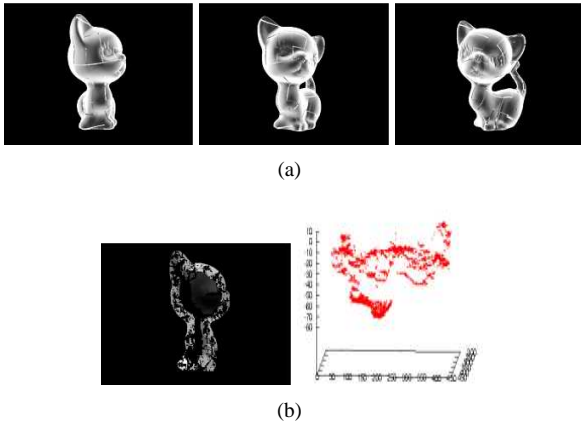


Figure 2: (a) The 3 views of the kitten model used in the analysis of different post-processing methods, (b) the disparity map corresponding to the left and centre view, shown as an image and as data points in disparity space.

It quickly becomes obvious that interpolating disparity data is not enough to generate high quality novel views. The data is too sparse and the regions missing too significant to be restored simply by interpolating the existing data. To show how much difference accurate and complete disparities make, we also generated the ground truth disparity map, using the kitten model. The corresponding view synthesis result is shown in Figure 3(d). By introducing the *a priori* knowledge about the shape of the object, we were able to synthesise a visually much more complete novel view, where the missing texture is not a consequence of missing disparity information, but instead lack of texture information in the reference views.

To gauge the accuracy of the view synthesis using each of the 4 disparity maps, we chose 500 3-D points on the kitten model and compared their image pixel coordinates when synthesised using the post-processed disparity maps with their image pixel coordinates when projected directly into the selected view. Figure 4 illustrates the experiment. The quantitative evaluation of the mean error and standard deviation in the 2-D Euclidean distance between the synthesised and projected point coordinates is shown in Table 1.

The results in Table 1 show that the model-based ground-truth disparity map clearly outperforms the other post-processing methods and also does better than the original disparity map, obtained with the correspondence search algorithm. This indicates that the quality of the wide-baseline disparity maps could be improved by fitting a generic *a-priori* model of the scene to the disparity information available from the correspondence search and using the resulting model as the new disparity map.

As we will show in the following sections, the suggested generic model does not need to be a 3-D model. The *a priori* shape can also be modelled in disparity space. By doing so, the model can be fit directly to disparity data rather than 3-D points and the disparity map for view synthesis is then readily available in the form of a deformed model after the

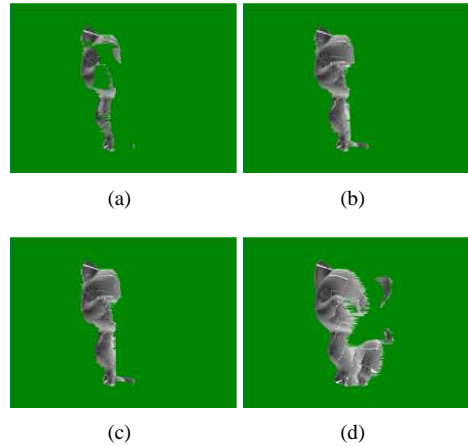


Figure 3: Synthesised view of the kitten. The viewpoint is equivalent to the right view shown in Figure 2(a). Original disparity map (a), linearly interpolated disparity map (b), cubic-spline interpolated disparity map (c), model-based disparity map (d).

| | Mean error | Standard deviation | Missing matches |
|-------------|------------|--------------------|-----------------|
| Original DM | 2.507091 | 10.431232 | 247 |
| Linear DM | 7.874941 | 15.513824 | 0 |
| Cubic DM | 7.348909 | 15.989594 | 0 |
| Model DM | 0.306740 | 0.263713 | 0 |

Table 1: Disparity map (DM) post-processing error statistics. The mean error and standard deviation were only measured for the points with known correspondences.

fit. By completing the data in disparity space, we skip a full step of 3-D reconstruction and reprojection which results in computational time savings (see the diagram in Figure 5 for illustration).

2 Related Work

Image and Video-Based Rendering addresses the problem of novel view synthesis by using image texture information to synthesise realistic novel views. When no other information about the scene is available, a large number of images acquired from fairly densely sampled viewpoints are required to produce realistic results. This approach results in elaborate image storage and retrieval requirements and is not very practical. The number of input images can be reduced by adding the information about the scene, in particular scene geometry.

The additional information can be introduced in different ways. The early work of Kanade *et al.* [16] makes extensive use of image-based stereo to produce Visible Surface Models which are then used to model the scene and render it from different viewpoints. Silhouette-based constraints are exploited by Matusik *et al.* [20], who compute a polyhedral visual hull

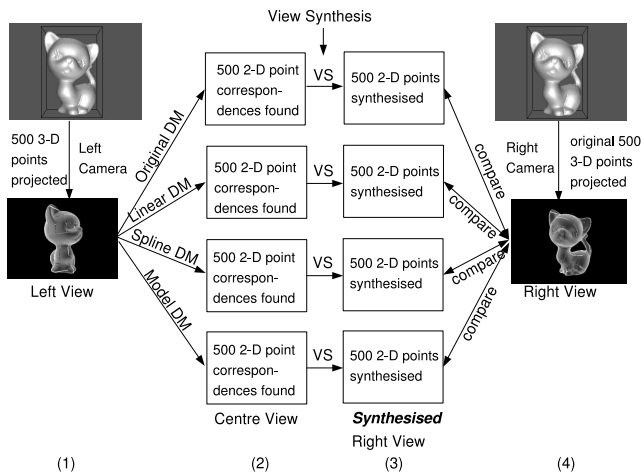


Figure 4: The view synthesis accuracy experiment. (1) 500 3-D points are projected in the left view. (2) Each of the 4 disparity maps is used to find corresponding points in the centre view. (3) The corresponding point pairs are used to synthesise the points in the right view. (4) The image location of the synthesised points is then compared to the image location of the 3-D points directly projected into the right view. The quantitative results are presented in Table 1.

and render it from novel views using view-dependent texture. Combining visual hulls and correspondence over time, Cheung *et al.* [8] compute a temporal visual hull from silhouette constraints and render it from novel viewpoints.

