On 3-D Scene Flow and Structure Recovery From Multiview Image Sequences

Ye Zhang and Chandra Kambhamettu, *IEEE, Member*

***Abstract*—Two novel systems computing dense three-dimensional (3-D) scene flow and structure from multiview image sequences are described in this paper. We do not assume rigidity of the scene motion, thus allowing for nonrigid motion in the scene. The first system, integrated model-based system (IMS), assumes that each small local image region is undergoing 3-D affine motion. Non-linear motion model fitting based on both optical flow constraints and stereo constraints is then carried out on each local region in order to simultaneously estimate 3-D motion correspondences and structure. The second system is based on extended gradient-based system (EGS), a natural extension of two–dimensional (2-D) optical flow computation. In this method, a new hierarchical rule-based stereo matching algorithm is first developed to estimate the initial disparity map. Different available constraints under a multiview camera setup are further investigated and utilized in the proposed motion estimation. We use image segmentation information to adopt and maintain the motion and depth discontinuities. Within the framework for EGS, we present two different formulations for 3-D scene flow and structure computation. One formulation assumes that initial disparity map is accurate, while the other does not. Experimental results on both synthetic and real imagery demonstrate the effectiveness of our 3-D motion and structure recovery schemes. Empirical comparison between IMS and EGS is also reported.**

***Index Terms*—Nonrigid motion, motion/depth boundary, scene flow, scene structure, stereo, 3-D affine motion model.**

I. Introduction

Motion and structure are fundamental problems in computer vision. Most motion estimation methods (e.g., [1]–[4]) compute optical flow, i.e., the apparent motion between several frames of an image sequence. In the recent past of visual-motion research, numerous applications including tracking, surveillance, recognition, etc., have been intensively utilizing optical flow information. However, optical flow only provides projected two-dimensional (2-D) motion information. It is clear that ambiguities exist when dynamic three–dimensional (3-D) objects/scenes are explained by using 2-D optical flow. This is why the counterpart of optical flow in 3-D space, 3-D scene flow, is introduced [5]–[7]. Like optical flow, 3-D scene flow is defined at every point in a reference image. The difference is that the velocity vector in scene flow field contains not only x, y, but z velocities. This also means that a multiview camera setup is usually required to compute reliable 3-D scene flow.

In this paper, we present two novel systems, integrated model-based system (IMS) and extended gradient-based system (EGS), which compute dense 3-D scene flow and structure from multiple synchronized video streams. These two systems were initially presented in [6] and [7]. In this journal version, we fully describe the formulations and report our recent experimental results and empirical comparison between IMS and EGS.

A. Previous Work

There has been considerable interest in recovering 3-D motion and structure from monocular view image sequences (e.g., [8]–[11]). Unfortunately, because the scene is viewed from only one camera, strong limitations are imposed on the types of motions that can be recovered and on the scenes that can be analyzed. There has also been a lot of work on stereo vision for the recovery of dense scene structure from multiview images (e.g., [12]–[14]). When monocular motion analysis and stereo vision are considered separately, each of them has its own inherent difficulties. Monocular motion analysis normally involves solving for point correspondences, or nonlinear equations. Thus the computation is very sensitive to noise. Moreover, the 3-D motion interpretation is difficult due to the structure ambiguities. On the other hand, stereo vision needs to solve *correspondence problem*, i.e., matching features between stereo image pairs. This problem, in general, is under-determined. Other heuristics from the scene are desirable. It is natural to consider integrating motion and stereo to complement each other's performance.

By assuming that the scene is rigid, some researchers have considered fusing motion and stereo to get better results. Richards [15] described the defects in stereo and motion parallax (i.e., structure from motion) respectively and integrated them to recover 3-D rigid shape. In his method the only goal was to recover 3-D structure—motion analysis did not benefit from stereo analysis. Ballard *et al.* [16], Huang *et al.* [17], Mutch *et al.* [18], Young *et al.* [19] and Balasubramanyam *et al.* [20] simply computed the rigid motion parameters assuming the depth was known or had been computed by stereo vision. Waxman *et al.* [21] used the difference between the flow fields of the left and right cameras to analyze the cases with unknown motion and structure. But their method assumed that the viewed surfaces were planar. Barron *et al.* [22] developed a relation between binocular velocity fields and the motion/structure parameters. A nonlinear method was then presented to simultaneously compute the motion and structure. Aloimonos *et al.* [23] used two cameras to recover surface structure, then

Manuscript received November 22, 2001; revised January 16, 2003. This work was supported by the National Science Foundation under Grants NSF IRI-9619240, CAREER IRI-9984842, and CISE CDA-9703088. This paper was recommended by Guest Editor H. Zha.

The authors are with the Video/Image Modeling and Synthesis (VIMS) Laboratory, Department of Computer and Information Sciences, University of Delaware, Newark, DE 19716 USA (e-mail: zhangye@cis.udel.edu; chandra@cis.udel.edu).

Digital Object Identifier 10.1109/TSMCB.2003.814284

utilized the positions of feature points in the stereo image pairs to decide the direction of translation. More recently, Dornaika *et al.* [24] recovered the stereo correspondence using one motion of a stereo rig. Like in [15], their approach did not refine and improve motion analysis through coupling stereo and motion. Li *et al.* [25] proposed a two-step fusing procedure. First, translational motion parameters were found from binocular image flows. Then the stereo correspondences were estimated with the knowledge of motion parameters. They relaxed the planarity assumption compared with [21]. However, the 3-D motion was still restricted to translational motion. Weng *et al.* [26] designed another two-step approach. A linear algorithm is first used for a preliminary estimate of rigid motion parameters, then an optimal objective function is minimized using the previous result as an initial guess.

As can be seen from many real-world examples (trees, human body parts, etc.), the presence of nonrigid motion is imperative and needs special attention in motion analysis. There has been very limited research on integration of nonrigid motion and stereo. Liao *et al.* [27] used a relation-based algorithm to co-operatively match features in both temporal and spatial domains. It therefore does not provide dense motion. Malassiotis *et al.* [28] used a grid deformable model to generalize the monocular approaches. However, model-based approach requires *a priori* knowledge of the scene. Kambhamettu *et al.* [29] coupled stereo and nonrigid motion analysis in a multi-resolution manner and designed a hierarchical framework to analyze time-varying cloud images. But they still computed motion and stereo correspondences in separate modules. Vedula *et al.* [5] defined and computed the nonrigid 3-D scene flow. They designed linear algorithms for three different scenarios. In their work, multiview optical flow was used to estimate scene flow. Then scene structure was estimated from scene flow. Their work is innovative and their method is efficient. However, stereo constraints were not fully utilized in the scene structure recovery. Neumann *et al.* [30] presented an algorithm in which the nonrigid object is modeled by a time-varying multi-resolution subdivision surface that is fitted to the image data using spatial-temporal stereo information and contour constraints. In their method, object space parameterization was utilized instead of dense scene parameterization.

A disadvantage of most of the above approaches is that the motion-stereo integrations were done in a biased way, i.e., using either structure to optimize motion or motion to optimize structure, rather than mutually benefiting both the analyses. Moreover, 3-D motion and stereo analyses were carried in different modules. The integration was not essentially coupled and was more of a post-processing nature.

Generally speaking, to estimate 3-D motion and structure from multiview image sequences, it is desirable to fuse stereo and motion constraints to some extent [21]. However, combining motion/stereo constraints from multiview image sequences requires extra caution. This is because some points in the reference image may not be visible (occluded) in other views. If the algorithm is not aware of this and still combines the motion/stereo constraints from the occluded view, the results could be very erroneous.

B. Our Approaches

In this paper, our goal is to accurately recover dense 3-D scene flow and scene structure from multiple synchronized video streams. No *a priori* knowledge of the scene is assumed, nor do we assume that the scene is rigid. We designed two systems to compute 3-D scene flow and structure.

The first system, IMS, integrates 3-D scene flow and structure recovery in order to complement each other's performance. We assume each local image region (normally 3×3) is undergoing similar motion which can be represented by a 3-D affine model. Non-linear least square method is used to fit the motion model at each small region. Main contributions of IMS include, 1) formulation to simultaneously recover 3-D motion and structure, and 2) seamless integration of 2-D motion and stereo constraints.

The second system, EGS, can be thought of as a natural extension of 2-D optical flow computation. We first design a hierarchical rule-based stereo matching algorithm employing image segmentation information to enforce the depth discontinuities. Then, we formulate 3-D scene flow estimation as an energy minimization problem based on optical flow constraints from different views. We show two different formulations for this problem. One formulation assumes that initial depth map computed by using our stereo matching algorithm is accurate while the other does not make this assumption. Main contributions of EGS include the following:

1) stereo matching algorithm producing smooth and detailed disparity maps with occlusions explicitly detected;
2) design of system extending traditional two dimensional optical flow computation to three dimensional case;
3) method utilizing image segmentation information to maintain reliable motion and depth discontinuities.

The rest of this paper is organized as follows. Section II describes integrated model-based system. Multiple camera geometry is briefly discussed in Section II.A. 3-D affine motion model and local model fitting are presented in Sections II.B and II.C, respectively. Initial guesses and egularization constraints are introduced in Section II.E. Complete recursive algorithm is proposed in Section II.F. Section III describes EGS. Image segmentation is briefly discussed in Section III.A. Our new stereo matching algorithm is described in Section III.B. Available motion constraints are investigated in Section III.C. Hard constraints are explained in Section III.D. Two different formulations of EGS are presented in Sections III-E and III.F, respectively. Section IV reports the experimental results of the proposed systems on both synthetic and real imagery. Empirical comparison between IMS and EGS is also reported. Section V presents conclusions and future work of this research.

II. IMS

The block diagram of IMS is presented in Fig. 1. We assume that the imaging cameras are calibrated. Optical flow, stereo constraints and regularization constraints are used to fit 3-D affine model for each small region. 3-D scene flow, 3-D correspondences and dense scene structure are simultaneously computed.

A. Multiple Camera Geometry

Several multiple camera algorithms for stereo analysis have been proposed in the past [14], [31], [21], [29]. In our system, we utilize multiple cameras in a manner similar to [14] and [31]. A pair of cameras are used as a *reference* or *basic stereo pair*. Other cameras provide extra information, thus contributing additional constraints.