Approaches using *a-priori* models of the scene are used as well. Starck and Hilton [26] deform a generic computer graphics model to match the stereo, silhouette and feature constraints. The model is then textured and an animation rendered from novel views. Plänkner and Fua [22] use an implicit surface model and recover its shape and motion from stereo and silhouette data. Carranza *et al.* [6] also estimate the pose of a generic 3-D model from image silhouettes and render novel views of the textured model.

Some approaches focus on using high-quality stereo information in the form of disparity maps. They invest a lot of effort in ensuring that the disparity maps are complete and accurate. Zitnick *et al.* [30] achieve very convincing results in dynamic scene rendering by using a colour-segmentation based stereo algorithm and extracting mattes for the areas near the depth discontinuities. In 3DTV context, Kauff *et al.* [17] post-process the consistency checked disparity maps by first segmenting them based on colour clustering and change detection and then using a variety of interpolation operators. In 3D Video, Waschbüsch *et al.* [29] use a “video brick” which acquires texture and structured light stereo images simultaneously. They then use a multi-window based matching algorithm with a subsequent subpixel disparity refinement to compute accurate disparity maps.

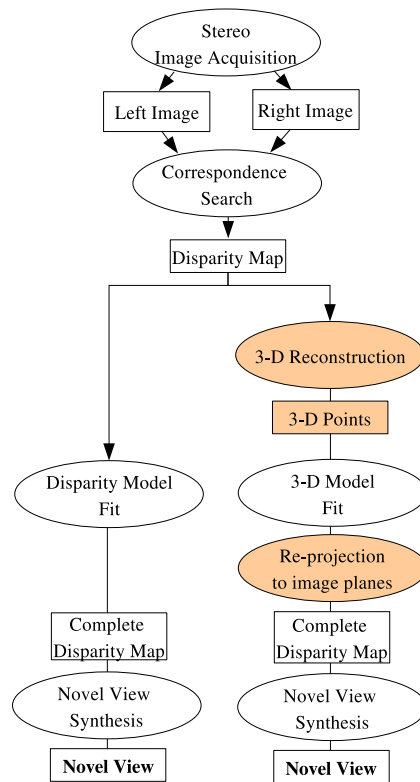


Figure 5: The diagram showing the comparison between fitting a model to disparity data and to reconstructed 3-D data. The equivalent steps in both approaches are shown on the same level. The colour-shaded parts of the diagram are the extra steps necessary to perform the fit to 3-D data.

Our approach combines the strengths of the model-based and pure stereo-based techniques. Our view synthesis algorithm is based on the trifocal tensor [3] and relies on the quality of the wide-baseline disparity map. We propose to improve the quality of the disparity map by fitting a generic *a-priori* model to the disparity data and using the deformed model to produce a complete disparity map, which is then used for novel view synthesis. Unlike other model-based approaches, we do not texture the resulting model but instead use it to produce a high-quality disparity map. View synthesis is then performed using the original image texture and the post-processed disparity map. As we are post-processing disparity data and not 3-D points, we also propose to implement this approach in disparity space rather than in 3-D. By doing so, we avoid the step of 3-D reconstruction. Our model-fitting approach results in a disparity space model which can itself be interpreted as a 3-D disparity map and used directly to synthesise novel views.

Disparity space related research can be roughly split in two subgroups - the *3-D disparity space* methods and the *disparity space image* methods. 3-D Disparity Space (x, y, d) is a more general concept of the two, a three-dimensional projective space defined as a projective transform of the 3-D space, where x and y are the column and row dimensions of the image plane, respectively, and d is the disparity of a left image point with respect to the corresponding right image point. *Disparity*

space image (DSI) is an image or a function defined over a continuous or discretised version of the disparity space, explicitly representing the matching space, and is constructed to facilitate the stereo matching process, like, for example, in [4, 23, 7].

In the area of dense disparity matching, Szeliski and Golland [27] formulate the problem of simultaneously recovering the disparities, true colours and opacities in a generalised 3-D (x,y,d) disparity space and solve it using iterative energy minimisation. Hong and Chen [13] describe a segment-based stereo matching algorithm which fits disparity planes to segments of disparity data. Mordohai and Medioni [21] perform tensor voting on the $\{x, y, d\}$ data to recover matches belonging to coherent disparity surfaces. Thakoor *et al.* [28] describe DSP implementation of an iterative segmentation-estimation framework for plane segmentation in disparity space.

In the area of motion estimation from stereo, Demirdjian *et al.* [9] describe rigid 3-D motion estimation in disparity space. Derpanis and Chang [10] report an extension with a closed form linear solution for rigid motion estimation. Agrawal and Konolige [1] present a mobile robot localisation system for outdoor environments which estimates the motion in disparity space and Hattori and Takeda [12] report dense stereo matching in disparity space used in an implementation of a side collision system for a road vehicle.

Our work differs from the mentioned research in that it addresses an articulated motion and articulated structure modelling in disparity space, to our knowledge not attempted so far. In particular, we show that the idea of rigid motion estimation in disparity space [9, 10] can be extended to estimating and modelling articulated motion and structure in disparity space, which can in turn be used to complete wide-baseline disparity data for high-quality novel view synthesis.

3 Disparity Space

Let us assume that a 3-D point $\mathbf{M} = (X, Y, Z, 1)^T$ is viewed by two distinct cameras, left camera with projection matrix P_l and right camera with projection matrix P_r , and that the image of the point \mathbf{M} is defined as $\mathbf{m}_l = (x_l, y_l, 1)^T \simeq P_l \mathbf{M}$ and $\mathbf{m}_r = (x_r, y_r, 1)^T \simeq P_r \mathbf{M}$ in the left and right camera's image plane, respectively, where “ \simeq ” denotes equality up to a scale factor. The corresponding points \mathbf{m}_l and \mathbf{m}_r are related by a *disparity* which, in a general case, is defined as:

$$d(\mathbf{m}_l, \mathbf{m}_r) = \mathbf{m}_l - \mathbf{m}_r = (x_l - x_r, y_l - y_r). \quad (1)$$

In the case of rectified images, the two corresponding points lie on the same scanline, and the disparity simplifies to a displacement along the scanline:

$$d(\mathbf{m}_l, \mathbf{m}_r) = x_l - x_r \quad (2)$$

For a rectified stereo pair of images, the disparity space is then defined as a three-dimensional space $\mathbb{D}^3 = \{x, y, d\}$. The so

defined disparity space is a projective space. This can be shown by deriving a projective transformation P_D between the 3-D space \mathbb{R}^3 and disparity space \mathbb{D}^3 , $P_D : \mathbb{R}^3 \rightarrow \mathbb{D}^3$, as follows.

We assume, without a loss of generality, a specific form of the rectified projection matrices [11]. Let the left and right rectified camera projection matrices, \tilde{P}_l and \tilde{P}_r , be written as:

$$\tilde{P}_l = \begin{pmatrix} p_{11}^l & p_{12}^l & p_{13}^l & p_{14}^l \\ p_{21}^l & p_{22}^l & p_{23}^l & p_{24}^l \\ p_{31}^l & p_{32}^l & p_{33}^l & p_{34}^l \end{pmatrix} \quad (3)$$

$$\tilde{P}_r = \begin{pmatrix} p_{11}^r & p_{12}^r & p_{13}^r & p_{14}^r \\ p_{21}^r & p_{22}^r & p_{23}^r & p_{24}^r \\ p_{31}^r & p_{32}^r & p_{33}^r & p_{34}^r \end{pmatrix}. \quad (4)$$

Projecting a point $\mathbf{M} = (X, Y, Z, 1)^T \in \mathbb{R}^3$ into left and right view gives the left and right image point, \mathbf{m}_l and \mathbf{m}_r :

$$\mathbf{m}_l = \begin{pmatrix} x_l \\ y_l \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{p_{11}^l X + p_{12}^l Y + p_{13}^l Z + p_{14}^l}{p_{31}^l X + p_{32}^l Y + p_{33}^l Z + p_{34}^l} \\ \frac{p_{21}^l X + p_{22}^l Y + p_{23}^l Z + p_{24}^l}{p_{31}^l X + p_{32}^l Y + p_{33}^l Z + p_{34}^l} \\ 1 \end{pmatrix} \simeq \tilde{P}_l \mathbf{M} \quad (5)$$

$$\mathbf{m}_r = \begin{pmatrix} x_r \\ y_r \\ 1 \end{pmatrix} = \begin{pmatrix} \frac{p_{11}^r X + p_{12}^r Y + p_{13}^r Z + p_{14}^r}{p_{31}^r X + p_{32}^r Y + p_{33}^r Z + p_{34}^r} \\ \frac{p_{21}^r X + p_{22}^r Y + p_{23}^r Z + p_{24}^r}{p_{31}^r X + p_{32}^r Y + p_{33}^r Z + p_{34}^r} \\ 1 \end{pmatrix} \simeq \tilde{P}_r \mathbf{M} \quad (6)$$

A point $\mathbf{D} \in \mathbb{D}^3$, written in homogeneous coordinates, is defined as:

$$\mathbf{D} = \begin{pmatrix} x_l \\ y_l \\ x_l - x_r \\ 1 \end{pmatrix} = \quad (7)$$

$$\begin{pmatrix} \frac{p_{11}^l X + p_{12}^l Y + p_{13}^l Z + p_{14}^l}{p_{31}^l X + p_{32}^l Y + p_{33}^l Z + p_{34}^l} \\ \frac{p_{21}^l X + p_{22}^l Y + p_{23}^l Z + p_{24}^l}{p_{31}^l X + p_{32}^l Y + p_{33}^l Z + p_{34}^l} \\ \frac{p_{11}^r X + p_{12}^r Y + p_{13}^r Z + p_{14}^r}{p_{31}^r X + p_{32}^r Y + p_{33}^r Z + p_{34}^r} - \frac{p_{11}^l X + p_{12}^l Y + p_{13}^l Z + p_{14}^l}{p_{31}^l X + p_{32}^l Y + p_{33}^l Z + p_{34}^l} \\ 1 \end{pmatrix}, \quad (8)$$

and the transformation P_D , for which

$$\mathbf{D} \simeq P_D \mathbf{M}, \quad \mathbf{M} \in \mathbb{R}^3, \mathbf{D} \in \mathbb{D}^3 \quad (9)$$

as

$$P_D = \begin{pmatrix} p_{11}^l & p_{12}^l & p_{13}^l & p_{14}^l \\ p_{21}^l & p_{22}^l & p_{23}^l & p_{24}^l \\ p_{11}^r - p_{11}^l & p_{12}^r - p_{12}^l & p_{13}^r - p_{13}^l & p_{14}^r - p_{14}^l \\ p_{31}^l & p_{32}^l & p_{33}^l & p_{34}^l \end{pmatrix} \quad (10)$$

The transformation P_D is the link between the 3-D space and disparity space which allows us to model the structure in 3-D space, where we can intuitively represent the geometry, and then convert the model into a disparity space representation for further manipulation.