At a given instance, a set of N cameras $C_0, C_1, \ldots C_{n-1}$ provide N images $I_0, I_1, \ldots I_{n-1}$, respectively. We use C_0, C_1 as the basic stereo pair. C_0 provides the basic view for which we intend to compute the 3-D scene flow and disparity map for each image point. A 3-D point $\mathbf{P}$ expressed in world coordinates with homogeneous coordinates $(x, y, z, 1)$ can be transformed to point $\mathbf{m}_i = (X_i, Y_i, 1)$ in the image plane of camera i by the relation

$$\mathbf{m}_i = \mathbf{J}_i \mathbf{W}_i \mathbf{P} = \mathbf{T}_i \mathbf{P} \tag{1}$$

where $\mathbf{J}_i$ is the *projection matrix*, $\mathbf{W}_i$ is the *camera position/orientation matrix* and $\mathbf{T}_i$ is the *camera calibration matrix*.

During the process of stereo analysis, each point $\mathbf{m}$ of Image I_0 is assigned a disparity d, or equivalently a depth z. We can transform $\mathbf{m}$ to a 3-D point $\mathbf{P_m}$ in world coordinates

$$\mathbf{P_m} = \mathbf{W}_0^{-1} \begin{pmatrix} \mathbf{m} \\ d \end{pmatrix}. \tag{2}$$

Therefore, for each base image point $\mathbf{m}$ and its disparity d, we have a set of $N-1$ re-projected stereo correspondences on the image planes of cameras $C_1, C_2, \ldots C_{n-1}$ which is represented by $\mathbf{R}$

$$\mathbf{R} = \{\mathbf{T}_i \mathbf{P_m}\}, \quad i \in [1, 2, \ldots n-1]. \tag{3}$$

B. Local Motion Model Selection

In order to describe 3-D motion without rigidity assumption, it is important to choose a motion model powerful enough to describe different kinds of nonrigid motion [32]. There have been many works which use 2-D affine motion model for image matching [33]. More recently, Ju *et al.* [34] and Bergen [33] have used 2-D affine model to estimate image motion. Li *et al.* [35] and Zhou *et al.* [10], [36], [37] have used 3-D affine model to analyze face and cloud nonrigid motion respectively, indicating the use of affine model in describing complex nonrigid motion. In our work, we utilize 3-D affine model to describe the underlying nonrigid motion in the scene.

Consider a 3-D point in the scene. In frame t, it is represented by a homogeneous vector $\mathbf{P}_\mathbf{m}^t = (x_m^t, y_m^t, z_m^t, 1)$. Assume that the point moves to a new position $\mathbf{P}_\mathbf{m}^{t+1} = (x_m^{t+1}, y_m^{t+1}, z_m^{t+1}, 1)$ in frame $t+1$. Then affine motion model can be represented as

$$\mathbf{P}_\mathbf{m}^{t+1} = \mathbf{M}^t \mathbf{P}_\mathbf{m}^t \tag{4}$$

where

$$\mathbf{M}^t = \begin{pmatrix} a_1^t, & b_1^t & c_1^t & d_1^t \\ a_2^t, & b_2^t & c_2^t & d_2^t \\ a_3^t, & b_3^t & c_3^t & d_3^t \\ 0, & 0 & 0 & 1 \end{pmatrix}. \tag{5}$$

One advantage of 3-D affine model is that it provides a simple way to combine nonrigid motion $(\mathbf{M}^t)$ and structure (z_m^t). The *dual* problem of motion and structure analyses can then be formulated into a *single* model fitting problem. Motion constraints and stereo constraints can be considered together during model fitting, thus integrating motion and structure analyses in a seamless manner. Unfortunately, although the mathematical form of the above motion model is simple, it is impossible to directly use (4) on the whole image in order to estimate 3-D motion. This is because $\mathbf{M}^t$ is, in general, point dependent during nonrigid motion. However, in practice, motion field is spatially smooth. Thus if we apply affine model (4) locally, we can assume $\mathbf{M}^t$ is point independent. This means we have to segment the images into local regions.

Obviously, the optimal segmentation should aggregate points having similar motion. Although we don't have any *a priori* knowledge of the motion in the scene, one may argue that it is still possible to segment the images according to optical flow information (e.g., [38]). Nonrigid motion segmentation (and thus image segmentation) is an ongoing research topic. We do not incorporate it in our system. In practice, we segment the images evenly. Through experiments, we found that if the local region is small enough, this approach generates good results.

To avoid overfitting and ensure convergence in each small region, we need more constraints during nonlinear model fitting. Zhou *et al.* [10] introduced stronger constraints by not only assuming spatial smoothness but also assuming temporal smoothness. Motion of each region in successive S frames is assumed to be temporally smooth but not necessarily of the same scale. This means that the difference between the motion matrix $\mathbf{M}^t$ of a local region in successive S frames can be defined by a scaling factor α^t. So, $\mathbf{M}^t$ in successive S frames can be represented as,

$$\mathbf{M}^t = \alpha^t \begin{pmatrix} a_1^\tau, & b_1^\tau & c_1^\tau & d_1^\tau \\ a_2^\tau, & b_2^\tau & c_2^\tau & d_2^\tau \\ a_3^\tau, & b_3^\tau & c_3^\tau & d_3^\tau \\ 0, & 0 & 0 & 1 \end{pmatrix}, \quad t \in [\tau, \tau + S). \tag{6}$$

Equation (6) reduces the number of unknowns in successive frames for each small region, thus improving the robustness of nonlinear fitting.

C. Motion Model Fitting

Balasubramanian *et al.* [39] and Zhou *et al.* [10] discussed how to fit affine models on local regions of monocular image sequences. Since we have multiview image sequences, our constraints are further enriched. In our system, Levenberg–Marquart (LM) [40] nonlinear method is used to estimate $\mathbf{M}^t$ and z_m^t in each local region. Other gradient search algorithms are also tested. We find that LM algorithm gives best results for the given formulation.

During model fitting, we eliminate the translation unknowns by fixing d_1^i, d_2^i, and d_3^i to small constants. This is to avoid *trivial solutions*, where all other unknowns are 0 except d_1^i, d_2^i, and d_3^i. Thus, if the local region size is $w \times h$ and we assume that motion is temporally smooth in successive S frames, we have $9 + w \times h + S - 2$ unknowns in (4) for each region. The unknown vector is represented by

$$\mathbf{U}_t = \left(a_1^\tau, a_2^\tau, \ldots, c_3^\tau, \alpha^{\tau+1}, \ldots \alpha^{\tau+S-2}, z_1, z_2, \ldots, z_{wh}\right)$$

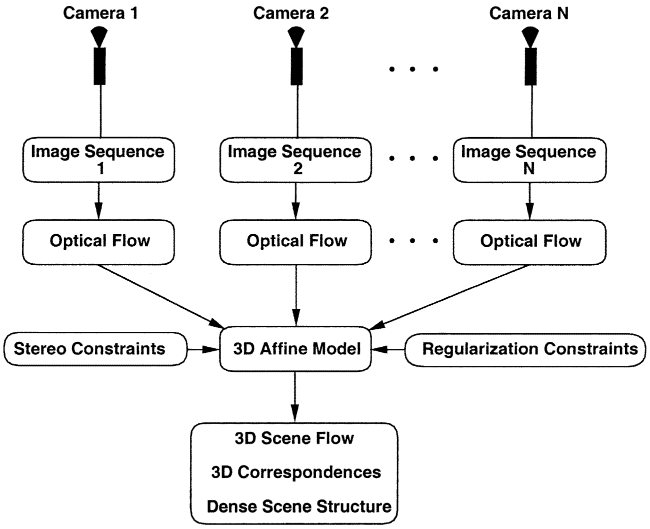


Fig. 1. Block diagram of IMS.

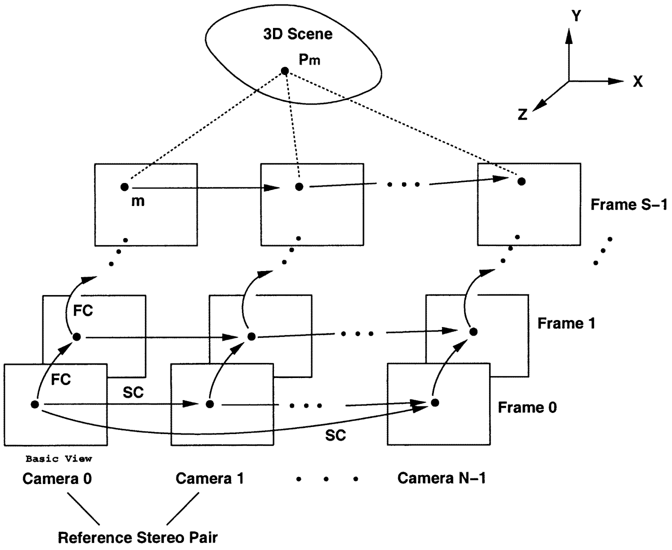


Fig. 2. Constraints for local fitting: SC denotes stereo constraints; FC denotes optical flow constraints.

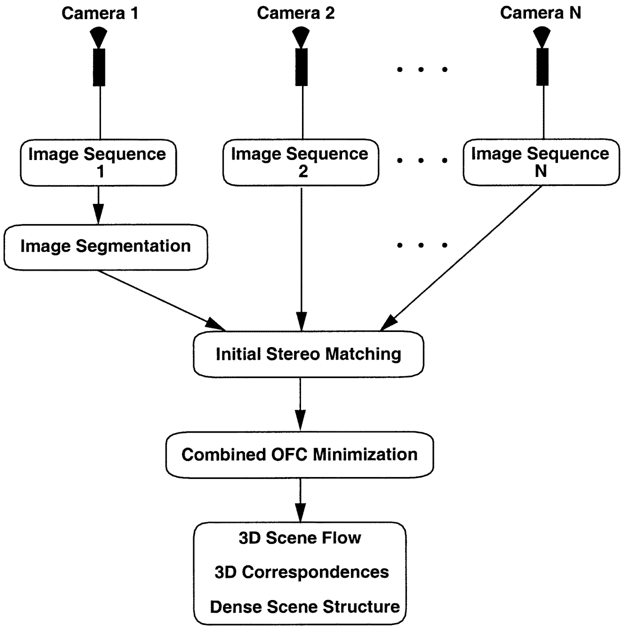


Fig. 3. Block diagram of EGS.

where $z_1, z_2, \ldots, z_{wh}$ is the depth for the first basic frame. The local model fitting can be formulated as,

$$\mathbf{U}_t^* = \arg\left(\min_t(\mathrm{EOF}(\mathbf{U}_t))\right) \tag{7}$$

where $\mathbf{U}_t^*$ is the optimal unknown vector and $\mathrm{EOF}(\mathbf{U}_t)$ is the *error-of-fitting* function which is to be minimized.