It can be assumed that the noise associated with $\{x, y, d\}$ is homoscedastic and fairly approximated by a covariance matrix

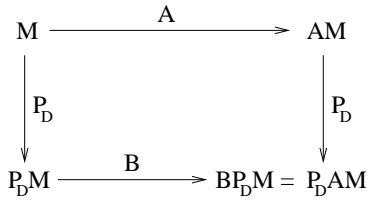


Figure 6: Transformation diagram

$\Lambda = \sigma^2 \mathbf{I}$, where $\sigma = \sigma_x = \sigma_y = \sigma_d$ and σ is typically 1 image pixel [9]. Unlike in the case of 3-D reconstructed points, where their heteroscedastic nature requires a special error-in-variables approach when estimating model parameters from data [19], the disparity data is statistically better behaved and can as such be used for more accurate parameter estimation [1, 9, 10].

4 Estimating and Modelling Articulated Motion in Disparity Space

In this section we describe how the rigid motion estimation in disparity space can be extended to an articulated motion. Let $M \in \mathbb{R}^3$, P_D be as in Equation (10), and $A = [R|t]$ be a known homogeneous transformation in \mathbb{R}^3 , transforming M into AM . Homogeneous transformation B , which transforms $D \in \mathbb{D}^3$ to $BD \in \mathbb{D}^3$, can then be derived as follows (see the transformation diagram in Figure 6):

$$\begin{aligned}
BP_D M &= P_D AM \\
BP_D &= P_D A \\
B &= P_D A P_D^{-1}
\end{aligned} \tag{11}$$

This result shows that the points can be rotated and translated in disparity space directly if, for every known homogeneous transformation A in 3-D space, a corresponding transformation B is computed as in Equation (11).

4.1 Articulated Body Pose Estimation in 3-D Space

The body pose of an articulated body model is estimated by identifying the homogeneous transformations defining the skeleton kinematic chain. The skeleton is defined as a set of transformation matrices which encode the position and orientation of every joint with respect to its parent joint in the hierarchy:

$$Skeleton = \{A_1^2, A_2^3, \dots, A_{N-1}^N\}, \tag{12}$$

where N is the number of joints in the skeleton and A_i^j is a homogeneous transformation matrix encoding the orientation of the coordinate system of joint j with respect to the coordinate system of joint i .

We estimate the body pose of the person imaged in the reference stereo pair by silhouette-constrained global optimisation, where the model generating the silhouette is a subdivision surface upper body model (see Figure 7(a)). The pose parameters are estimated using a Particle Swarm Optimisation (PSO) method (see Section 4.3).

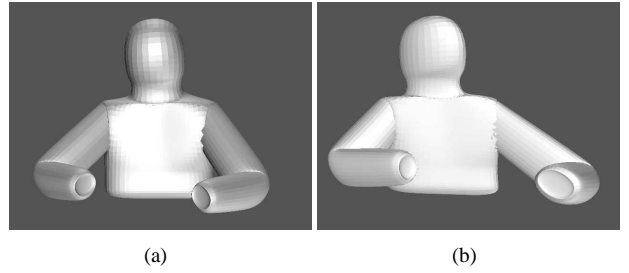


Figure 7: Generic subdivision surface upper body model: (a) 3D model; (b) Disparity space model.

4.2 Articulated Body Pose Estimation in Disparity Space

The body pose can also be estimated in disparity space by using a disparity space body model, defined as follows.

The skeleton of the disparity space model is defined as a set of disparity space transformation matrices which encode the position and orientation of every joint with respect to its parent joint in the hierarchy:

$$Skeleton = \{B_1^2, B_2^3, \dots, B_{N-1}^N\}, \tag{13}$$

where N is the number of joints in the skeleton and B_i^j is a homogeneous disparity space transformation matrix, encoding the orientation of the coordinate system of joint j with respect to the coordinate system of joint i , and defined as in Equation (11). The model’s skin, the subdivision surface mesh, is also transferred to the disparity space using Equation (9) on the mesh vertices.

The pose is then estimated in disparity space directly, with silhouettes generated by the disparity space model (see Figure 7(b)) and Particle Swarm Optimisation, just like in the 3-D case. The “transfer” of the model from 3-D space to disparity space is only necessary to define the disparity space model, and only needs to be done when specifying the model to use. If we were able to model structure in disparity space directly, this step would not be necessary.

4.3 Pose Estimation with PSO

Particle Swarm Optimisation (PSO) is an evolutionary optimisation technique developed by Kennedy and Eberhart [18]. It models the swarming behaviour exhibited by bird flocks and fish schools.

Every particle in the swarm represents one possible solution to the optimisation problem. When estimating the body pose, each particle represents a possible skeleton configuration:

$$X_i = (r_x, r_y, r_z, \alpha_x^0, \beta_y^0, \gamma_z^0, \alpha_x^1, \beta_y^1, \gamma_z^1, \dots, \alpha_x^N, \beta_y^N, \gamma_z^N), \tag{14}$$

where N is the number of joints, r_x, r_y, r_z denote the position of the root joint with respect to the world coordinate system, α_x^0 denotes a rotation around the x -axis of the root joint coordinate system for angle α , γ_z^1 denotes a rotation around the z -axis of the next joint in the hierarchy for angle γ , etc.

Each particle explores the search space on its own and in interaction with other particles. The solution is found when all particles converge. A more detailed description of the body pose estimation with PSO can be found in [14].

5 Fitting the Model to Disparity Data

5.1 3D Space

Post-processing disparity data in 3-D space first requires the correspondence pairs to be reconstructed as 3-D points via triangulation. The model is then deformed to fit the reconstructed points using the quasi-interpolation method (see Section 5.3), and the resulting model projected onto the image planes of the reference stereo cameras to obtain the post-processed disparity map. This process is illustrated in the right part of the diagram in Figure 5.

5.2 Disparity Space

In disparity space, the disparity model can be fit to the disparity data directly, using the same fitting method. Once deformed, the model can then be used as a three-dimensional disparity map itself, or transformed into a conventional 2-D disparity map by orthographic projection onto the left camera’s image plane. Fitting to disparity data directly is illustrated in the left part of the diagram in Figure 5.