It is crucial to define a good EOF function. The rest of this section addresses this problem. First, we introduce the local constraints (i.e., optical flow and stereo constraints), then the regularization constraints are presented. Finally, a complete recursive algorithm which incorporates all the available constraints is presented.

1) Optical Flow and Stereo Constraints: The optical flow for each image sequence gathered by each camera is first computed. We use the method described in [3] to preserve the discontinuity in motion field. We denote the optical flow of point $\mathbf{m}_j$ on image plane j as $\mathbf{U}_j(\mathbf{m}_j) = (u, v)$. The next step is to design the EOF for local motion model fitting according to optical flow and stereo constraints. In frame t, once a base image point $\mathbf{m}$ is assigned a disparity value d, it can be transformed to a world coordinate 3-D point $\mathbf{P}^t_{\mathbf{m}}$ in the scene by (2). Clearly, from frame t to frame $t+1$, the 2-D motion of the projective point of $\mathbf{P}^t_{\mathbf{m}}$ on the image plane of camera j can be computed as

$$\mathbf{V}_j(\mathbf{m}, t) = \mathbf{H}\left(\mathbf{T}_j\left(\mathbf{M}^t\mathbf{P}^t_{\mathbf{m}} - \mathbf{P}^t_{\mathbf{m}}\right)\right) \tag{8}$$

where

$$\mathbf{H}\left(\begin{pmatrix} x \\ y \\ w \end{pmatrix}\right) = \begin{pmatrix} \frac{x}{w} \\ \frac{y}{w} \end{pmatrix} \tag{9}$$

is homogenizing function.

The optical flow of the projective point of $\mathbf{P}^t_{\mathbf{m}}$ on the image plane of camera j is represented by function $\mathbf{F}_j$

$$\mathbf{F}_j(\mathbf{m}, t) = \mathbf{U}_j\left(\mathbf{T}_j\mathbf{P}^t_{\mathbf{m}}\right). \tag{10}$$

It is evident that the optical flow and the projected 2-D motion of $\mathbf{P}^t_{\mathbf{m}}$ should be compatible. Thus from (8) and (10), the optical flow constraint can be represented as

$$\|\mathbf{V}_j(\mathbf{m}, t) - \mathbf{F}_j(\mathbf{m}, t)\| \to 0. \tag{11}$$

The stereo constraint is essentially the similarity measurement between the potential stereo correspondences. In our work, we use cross-correlation measure. If the potential stereo correspondence of $\mathbf{m}_1$ of camera i is $\mathbf{m}_2$ of camera j, we denote their cross-correlations as $\mathrm{Corel}_{\mathrm{i,j}}(\mathbf{m}_1, \mathbf{m}_2)$. The range of Corel is $[0, 1]$ and "1" means well correlated. Thus if $\mathbf{m}$ is assigned a good disparity, at frame $t+1$ we have

$$\mathrm{Corel}_{i,j}\left(\mathbf{T}_i\mathbf{M}^t\mathbf{P}^t_{\mathbf{m}}, \mathbf{T}_j\mathbf{M}^t\mathbf{P}^t_{\mathbf{m}}\right) \to 1, \quad i, j \in [0, N). \tag{12}$$

Optical flow and stereo constraints are illustrated in Fig. 2. In a local region A, the *error-of-fitting* for successive S frames can be defined at the bottom of the page as (13), where w is a weight.

To ensure robust convergence, this weight is decided adaptively. Generally speaking, when the variation of local optical flow is too small, or the error from the optical flow constraint is too large, w should be increased. This prevents the motion field from overwhelming the similarity measurement.

Equation (13) is then used to solve (7) in our recursive algorithm. Clearly, in this EOF formulation, optical flow and stereo constraints are considered together in order to estimate 3-D motion and structure simultaneously (as shown in (13) at the bottom of the page.)

D. Initial Guesses

We use LM algorithm to solve (7). As mentioned before, other numerical algorithms (e.g., Powell algorithm) have also been tested. However, we find that LM algorithm gives us the best results. To solve (7) numerically, initial guess for the unknown vector $\mathbf{U}_t$ is needed. If we assume small motion between two adjacent frames (this assumption holds in most cases), the motion parameters can be initialized as $a_1^\tau = 1, b_1^\tau = 0, c_1^\tau = 0, a_2^\tau = 0, b_2^\tau = 1, c_2^\tau = 0, a_3^\tau = 0, b_3^\tau = 0, c_3^\tau = 1, \alpha^{\tau+1} = \cdots = \alpha^{\tau+S-2} = 1$. We also need the initial depth guess for frame 0. Zhou *et al.* in [10] simply assume that the accurate depth for the first frame is given. In our case, the initial first frame depth can be computed by any stereo algorithm.

E. Regularization Constraints

In the above optimization scheme, the affine model is fitted for each small region independently. Thus, there is a need to regularize noisy data. Since $x-y$ motion field has been regularized during optical flow computation, we only need to deal with the z velocity. One of the most frequently adopted regularization constraint is *motion smoothness*, which has been widely used to compute ill-posed optical flow. However, it is well known that smoothness constraints lose motion discontinuities. Many researchers (e.g., [3], [41]) addressed how to preserve discontinuity in optical flow computation. Experiments have shown [41] that the partial derivatives of image intensity provides a reliable measure of goodness of regularization. If the partial derivatives are small at some image point, high amount of regularization should be performed to propagate the flow vectors to that point from neighboring points. Otherwise, the regularization term should be kept small. This means that in order to regularize accurately, it is necessary to apply data-weighted smoothness. Intuitively, the discontinuity preserving smoothness term can be defined as

$$C'_R = \frac{\lambda}{\|\nabla I\| + \|\nabla D\|}\|\nabla V_z\| \tag{14}$$

where I, D, V_z denote the image intensity, the disparity and the z velocity at image point $\mathbf{m}$, respectively and λ is a small constant.

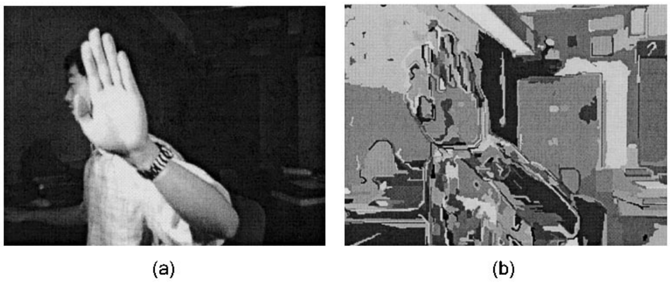


Fig. 4. (a) Input image. (b) Segmentation result.

However, the above constraint cannot be directly used because we want to smooth the motion across the local regions. Thus we redefine this constraint as

$$C_R = \frac{\lambda}{\|\nabla I\| + \|\nabla D\|}\|V_z - \bar{V}_z\| \tag{15}$$

where $\bar{V}_z$ is the average z velocity in Q adjacent regions in the previous iteration.

Equation (15) is applied in a recursive manner: in the first iteration, it is not used. In the following iterations, it is added into the EOF. It is well known that numerical solution suffers from local minima. In our case, local minima may happen if the search range of z is not confined. This is due to structure ambiguities. According to small motion assumption, we define a penalty constraint

$$C_P = \gamma \min(|z_{i+1} - z_i| - r, 0.0) \tag{16}$$

where z_i and z_{i+1} are the depth values of corresponding points in frames i and $i+1, r$ is a positive constant indicating the specified range and γ is a large constant. This constraint is added into the EOF during local model fitting. Clearly, if $z_{i+1} > z_i + r$ or $z_{i+1} < z_i - r$, the EOF is penalized.

F. Recursive Algorithm

To incorporate all the above constraints, a recursive algorithm is designed as Algorithm 1. In our experiments, this algorithm converges in 3–4 iterations.

Algorithm 1: A Recursive Algorithm for 3-D Scene Flow and Structure Recovery

```
begin
    Initialize depth map and motion parameters.
    Set flag := 0.
    while (regularization constraint is greater than
        a threshold and maximum number of
        iterations has not been exceeded) do
        for $i$ := 1 to $n$ regions step 1 do
```

$$\text{EOF} = \sum_{t=\tau}^{\tau+S-1}\sum_{\mathbf{m}\in A}\sum_{j=0}^{N-1}\|\mathbf{V}_j(\mathbf{m},t) - \mathbf{F}_j(\mathbf{m},t)\| - w\sum_{t=\tau}^{\tau+S-1}\sum_{\mathbf{m}\in A}\sum_{i,j=0,i\neq j}^{N-1}\underset{i,j}{\text{Corel}}\left(\mathbf{T}_i\mathbf{M}^t\mathbf{P}_{\mathbf{m}}^t, \mathbf{T}_j\mathbf{M}^t\mathbf{P}_{\mathbf{m}}^t\right) \tag{13}$$

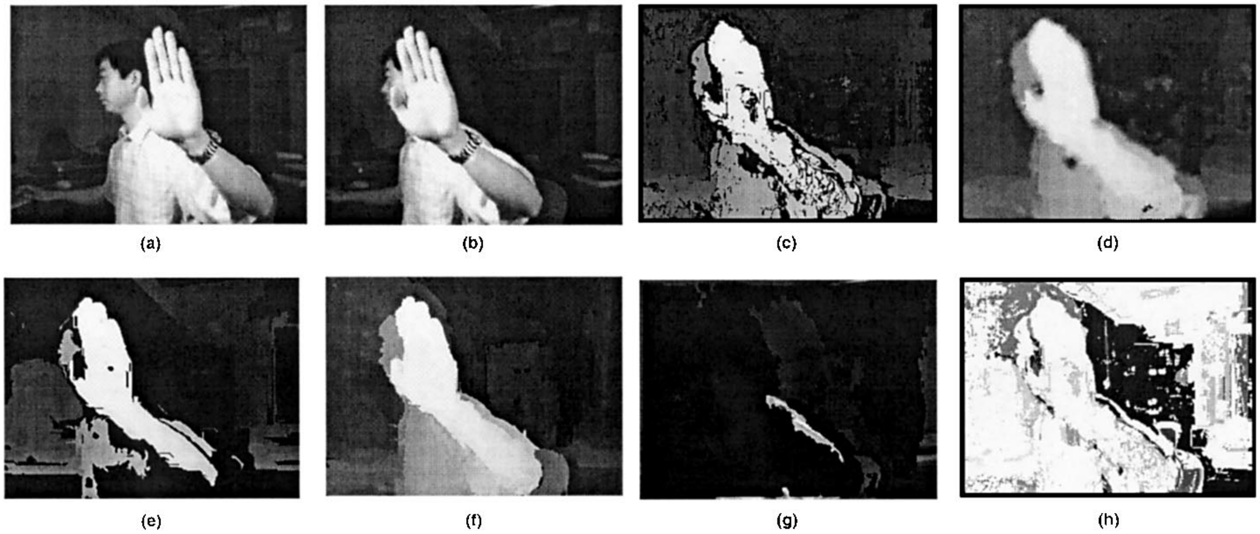


Fig. 5. Stereo results on a snapshot of Point Grey Research, Inc. Triclops system. (a) and (b) Left and right (reference) views, respectively. Top view is not shown here. (c) Sparse disparity map generated by merging the valid points from three views and two level Gaussian pyramid. (d)Dense disparity map generated by a direct method. Second row shows the output of our stereo matching algorithm. (e) Segments labeled as *VALID* after we applied Rule 1 and 2. (f) Final dense disparity map. (g) Detected regions in the reference view which are occluded in the left view. (h) Confidence measurement map of each point—brighter means higher confidence.

```
    if (flag = 1)
      then
        Local model fitting: add (15),
        (16) into (13), then solve
        (7) in region $i$;
      else
        Local model fitting: add (16)
        into (13), then solve (7)
        in region $i$; od fi
    Compute $\bar{V}_z$ in adjacent $Q$ regions.
    Set flag := 1. od
end
```

III. Extended Gradient-Based System (EGS)

The block diagram of EGS is shown in Fig. 3. Combined optical flow constraints are minimized to compute dense 3-D scene flow and structure.

In the following description, some assumptions have been made without loss of generality. First, we assume that all the other cameras are in standard (parallel) set up with the reference camera C_0. Also, all the camera parameters are known and the image sequences captured from different cameras are well rectified. This makes it easy to discuss how to combine constraints from different views in EGS. Second, we assume that the motion and depth in each image segment are smooth, thus justifying the enforcement of smoothness constraint within each image segment. It is worthwhile to note that the smoothness constraint is not unconditionally applied to the entire image. This is the reason why our method can maintain sharp motion and depth boundaries. The images captured by camera k is denoted as $\mathbf{I_{k,t}}(I_{k,0}, I_{k,1}, \ldots)$, where t represents time. The disparity value at point P in frame t in the reference view is denoted as d_t. 3-D scene flow at point P is denoted as (u, v, w), where u, v are actually the components of optical flow vector. w is defined as the disparity motion $d_{t+1} - d_t$.

A. Image Segmentation

For the experiments performed by using EGS we have used the graph-based image segmentation proposed in [42]. As discussed before, we assume that there is no large motion or disparity discontinuities within an image segment so that smoothness constraint can be applied to each image segment. This guarantees the smoothness in textureless regions because a textureless region tends to be grouped as one segment. On the contrary, we do not enforce smoothness across the boundaries for actually smooth but highly textured regions because these regions are easily over-segmented. However, this is usually not a problem since motion and structure estimation tends to be reliable in textured regions even without the smoothness constraint. In other words, over-segmentation can be tolerated to some extent in our framework. Since smoothness is not applied to the entire image, sharp motion/depth boundaries can be maintained. A typical result of this image segmentation algorithm is shown in Fig. 4.

B. Initial Stereo Matching

To get reliable 3-D scene flow, we require that the initial disparity map be smooth and detailed. We also hope that continuous and even surfaces produce a region of smooth disparity values with their boundary precisely delineated, while small surface elements are detected as separate distinguishable regions. Furthermore, an ideal initial stereo matching algorithm should explicitly identify and report the occluded area and provide a confidence measurement of the computed disparity at every image point. As will be discussed later, the occlusion and confidence information is important in 3-D scene flow estimation. Though obviously desirable, it is not easy for a stereo algorithm to satisfy all these requirements at the same time.

Inspired by the work of cross-validation [43], image segmentation, and prediction error testing [44], we propose a hierarchical rule-based stereo matching algorithm. A set of rules for

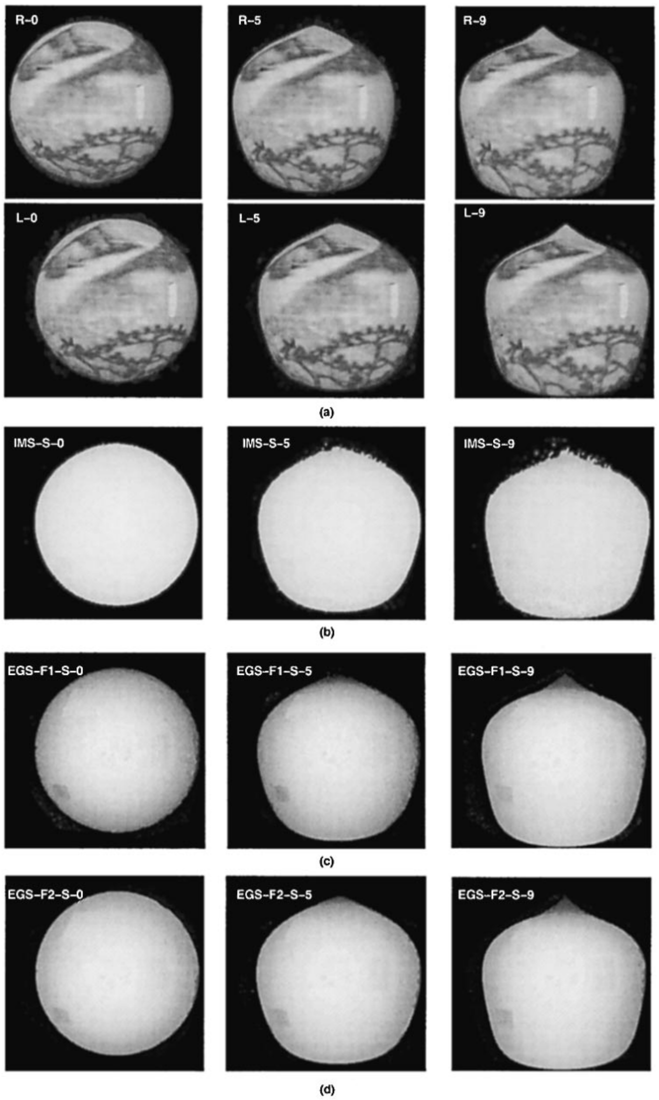


Fig. 6. Structure results on the synthetic image sequence. (a) Two-view synthetic image sequences; (b)–(d) illustrate the scene structure recovered by IMS, EGS-F1, and EGS-F2, respectively.

depth hypothesis is defined to guide the matching process. The output includes a disparity map, an occlusion map, and a confidence map.

First, a correlation volume $\mathcal{C}(x, y, d)$ is computed between image $I_{0,0}$ and $I_{1,0}$ along the epipolar line. The measured disparity is the one with the largest matching score. We perform the correlation twice by reversing the roles of the two images and consider as valid only those matches for which we measure the same depth at corresponding points when matching from $I_{0,0}$ to $I_{1,0}$ and from $I_{1,0}$ to $I_{0,0}$. Following this method, we compute the valid disparities between $I_{0,0}$ and $I_{2,0}, I_{3,0}, \ldots, I_{N-1,0}$, respectively. Then we merge the results together and get a sparse initial "valid" disparity map. If two sets of views produce different valid disparities, then the one with higher matching score wins. Also, the matching score in the correlation volume is updated accordingly. To further increase the density of the initial valid disparity map, a pyramid control strategy is employed as suggested in [43]. Valid matches from two resolution levels are merged. To further increase the *Signal/Noise* ratio, we filter out those valid points that are either isolated or have very large standard deviation in a small neighborhood. Finally, we get a initial disparity map with few errors.

Second, we segment the reference image into small regions. We then label each image segment as following:

$$L(s) = \begin{cases} \text{VALID} & \text{if } r \geq \alpha_1; \\ \text{SEMIVALID} & \text{if } \alpha_2 \leq r < \alpha_1; \\ \text{INVALID} & \text{if } r < \alpha_2, \end{cases} \tag{17}$$

where r is the ratio of valid disparity points in segment s, α_1 and α_2 lie between 0 and 1 (in our experiments, we set $\alpha_1 = 0.9$ and $\alpha_2 = 0.6$), and *VALID*, *SEMIVALID*, and *INVALID* are all symbolic values. *VALID* means that we have high confidence on the disparity map within segment s. *INVALID* means low confidence, and *SEMIVALID* means medium confidence. This labeling method reflects an assumption we have made: image segments where the valid disparity points are dense are more reliable. Generally, experiments have shown that this assumption holds [43].

After image segmentation and segment labeling, we identify which segments have more reliable information for stereo matching, i.e., which segments are more valid. Thus we can define a set of rules that guides the stereo hypothesis process according to the different labels of the segments. The principle behind these rules is to first process more valid segments, then utilize information in valid segments to hypothesize those neighboring segments with less reliable information. Accordingly, we first define Rule 1 for hypothesizing *VALID* segments, then Rule 2 for *SEMIVALID* segments with *VALID* neighbors, finally Rules 3 and 4 for the rest *SEMIVALID* and *INVALID* segments.

Rule 1 If $L(s)$ equals *VALID*, then the disparity in this segment is filled by interpolating segment s.