5.3 Fitting Method

The subdivision surface model is fit to the data in an iterative manner, using quasi-interpolation. Quasi-interpolation is a technique which finds the subdivision control polygon (base mesh) for which the resulting subdivision limit surface best approximates the given data. In 1-D, it is applied locally as follows:

$$p_i = -\frac{1}{6}f(i-1) + \frac{4}{3}f(i) - \frac{1}{6}f(i+1), \quad \forall i \in \mathbb{Z}, \quad (15)$$

where p_i is the control polygon point, $f(i)$ is a function sample (data point), and $[-\frac{1}{6}, \frac{4}{3}, -\frac{1}{6}]$ is a 1-D quasi-interpolation stencil. The stencil is generalised to 2-D via the tensor product. A detailed description of the technique is outside the scope of this paper and can be found in [15], where further details and results are presented.

6 Experiments

We performed experiments on synthetic data to compare the performance of both proposed approaches. Results on real images are presented in Section 6.2.

6.1 Comparison of Post-Processing in 3-D and Disparity Space

We performed an experiment on a synthetic stereo pair of images to compare the performance of the post-processing algorithm in 3-D space with the one in disparity space. Every care was taken to ensure that the conditions were kept the same for both approaches and a fair comparison was possible.

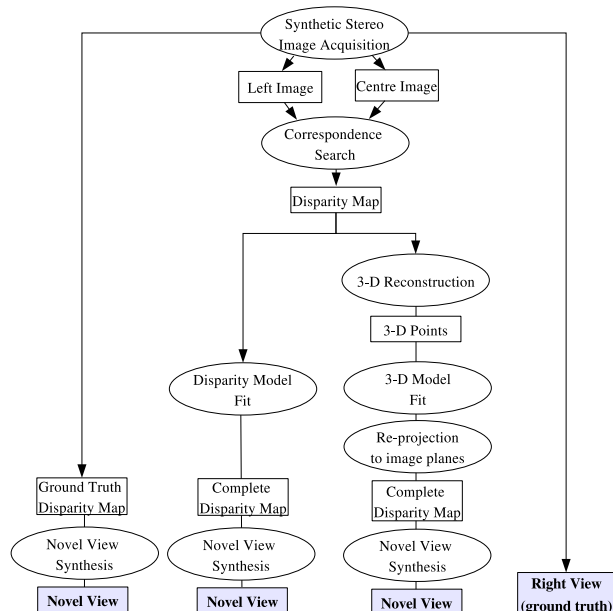


Figure 8: The flowchart of the experiment where we compared the post-processing in 3-D space with post-processing in disparity space. The shaded boxes are the views which were used for comparison.

The model-data correspondence information was determined for the 3-D model and then used for both 3-D and disparity space deformation to guarantee consistency. Figure 8 shows the diagram of the experiment.

As shown in Figure 8, three views were acquired with an OpenGL synthetic camera setup (see Figure 9). The left and centre view were used to obtain disparity data and the right view was used as ground truth for the view synthesis.

Figure 10 illustrates the comparison process. The left column refers to the disparity space algorithm and the right column to the 3-D space algorithm. Row (a) shows the *a-priori* disparity model and 3-D model. Row (b) shows the disparity data as three-dimensional data in disparity space and as reconstructed 3-D points using triangulation. The 3-D and disparity model after fitting to corresponding data are shown in row (c). Row (d) shows the resulting disparity maps and row (e) the synthesised novel view. Row (f) contains the difference images between the synthesised views and the ground truth right view.

Table 2 lists the sum of absolute differences (SAD) measure for each of the synthesised views compared to the ground truth view. The ground truth disparity map is the closest in terms of view synthesis quality, which is to be expected because the disparity map generation via the correspondence search necessarily introduces some noise in the data. Of the remaining two, the disparity space algorithm slightly outperforms the 3-D space approach. This difference is most likely due to the fact that an unavoidable approximation error was introduced when reconstructing the 3-D data in order to perform the fit in 3-D space, and then projecting the resulting model back onto the

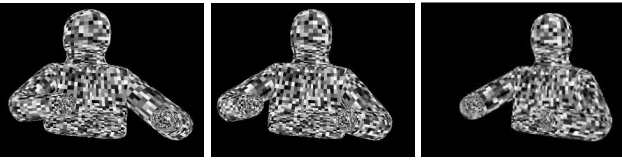


Figure 9: The three synthetic views generated for the comparison experiment.

| | Ground Truth View (Right View) |
|-----------------|--------------------------------|
| Ground Truth DM | 2659991 |
| Disparity Space | 3915826 |
| 3-D Space | 3924308 |

Table 2: Pixel-by-pixel sum of absolute differences (SAD) comparison between the ground truth view (the right view) and the synthesised novel views. The synthesised views were computed using the ground truth disparity map, the disparity space post-processed disparity map and the 3-D space post-processed disparity map.

image plane to compute the disparity map.

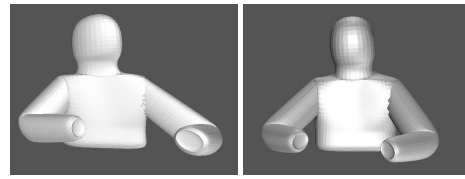
As demonstrated, disparity space algorithm not only saves on computational complexity, as illustrated in Figure 5, but also has a performance which is at least equivalent to the 3-D approach. By the latter we mean that, on top of avoiding the 3-D reconstruction error, there is a potential for the disparity space approach to be more accurate than the 3-D space approach, if care is taken to preserve the data’s homoscedastic nature [19]. We do not take advantage of this in our algorithm because we do not estimate the pose parameters directly from disparity data as it proves too sparse to serve as a reliable constraint (we use the silhouettes instead). However, should the data be available and used in this way, the pose parameter estimation in disparity space would be preferable.

6.2 Experiments with Real Data

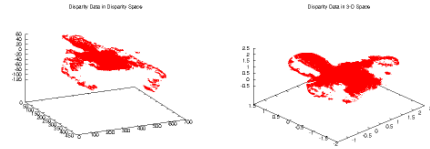
Our camera setup consists of 4 IEEE 1394 webcams acquiring in 640×480 RGB mode, three positioned in front of the person and one above. The baseline between each pair of cameras located in front of the person is approximately 20° . The camera above proves crucial in pose estimation as it resolves a lot of ambiguity, but it is not used to gather any stereo data because the overlap in the field of view is insignificant. Additional cameras are used as ground truth test cameras, which we use to gauge the quality of the view synthesis.

In this paper, we present the results of the disparity space approach on real data. A systematic quantitative and qualitative comparison of both algorithms on real data to complement the synthetic data experiment presented in this paper is the scope of future work.

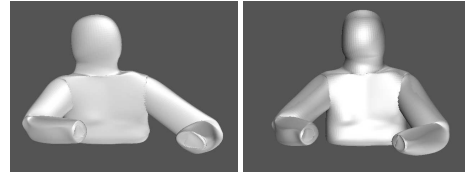
The disparity space approach was tested on a set of real



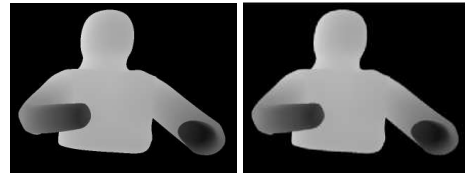
(a) Disparity model and 3-D model



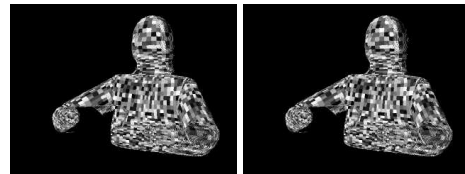
(b) Disparity data and 3-D data



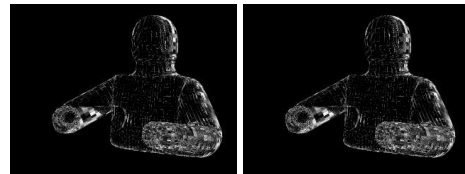
(c) Deformed disparity model and deformed 3-D model



(d) Disparity map, generated from the disparity model and from the 3-D model



(e) Novel view synthesised with the disparity map post-processed in disparity space and 3-D space



(f) Difference image for disparity space and 3-D space novel view

Figure 10: Illustration of the comparison steps performed in the experiment with synthetic data. Visually, the performance of the two compared approaches is almost exactly the same. Only a quantitative analysis reveals that there is a slight difference in quality.

images showing a person seated at a table, typical of videoconferencing scenarios. In order to complete the data with the *a-priori* subdivision model of the upper body, the correct pose of the model had to be determined first. This was achieved by optimising for skeleton joint transformations

using a Particle Swarm Optimisation algorithm, constrained by the silhouettes extracted from the 4 camera images.

At this stage, our upper body model did not include hands as they are an articulated structure in themselves, requiring a separate kinematic model, which was outside the scope of this work.

Figure 11(a,b) shows the results of view synthesis for two different poses, obtained using the disparity-space post-processed disparity map. The first row shows the original stereo pair, the second and third row show the comparison between the view synthesis using the original, patchy disparity map and the disparity map completed with our algorithm. Two different virtual camera views are shown, the first located in between the original stereo pair of cameras, below the baseline and the second located to the right of the stereo pair and above the baseline.

In Figure 12, we tested the view synthesis for a sequence of views extrapolated to the right of the reference stereo pair. The results are convincing, although extrapolation does reveal the lack of realism in the simple *a-priori* model that we are using.

Figure 13 compares the view synthesis performance when the disparity map is completed using various interpolation techniques, such as linear interpolation, bilinear interpolation and cubic spline interpolation, with the performance of our model-based approach. The fact that we do not model hands becomes a clear disadvantage in this case, as the hands move as a part of the body in the synthetic view. This is a disadvantage of a model-based approach in comparison with simple interpolation. Ignoring the hands, the model-based synthesised view seems much more consistent and less noisy than the views obtained with interpolated disparity maps, especially in the area occluded by the arms.

7 Conclusions and Future Work

We presented an approach to modelling articulated structure and motion in disparity space as an alternative to the traditional 3-D space modelling.

The aim was to use the model to post-process the disparity data used for view synthesis. In achieving that, the disparity space approach is computationally more efficient and also potentially more accurate than its 3-D alternative, as we have shown with experiments on synthetic data.

The approach suffers from the same drawbacks as other model-based approaches, primarily concerning the realism of the model and possible generalisation. The results could be improved further by using a more realistic and detailed model. Hand modelling is also a big challenge in an application like ours and requires further research. Using our approach on images of different people requires the model to be initialised to their body proportions. We currently initialise the model dimensions manually, however, attempts have been made by researchers to customise the *a-priori* model's shape to that of a specific person automatically [6, 26].

The model in the correct body pose estimated from the silhouettes can be used as a search interval constraint in the stereo correspondence algorithm. This should produce a much denser original disparity map and would present an advantage as the model's fidelity to the data after the fit could also be greater. At the moment, the level of model's skin deformation is restricted to a small number of iterations because the data is known to be noisy and very patchy. Future work will address the use of the model as a stereo constraint.

Although the disparity space modelling approach is based on the concept of "modelling in 3 dimensions", it is somewhat closer to the image-based rendering concept than its 3-D counterpart, in the sense that the step of 3-D reconstruction from corresponding pairs of points is avoided. The approach is applicable to any problem which currently uses a 3-D model for the purpose of view synthesis.

The disparity space algorithm is not restricted to only two views. It was presented on a stereo pair for the reasons of clarity and can be generalised to as many views as required. Just like additional 3-D points can be generated by adding new views and finding correspondences, those correspondences can be added to the original disparity space via the trilinear tensor transfer, without the need for 3-D reconstruction.