Rule 2 If $L(s)$ equals *SEMIVALID*, then search all the neighbors of segment s. If one neighbor n satisfies the following criteria:

a) $L(n)$ equals *VALID*;
b) disparity of n is very similar to that of s;
c) intensity of n is very similar to that of s;

then store the segment number in an array K. After checking all the neighbors, if K is not empty, find the segment k in K which has the most similar disparity compared with segment s. Then the disparity in segment s is filled by interpolating segments s and k as if they were one large segment. Set $L(s)$ to *VALID*.

Rule 3 If $L(s)$ equals *INVALID* or *SEMIVALID*, then hypothesize the disparity in segment s by using one of the *VALID* neighbors k. Then warp the image of this segment to other views by using the hypothesized disparity [44]. The matching score M $(0 \leq M \leq 1)$ between the warped image and original image is stored. If the highest matching score corresponding to a *VALID* neighbor (segment k) satisfies $M > T$ (T is a positive constant. In our experiments, we set it as 0.8.), then the disparity in segment s is filled by

TABLE I
STRUCTURE RECOVERY EVALUATION

Frame Number	**IMS** Initial/Final % Disparity Correct	**EGS-F1** Initial/Final % Disparity Correct	**EGS-F2** Initial/Final % Disparity Correct
0	94.67/96.18	96.75/96.75	96.75/97.12
1	94.29/95.88	97.01/97.01	97.01/97.23
2	94.01/94.71	96.45/96.45	96.45/96.99
3	93.89/94.58	96.59/96.59	96.59/96.87
4	93.56/94.42	96.77/96.77	96.77/97.15
5	93.29/94.40	96.83/96.83	96.83/97.01
6	93.05/94.13	97.12/97.12	97.12/97.21
7	92.97/93.71	96.87/96.87	96.87/97.07
8	92.93/93.52	97.01/97.01	97.01/97.18
9	92.90/93.45	97.11/97.11	97.11/97.21

interpolating segments s and k as if they were one large segment. Set $L(s)$ to *VALID*.

Rule 4 Same as Rule 3 except set T equals 0.

The interpolation used in our algorithm is a membrane model. More details can be found in [7].

In our algorithm, Rule 1 to Rule 4 are applied sequentially on each image segment. Rules 2–4 need to be applied iteratively until there is no more updated segment before moving to the next Rule. In Rule 2, the similarity measurement of intensity/disparity between two segments is simply the absolute difference of their average values. This is because image segmentation guarantees that the intensity within one segment is very similar. Also, if the image is not seriously under-segmented, the disparity variation within a segment should not be large. The worst case happens when very slanted surface exists in the scene. In that case disparity variation in some segments may be large, and Rule 2 tends to reject the hypothesis from the neighborhood. However, this does not matter because Rule 3 or 4 still have good chances to make correct hypothesis. In fact, Rule 2 is designed to deal with over-segmentation. Rule 3 and Rule 4 are separated because we want to give higher priority to those *INVALID* segments with larger matching score after warping. It is to be emphasized that the sequence of applying these rules are important. The general rule is that segments with more valid information should be processed first.

It is straightforward to decide on the confidence map and occlusion map from the above algorithm. The disparity confidence measurement at each point is assigned in a hierarchical manner: we first assign maximum confidence values to three different segments (e.g., assign 1.0 to *VALID* segments, 0.8 to *SEMI-VALID* segments and 0.5 to *INVALID* segments). Each point's confidence should not exceed the maximum confidence of the segment containing this point. Within each segment, we assign the confidence measurement according to the matching score.

After we get the confidence map, the occluded areas are detected by setting a threshold on the confidence map. If the confidence measurement of a point is below the threshold (e.g., 0.45), we label this point as occluded. The results of our stereo matching algorithm is shown in Fig. 5, where the image dimension is 320 by 240 pixels, and the width of the occluded area (widest part) is over 75 pixels. We can see that even with large occlusions, our algorithm still provides good stereo estimates and detects the occluded areas.

C. Motion Constraints

The motion constraints we use in our algorithm actually combine optical flow constraints from every single camera. The optical flow constraint in camera i can be represented as

$$\frac{\partial I_i}{\partial x}u_i + \frac{\partial I_i}{\partial y}v_i + \frac{\partial I_i}{\partial t} = 0. \tag{18}$$

Since we suppose other cameras are all in standard set up with the reference camera, it is easy to derive a way to combine the optical flow constraints. Suppose the focal length of camera 0 is f. At frame t, a 3-D object point $\mathbf{P}$ is at (X_t, Y_t, Z_t). If we use the camera coordinate of the reference camera as the world coordinate, then the projection position of point $\mathbf{P}$ on the image plane of camera C_0 at frame t is

$$x_t = \frac{X_t f}{Z_t}, \quad y_t = \frac{Y_t f}{Z_t}. \tag{19}$$

Now we use a two camera set up as an example to show how to combine optical flow constraints from different cameras. Suppose cameras C_0 and C_1 form a standard set up and have the

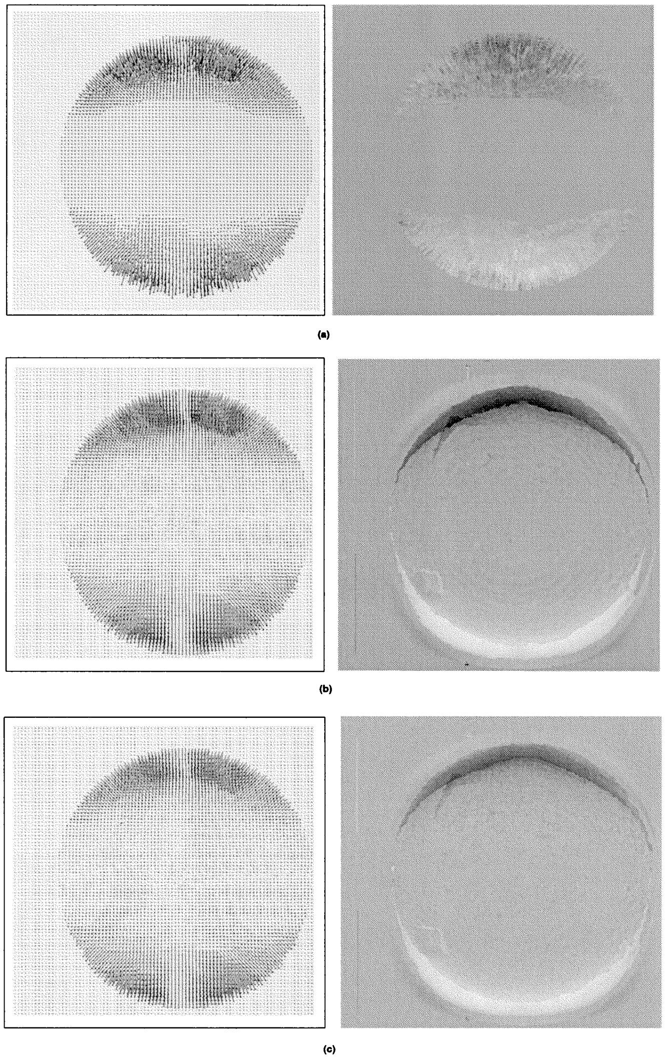


Fig. 7. Recovered 3-D scene flow of the synthetic sequence. Left images show projected 3-D scene flow. Right images show 3-D scene flow along z direction: darker means moving away from the camera; brighter means moving toward the camera. (a), (b), and (c) are the results produced by IMS, EGS-F1, and EGS-F2, respectively.

same focal length f and the base line b is along X axis. If at frame t, 3-D point $\mathbf{P}$ is projected at (x_t, y_t) in the reference view, then the projection position (x'_t, y'_t) of $\mathbf{P}$ on camera C_1 is

$$x'_t = x_t + \frac{bf}{Z_t}, \quad y'_t = y_t. \tag{20}$$

At frame $t+1$, point $\mathbf{P}$ is projected at $(x_t + u, y_t + v)$ on camera C_0, and the projection position (x'_{t+1}, y'_{t+1}) of $\mathbf{P}$ on camera C_1 is

$$x'_{t+1} = x_t + u + \frac{bf}{Z_{t+1}}, \quad y'_{t+1} = y_t + v. \tag{21}$$

Since $d_t = (bf)/(Z_t)$ and $w = d_{t+1} - d_t$, the optical flow (u', v') of point $\mathbf{P}$ on camera C_1 is

$$u' = u + \frac{bf}{Z_{t+1}} - \frac{bf}{Z_t} = u + w, \quad v' = v. \tag{22}$$

So, the combined motion constraint of camera C_0 and C_1 can be represented as

$$\begin{aligned}\mathcal{E}_m = &\left(\left.\frac{\partial I_0}{\partial x}\right|_{x,y} u + \left.\frac{\partial I_0}{\partial y}\right|_{x,y} v + \left.\frac{\partial I_0}{\partial t}\right|_{x,y} \right)^2 \\ &+ \kappa \left(\left.\frac{\partial I_1}{\partial x}\right|_{x+d,y} (u+w) \right. \\ &\left. + \left.\frac{\partial I_1}{\partial y}\right|_{x+d,y} v + \left.\frac{\partial I_1}{\partial t}\right|_{x+d,y} \right)^2\end{aligned} \tag{23}$$

where κ is the confidence measurement of disparity (obtained from the stereo matching algorithm) if $\mathbf{P}$ is visible in camera C_1, otherwise it is 0.

It is straightforward to extend this combination to situations where more than two cameras are utilized. Ideally, the motion constraints should combine all the optical flow constraints from all the cameras. However, under a multiview setup, occlusion is almost unavoidable. Only the optical flow constraints from some of the cameras are usable. Note that if the object point $\mathbf{P}$ is not visible in other cameras except the reference camera, the combined motion constraint degrades to normal optical flow constraint.

D. Hard Constraints

In initial stereo matching, we can easily add hard constraints during interpolation by setting large weights for the points with high confidence measurement. This makes the disparity map more accurate. Similarly, hard constraints can be added in the temporal domain. We perform correlation twice on two consecutive frames. Suppose P is a point in the reference frame t. We first search the correspondence of point P in frame $t+1$ using correlation. The search area is within a window delimited by the possible maximum motion. We then exchange the roles of the two frames and find the correspondence again. If at the corresponding points we have the same motion measurement in both cases, we consider the 2-D motion (u_h, v_h) at the image point as valid motion. The hard constraint at a point in temporal domain can be represented as

$$\mathcal{E}_h = \mu c((u - u_h)^2 + (v - v_h)^2), \tag{24}$$

where (u_h, v_h) is the valid motion found by cross-validation, and c is a large constant. μ is the normalized matching score while searching for motion correspondence if valid motion has been measured, otherwise it is 0.