A potential application for our method is the immersive videoconferencing which requires efficient algorithms running in real-time. We must emphasise that the described approach has only been implemented as a proof of concept and further research is necessary to make it suitable for a real-time environment. The aim of our work was to show that, should there be a time-constrained application running in near-real time and making use of the presented concepts in 3-D space, substituting 3-D with the proposed disparity space approach is likely to further improve its performance.

Acknowledgments

The authors would like to thank the anonymous reviewers for their considered and helpful comments.

References

- [1] M. Agrawal and K. Konolige. Real-time localization in outdoor environments using stereo vision and inexpensive gps. In *International Conference on Pattern Recognition*, pages 1063 – 1068, 2006.
- [2] A.I.M.A.T.S.H.A.P.E. <http://www.aim-at-shape.net>.
- [3] S. Avidan and A. Shashua. Novel view synthesis by cascading trilinear tensors. *IEEE Transactions on Visualization and Computer Graphics*, 4(4):293–306, 1998.
- [4] A.F. Bobick and S.S. Intille. Large occlusion stereo. *International Journal of Computer Vision*, 33(3):181–200, 1999.

- [5] M.Z. Brown, D. Burschka, and G.D. Hager. Advances in computational stereo. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(8):993 – 1008, 2003.
- [6] J. Carranza, C. Theobalt, M.A. Magnor, and H.P. Seidel. Free-viewpoint video of human actors. In *Proceedings of ACM SIGGRAPH*, pages 569 – 577, 2003.
- [7] J. Cech and R. Sara. Efficient sampling of disparity space for fast and accurate matching. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, pages 1 – 8, 2007.
- [8] G.K.M. Cheung, S. Baker, and T. Kanade. Visual hull alignment and refinement across time: A 3d reconstruction algorithm combining shape-from-silhouette with stereo. In *Proceedings of IEEE Computer Vision and Pattern Recognition, Volume II*, pages 375 – 382, 2003.
- [9] D. Demirdjian and T. Darrell. Using multiple-hypothesis disparity maps and image velocity for 3-d motion estimation. *International Journal of Computer Vision*, 47(1/2/3):219–228, 2002.
- [10] K.G. Derpanis and P. Chang. Closed-form linear solution to motion estimation in disparity space. In *Intelligent Vehicles Symposium*, pages 268 – 275, 2006.
- [11] A. Fusiello, E. Trucco, and A. Verri. A compact algorithm for rectification of stereo pairs. *Machine Vision and Applications*, 12(1):16–22, 2000.
- [12] H. Hattori and N. Takeda. Dense stereo matching in restricted disparity space. In *Intelligent Vehicles Symposium*, pages 118 – 123, 2005.
- [13] L. Hong and G. Chen. Segment-based stereo matching using graph cuts. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, volume 1, pages 74 – 81, 2004.
- [14] S. Ivekovic and E. Trucco. Human body pose estimation with pso. In *World Congress on Computational Intelligence, WCCI 2006*, pages 1256 – 1263, 2006.
- [15] S. Ivekovic and E. Trucco. Fitting subdivision surface models to noisy and incomplete 3-d data. In *Proceedings of Mirage 2007*, pages 542 – 554, 2007.
- [16] T. Kanade, P. Rander, and P.J. Narayanan. Virtualized reality: Constructing virtual worlds from real scenes. *IEEE Multimedia*, 4(1):34 – 47, 1997.
- [17] P. Kauff, N. Atzpadin, C. Fehn, M. Mller, O. Schreer, A. Smolic, and R. Tanger. Depth map creation and image-based rendering for advanced 3dtv services providing interoperability and scalability. *Elsevier Signal Processing: Image Communication*, 22(2):217 – 234, 2007.
- [18] J. Kennedy and R. Eberhart. Particle swarm optimization. In *Proceedings of the IEEE International Conference on Neural Networks*, volume 4, pages 1942–1948, 1995.
- [19] B. Matei and P. Meer. A general method for errors-in-variables problems in computer vision. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, pages 18 – 25, 2000.
- [20] W. Matusik, C. Buehler, and L. McMillan. Polyhedral visual hulls for real-time rendering. In *Proceedings of the 12th Eurographics Workshop on Rendering Techniques*, pages 115 – 126, 2001.
- [21] P. Mordohai and G. Medioni. Stereo using monocular cues within the tensor voting framework. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(6):968 – 982, 2006.
- [22] R. Plaenkers and P. Fua. Articulated soft objects for multi-view shape and motion capture. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 25(10), 2003.
- [23] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision*, 47(1/2/3):7 – 42, 2002.
- [24] D. Scharstein and R. Szeliski. Middlebury stereo vision page. <http://vision.middlebury.edu/stereo/>, 3. August 2007.
- [25] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, pages 519 – 526, 2006.
- [26] J. Starck and A. Hilton. Model-based multiple view reconstruction of people. In *IEEE International Conference on Computer Vision*, pages 915 – 922, 2003.
- [27] R. Szeliski and P. Golland. Stereo matching with transparency and matting. *International Journal of Computer Vision*, 32(1):45–61, 1999.
- [28] N. Thakoor, S. Jung, and J. Gao. Real-time planar surface segmentation in disparity space. In *Proceedings of IEEE Computer Vision and Pattern Recognition*, pages 1–8, 2007.
- [29] M. Waschbüsch, S. Würmlin, D. Cotting, and M. Gross. Point-sampled 3d video of real-world scenes. *Image Communication*, 22(2):203 – 216, 2007.
- [30] C.L. Zitnick, S.B. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. In *ACM Transactions on Graphics 23(3), Proceedings of SIGGRAPH 2004*, pages 600 – 608, 2004.