E. Formulation 1: 3-D Scene Flow Estimation

If we assume that the initial disparity map is accurate enough, we can formulate the problem as: given N image sequences captured by N different cameras and an accurate initial disparity map, compute 3-D scene flow (u, v, w) at every point in the reference image. We denote this formulation as EGS-F1. Combining the motion constraints and hard constraints, we have the following energy function

$$\mathcal{E}_1 = \iint (\mathcal{E}_m + \mathcal{E}_h + \mathcal{E}_s)\, dx\, dy \tag{25}$$

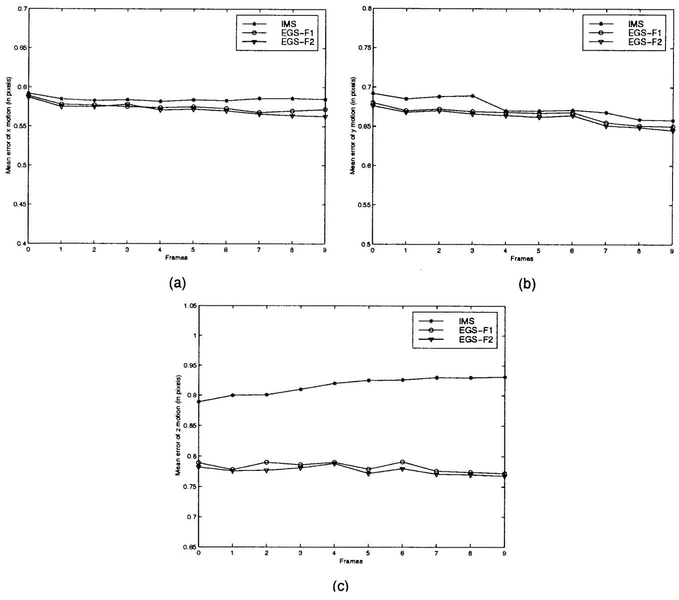


Fig. 8. Three-dimensional scene flow recovery evaluation. (a) Mean error of x motion. (b) Mean error of y motion; (c) Mean error of z motion.

where $\mathcal{E}_m$ and $\mathcal{E}_h$ are defined in previous sections. $\mathcal{E}_s$ is the smoothness term defined as

$$\mathcal{E}_s = \gamma \left(\left(\frac{\partial u}{\partial x} \right)^2 + \left(\frac{\partial u}{\partial y} \right)^2 + \left(\frac{\partial v}{\partial x} \right)^2 + \left(\frac{\partial v}{\partial y} \right)^2 + \left(\frac{\partial w}{\partial x} \right)^2 + \left(\frac{\partial w}{\partial y} \right)^2 \right). \quad (26)$$

In our algorithm, we want to minimize the above energy function within individual image segments to enforce motion boundaries. However, the recovery of segmented or piecewise smooth flow field is notably difficult [45]. Under a multiview setup, we may have more constraints than what we have by using only one camera. But we also have more unknowns (z motion) to solve. There are even more difficulties when a segment in the reference image is invisible in a lot of other cameras. In an extreme case, a segment in the reference camera is invisible in all the other cameras. This means we may have only two constraints (one optical flow constraint and one smoothness constraint) at a point in this segment in order to solve three unknowns (u, v, w). Mathematically speaking, w is not well defined by these constraints and the algorithm may converge incorrectly. Thus we need more information from the neighborhood to propagate into these points.

To deal with the occluded regions, we adopt a multi-resolution strategy to minimize the energy. First, we minimize the energy function on the entire image. This forces more neighborhood information to propagate toward the occluded points. Second, we use the results from the first step as initial values to minimize energy function within each segment. Obviously, for segments that are totally occluded in all the other views, the second step is not necessary. Experiments (Fig. 11) show that this strategy can still maintain reasonable motion boundaries.

As suggested in [1] and [46], an iterative relaxation method or conjugate gradient method may be used to minimize (25). In our experiments, conjugate gradient search is used. It is also clear that the energy function is very similar to gradient-based optical flow energy function. In fact, this formulation can be thought of as a natural extension of optical flow computation under a multiview setup. The difference is that all the optical flow constraints from different views contribute to the minimization wherever possible. Also, newly added hard constraints makes the algorithm more stable and accurate.

F. Formulation 2: Integrated 3-D Scene Flow and Structure Estimation

The previous formulation assumes that we have already got very accurate disparity map in initial stereo matching. However, initial stereo matching may be inaccurate and noisy because stereo matching is essentially an under-constrained problem due to occlusion, lack of texture, etc. Furthermore, as we discussed before, stereo matching algorithm normally ignores temporal information. It is reasonable to think that by considering motion constraints, we may get better disparity map. This means we formulate the problem as computing a four dimensional vector (u, v, w, d) at every point on the reference image, where the initial disparity is used as an initial guess. We denote this formulation as EGS-F1. However, with serious occlusion and limited number of cameras, this formulation is even more difficult because we now need to solve for four unknowns at every point. We need at least four independent constraints to make the algorithm stable. This means if we use (25), at least three cameras should be used for this formulation so that we can have three optical flow constraints, one smoothness constraint, and maybe one hard constraint. Thus new constraints in addition to those used in (25) are desired.

Considering the correlation volume $\mathcal{C}(x, y, d)$ we computed from initial stereo matching, a new constraint on disparity can be established. For image point (x_p, y_p), by using planes $x = x_p$ and $y = y_p$ to carve the correlation volume, we can get a one dimensional function $\mathcal{C}_{x_p, y_p}(d)$. Obviously the disparity should maximize the value of $\mathcal{C}_{x_p, y_p}(d)$. Thus another energy term, stereo constraint, can be defined as

$$\varepsilon_c = -\tau \mathcal{C}_{x,y}(d) \quad (27)$$

where τ is a positive constant if the point is not occluded in the corresponding camera; otherwise it is 0. Also, if we have high confidence measurement for initial disparity map, we can add another energy term

$$\varepsilon_i = \zeta (d - d_i)^2 \quad (28)$$

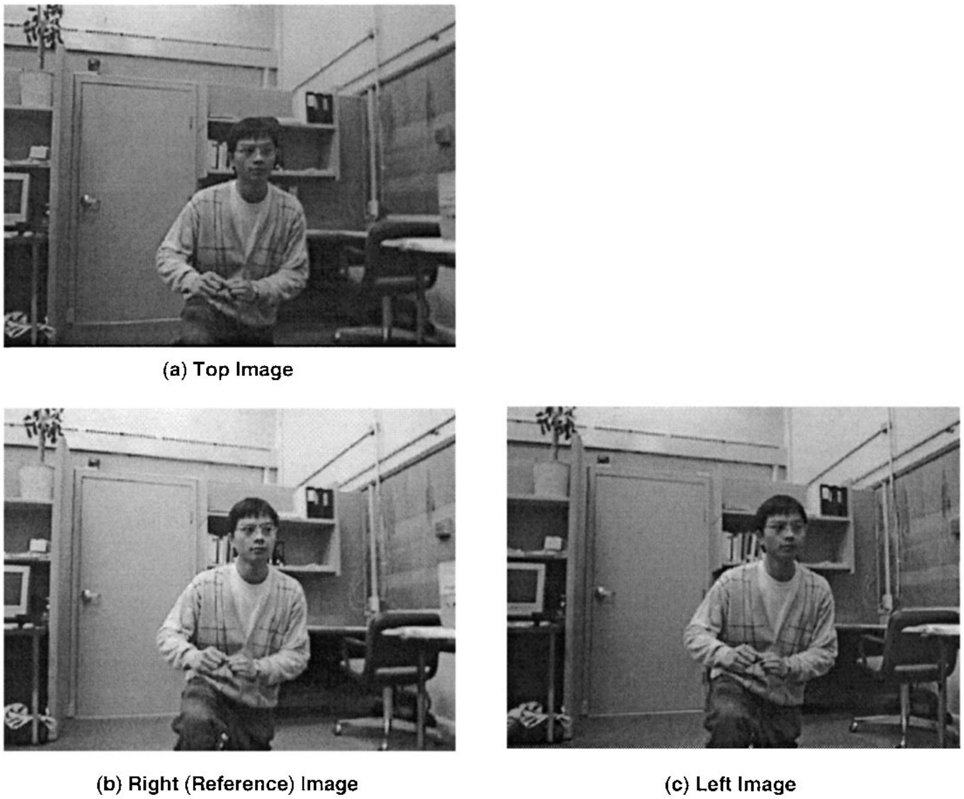
where d_i is the initial disparity and ζ is its confidence measurement.

Thus, the new energy function can be defined as

$$\mathcal{E}_2 = \iint (\mathcal{E}_m + \mathcal{E}_h + \mathcal{E}_s + \mathcal{E}_c + \mathcal{E}_i) \, dx \, dy. \quad (29)$$

Another change of energy function in this formulation is that the smoothness term ε_s should include the smoothness measurement of disparity defined as

$$\left(\frac{\partial d}{\partial x} \right)^2 + \left(\frac{\partial d}{\partial y} \right)^2. \quad (30)$$


(a) Top Image

(b) Right (Reference) Image

(c) Left Image

Fig. 9. Snapshot of images acquired by Triclops.

Again, we can use the multiresolution strategy discussed in Section II-E to minimize the energy function in each image segment.

In both formulations, to take advantage of the state of the art in optical flow estimation, we can first compute the optical flow in the reference view, then use the results as the initial guess while minimizing the above energy functions. This makes the algorithm converge faster and be more robust.

IV. Experiments

To demonstrate the effectiveness of our systems we have applied them to both synthetic and real sequences. To show the advantages of our rule-based stereo matching algorithm, we employed several different stereo matching algorithms in our systems to compute the initial disparity maps.

A. Synthetic Scene

We used *O*penInventor to generate a two-view image sequence (ten frames) consisting of a deformable sphere. Since the ground truth is available, we can quantitatively evaluate the systems. Fig. 6(a) shows the generated synthetic input sequence: The upper part of the sphere moves down and away from the camera (shrinks), while the lower part of the sphere moves down and toward the camera (expands). It is also worth mentioning that in order to test the capability of our systems in maintaining the details of the object, some artificial pits and bumps are added to the sphere surface.

The recovered structure computed by IMS is shown in Fig. 6(b). To emphasize the recovered structure and motion of the object of interest (the deformable sphere), we only show the results in the object area for this synthetic sequence (i.e., the background, which contains no information, is discarded). The initial depth guess in IMS for frame 0 was computed by using the stereo matching algorithm proposed in [14]. As can be seen, most part of the structure is correctly recovered. However, the artificial pits and bumps are not clear. Fig. 6(c) and (d) illustrate the results computed by using two formulations of EGS, respectively. The initial disparity maps for both formulations of EGS were computed by using the rule-based stereo algorithm proposed in Section III.B. Since EGS-F1 assumes that the initial disparity maps are accurate, from Fig. 6(c), we can immediately notice that the artificial pits and bumps are clearly preserved in the recovered structure of our rule-based stereo matching algorithm. These results demonstrate the advantages of our stereo algorithm: it produces smooth disparity maps and clearly maintains the details (discontinuities) of the scene. Examining the results closely, we can also notice that EGS-F2 produces slightly better scene structure than what EGS-F1 does (e.g., the top part of the sphere is lost in EGS-F1-S-5, while preserved in EGS-F2-S-5. Data in Table I also suggest this.). This means that by incorporating the initial disparities into the motion minimization framework, we can actually refine the initial scene structure. This refinement demonstrates that motion constraints help structure estimation in EGS-F2.

Table I shows the quantitative evaluation of the recovered structure of our systems on each frame. We deem that a disparity is correctly recovered if the difference between the estimated

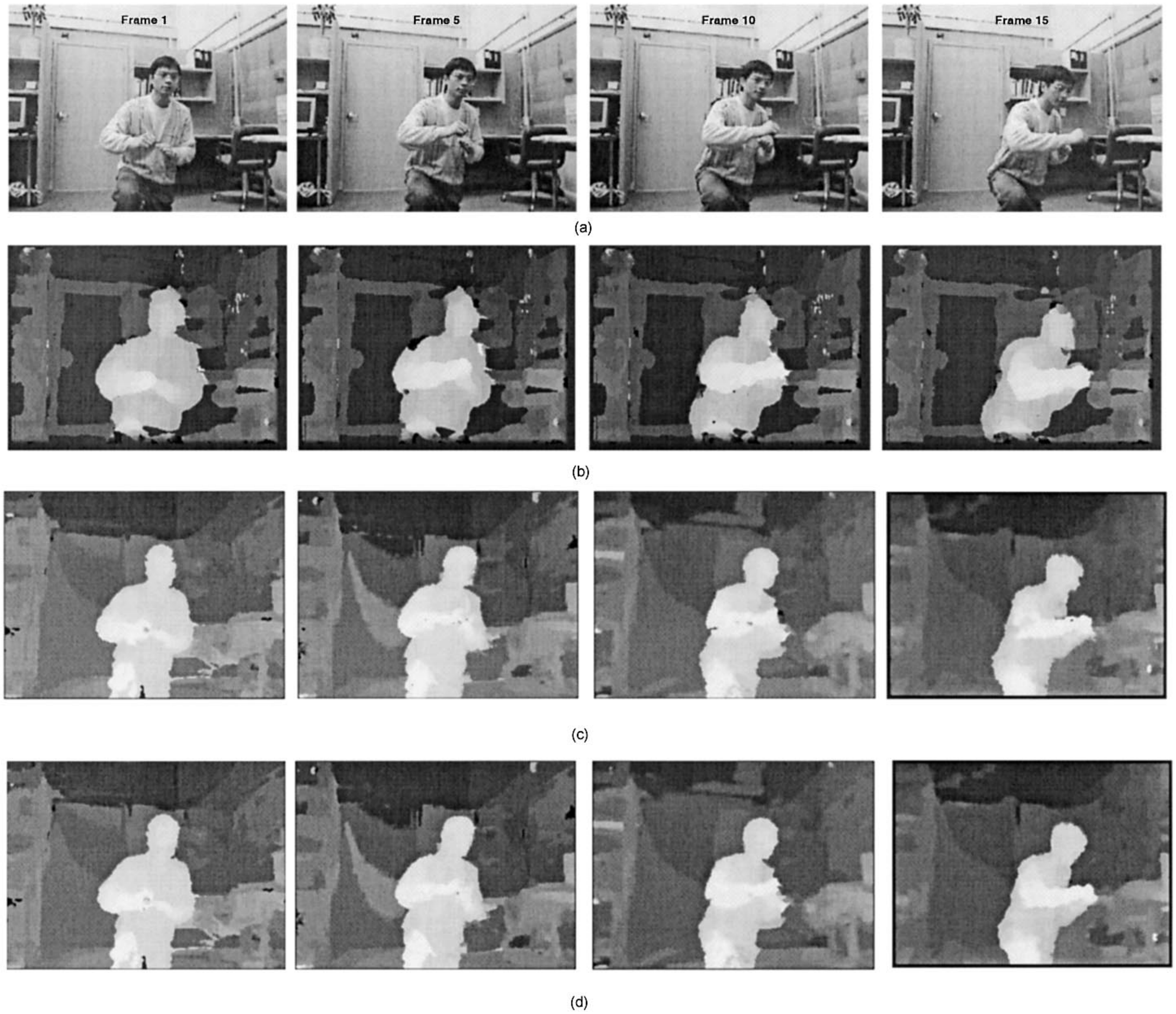


Fig. 10. Structure results on real image sequences. (a) Image sequence gathered with reference camera (one of the 3-view sequences); (b), (c), and (d) illustrate the recovered scene structure by IMS, EGS-F1, and EGS-F2, respectively.

disparity and the ground truth is less than 0.5 pixel. Then the percentage of disparities recovered correctly is used as a performance evaluation criterion. As can be seen from the table, disparities recovered by either IMS or EGS are at least 92% correct. Note that there is considerable improvement of results by using our rule-based stereo matching algorithm, over the initialization performed using the algorithm described in [14]. The quantitative results show that EGS-F2 achieves the best structure results.

The recovered 3-D scene flow results computed by IMS, EGS-F1, and EGS-F2 are illustrated in Fig. 7(a)–(c), respectively. Fig. 8(a), (b) and (c) show the quantitative evaluation of the recovered 3-D scene flow (x, y, and z components, respectively) on each frame. Qualitatively speaking, both IMS and EGS-F1 tracked most parts of the sphere correctly. All the results indicate that the upper part of the sphere moves down and away from the camera, while the lower part moves down and toward the camera. For x and y components, both IMS and EGS produce comparable results. However, it is clear to see that EGS tracked the z motion more accurately and EGS-F2 again achieves the best motion results.

According to the experiment, both IMS and EGS are able to recover 3-D scene flow and structure correctly on most parts of the synthetic sequence. However, EGS generally produces better results (motion and structure) than IMS, both quantitatively and qualitatively. The main reason is that EGS utilizes our rule-based stereo matching algorithm which produces more accurate initial disparity matches and is able to report the occluded areas. Between EGS-F1 and EGS-F2, we can see that EGS-F2 achieves more accurate results by incorporating initial disparity into the motion minimization framework.

B. Real Scene

In order to test and evaluate our systems in practice, we have performed experiments with real scene sequences. We acquired real image sequences with Point Grey Research Inc. Triclops system: a three-eye stereo camera connected with Matrox MeteorIIMC real time video capture card. This device provides us real-time (around 15 frame/sec) rectified image sequences and camera calibration parameters. To test the robustness of our

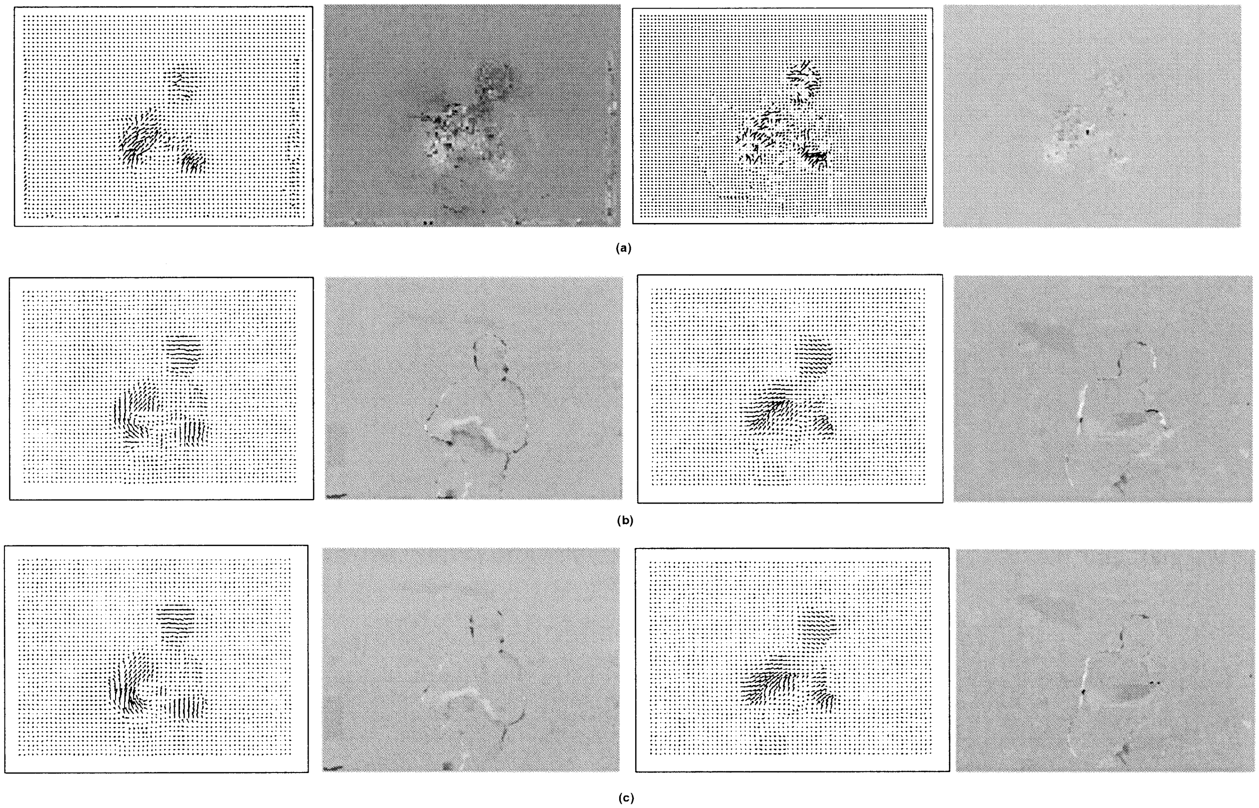


Fig. 11. Three-dimensional scene flow at frames 1 and 10 recovered by (a) IMS, (b) EGS-F1, and (c) EGS-F2. Needle graphs represent the 2-D projection, and intensity maps represent the z velocity.

system, the image sequences were captured under poor illumination conditions (thus our initial stereo correspondences and optical flow were noisy). Fig. 9 shows a snapshot captured by Triclops.