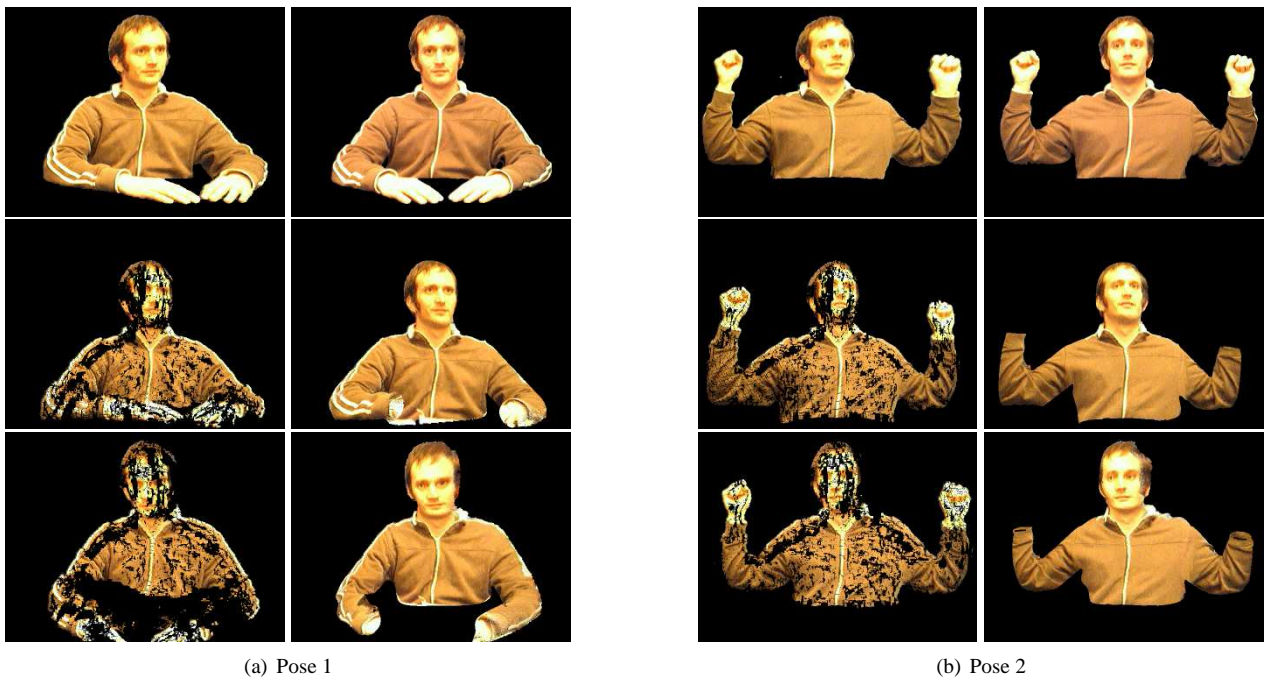


Figure 11: View synthesis with post-processed disparities. For each pose: original stereo pair in row 1; rows 2 and 3 compare the novel view synthesised with original disparities (on the left) with the novel view synthesised with the model-post-processed disparities (on the right). 2 virtual views are shown, in row 2 interpolated below, and in row 3 extrapolated above the baseline.



Figure 12: View synthesis with an extrapolated virtual view. The synthesised results look convincing, however, the effects of the model-based approach, such as the unrealistic shape of the head, also become obvious when extrapolating.

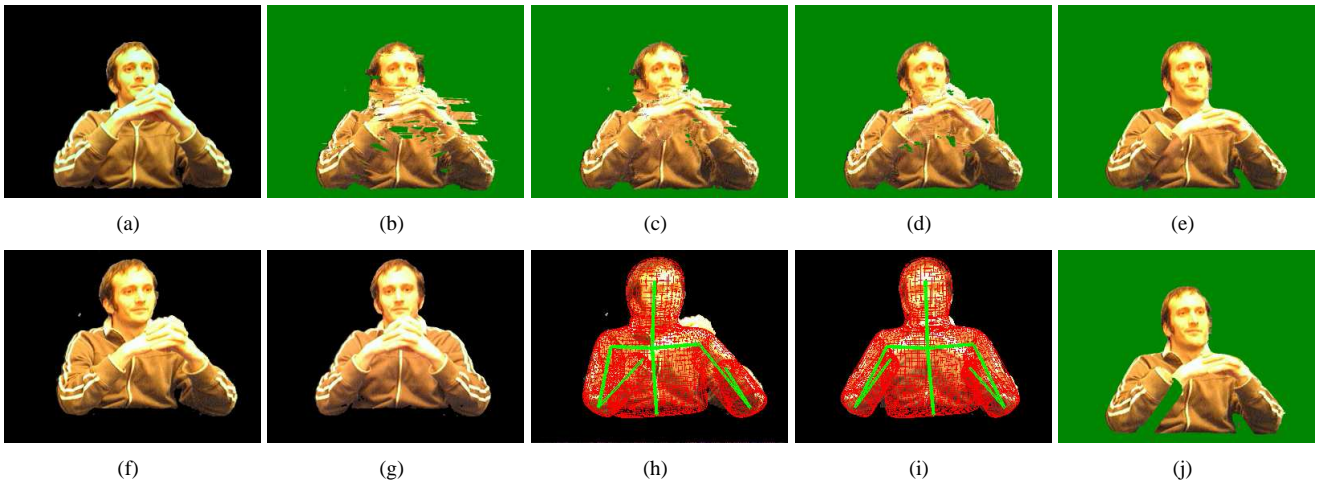


Figure 13: Comparison of view synthesis results for the ground truth test view (a) and disparity maps post-processed using (b) linear interpolation, (c) bilinear interpolation, (d) spline interpolation and (e) disparity space post-processing; (f-g) the original stereo pair, (h-i) the disparity model in the correct pose, (j) partially synthesised view with texture only from the left camera explains the arm artifact in (e) that appears when the right camera's texture is added to complete the view.