The three-view image sequence used in our experiment is illustrated in Fig. 10(a) (only the images captured by the reference camera are shown). Fig. 10(b) shows the recovered scene structure by IMS. The initial disparity maps are computed by using a simple direct method (window based matching) provided by the Triclops SDK. It is obvious that the direct method has the well known "foreground fattening" effect, i.e., the recovered structure of the person in the scene is fattened due to the nonhorizontal depth discontinuities. This not only makes the recovered structure inaccurate, but also makes the recovered z motion noisy. Fig. 10(c) and (d) show the recovered scene structure by EGS-F1 and EGS-F2, respectively. The initial disparity maps were computed by using our rule-based stereo matching algorithm. We can easily see that the fattening effect is heavily reduced and the depth discontinuities are clearly preserved. As we expect, EGS-F2 still performs slightly better compared with EGS-F1. For example, on frame 1, some mismatches under the right hand of the person and close to the left side of the face are refined. On frame 10, some mismatches close to the right fist of the person are corrected. It is interesting to notice that all these refined mismatches happen in the moving parts of the person. Especially, most of them are close to the intensity boundaries. At those places, motion constraints are much more reliable due to additional texture information. This further suggests that in practice, stereo and motion constraints do help each other. Fig. 11(a)–(c) illustrate the 3-D scene flow at frame 1 and 10 recovered by IMS, EGS-F1, and EGS-F2, respectively. The moving parts in the scene (such as the arms of the subject) were successfully tracked by both IMA and EGS. Also, the recovered structure at the moving parts preserved the correct shape (e.g., the arm can be distinguished). However, it can be seen that EGS tracked the motion more accurately. Results recovered by IMS become noisy at frame 10 due to inaccurate disparity and error accumulation. Close examination also shows that EGS-F2 still achieves the best results, showing evidence for the stability of a tightly coupled stereo and motion constraints based system.

V. Conclusion

There are numerous potential applications for 3-D scene flow and structure such as robust scene structure recovery, dynamic scene interpretation, dynamic rendering, etc. In this paper, we have presented two automatic systems, IMS and EGS, for 3-D scene flow and structure recovery. IMS utilizes 3-D affine motion model fitting techniques to simultaneously compute 3-D motion and structure. EGS extends traditional two dimensional gradient-based optical flow techniques to three dimensional case. In doing so, two different formulations of EGS are developed. We have also noticed that it is critical to have an accurate initial structure estimation and occlusion detection when fusing motion and stereo. We therefore designed a novel rule-based stereo matching algorithm. This algorithm produces accurate disparity maps. The fattening effect is heavily reduced and the depth discontinuities are well preserved. Also, the occluded areas are explicitly reported. Image segmentation information is also used to preserve the motion and depth discontinuities.

In our systems, we integrated 2-D motion and stereo constraints tightly to compute 3-D motion and structure. We have shown that motion and stereo benefit from each other in both ideal (noise-free) and real situations. This makes 3-D motion and structure analyses more stable. We conducted experiments on both synthetic and real imagery. Promising results can be seen on both synthetic and real data. The empirical comparison has suggested that EGS-F2 produces the most accurate results. This is due to the concerted efforts of rule-based stereo algorithm and tight integration of motion/stereo constraints.

REFERENCES

[1] B. Horn and B. Schunck, "Determining optical flow," *Artif. Intell.*, vol. 17, pp. 185–203, 1981.

[2] P. Anandan, "A computational framework and an algorithm for the measurement of visual motion," *Int. J. Comput. Vis.*, vol. 2, no. 3, pp. 283–310, Jan. 1989.

[3] M. Black and P. Anandan, "The robust estimation of multiple motions: Parametric and piecewise-smooth flow-fields," *Comput. Vis. Image Understand.*, vol. 63, no. 1, pp. 75–104, Jan. 1996.

[4] J. Barron, D. Fleet, and S. Beauchemin, "Performance of optical flow techniques," *Int. J. Comput. Vis.*, vol. 12, no. 1, pp. 43–77, Feb. 1994.

[5] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade, "Three-dimensional scene flow," *Proc. IEEE Computer Soc. Int. Conf. Computer Vision*, pp. 722–729, 1999.

[6] Y. Zhang and C. Kambhamettu, "Integrated 3-D scene flow and structure recovery from multiview image sequences," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. II:674–II:681, 2000.

[7] ——, "On 3-D scene flow and structure estimation," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. II:778–II:785, 2001.

[8] A. Pentland and B. Horowitz, "Recovery of nonrigid motion and structure," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 13, pp. 730–742, July 1991.

[9] M. Black, "Explaining optical flow events with parameterized spatio-temporal models," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. I:326–I:332, 1999.

[10] L. Zhou, C. Kambhamettu, and D. Goldgof, "Extracting nonrigid motion and 3-D structure of hurricanes from satellite image sequences without correspondences," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. II:280–II:285, 1999.

[11] M. Penna, "The incremental approximation of nonrigid motion," *Comput. Vis., Graph. Image Process.*, vol. 60, no. 2, pp. 141–156, Sept. 1994.

[12] U. Dhond and J. Aggarwal, "Structure from stereo: A review," *IEEE Trans. Syst., Man, Cybern.*, vol. 19, pp. 1489–1510, Nov. 1989.

[13] D. Scharstein and R. Szeliski, "A taxonomy and evaluation of dense two-frame stereo correspondence algorithms," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 7–42, Apr. 2002.

[14] S. Roy and I. Cox, "A maximum-flow formulation of the n-camera stereo correspondence problem," *Proc. IEEE Computer Soc. Int. Conf. Computer Vision*, pp. 492–499, 1998.

[15] W. Richards, "Structure from stereo and motion," *J. Op. Soc. Ameri. A*, vol. 2, no. 2, pp. 343–349, Feb. 1985.

[16] D. Ballard and O. Kimball, "Rigid body motion from depth and optic flow," *Comput. Vis., Graph. Image Process.*, vol. 22, no. 1, pp. 95–115, Apr. 1983.

[17] T. Huang and S. Blostein, "Robust algorithms for motion estimation based on two sequential stereo image pairs," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 518–523, 1985.

[18] K. Mutch, "Determining object translation information using stereoscopic motion," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. PAMI-8, pp. 750–755, Nov. 1986.

[19] G. Young and R. Chellappa, "3-D motion estimation using a sequence of noisy stereo images: Models, estimation, and uniqueness results," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 735–759, Aug. 1990.

[20] P. Balasubramanyam, "Computation of motion in depth parameters: A first step in stereoscopic motion interpretation," in *Proc. Image Understanding Workshop*, 1988, pp. 907–920.

[21] A. Waxman and J. Duncan, "Binocular image flows: Steps toward stereo-motion fusion," *IEEE Trans. Pattern Anal. Machine Intell.* , vol. PAMI-8, pp. 715–729, Nov. 1986.

[22] J. Barron, A. Jepson, and J. Tsotsos, "Determination of egomotion and environmental layout from noisy time-varying velocity in binocular image sequences," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1987, pp. 822–825.

[23] Y. Aloimonos and A. Basu, "Shape and 3-D motion from contour without point to point correspondences: General principles," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 518–527, 1986.

[24] F. Dornaika and R. Chung, "Stereo correspondence from motion correspondence," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. I:70–I:75, 1999.

[25] L. Li and J. Duncan, "3-D translational motion and structure from binocular image flows," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 657–667, July 1993.

[26] J. Weng, N. Ahuja, and T. Huang, "Optimal motion and structure estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, pp. 864–884, Sept. 1993.

[27] W. Liao, S. Aggarwal, and J. Aggarwal, "The reconstruction of dynamic 3-D structure of biological objects using stereo microscope images," *Mach. Vis. Applicat.*, vol. 9, no. 4, pp. 166–178, 1997.

[28] S. Malassiotis and M. Strintzis, "Model-based joint motion and structure estimation from stereo images," *Comput. Vis. Image Understand.*, vol. 65, no. 1, pp. 79–94, Jan. 1997.

[29] C. Kambhamettu, K. Palaniappan, and A. Hasler, "Coupled, multi-resolution stereo and motion analysis," *Proc. IEEE Symp. Computer Vision*, pp. 43–48, 1995.

[30] J. Neumann and Y. Aloimonos, "Spatio-temporal stereo using multi-resolution subdivision surfaces," *Int. J. Comput. Vis.*, vol. 47, no. 1–3, pp. 181–193, Apr. 2002.

[31] I. Cox, S. Hingorani, S. Rao, and B. Maggs, "A maximum-likelihood stereo algorithm," *Comput. Vis. Image Understand.*, vol. 63, no. 3, pp. 542–567, May 1996.

[32] *Handbook of Pattern Recognition and Image Processing: Computer Vision*, Academic, San Diego, 1994, pp. 405–430. Nonrigid motion analysis.

[33] J. Bergen, P. Anandan, K. Hanna, and R. Hingorani, "Hierarchical model-based motion estimation," in *Proc. Eur. Conf. Computer Vision*, 1992, pp. 237–252.

[34] S. Ju, M. Black, and A. Jepson, "Skin and bones: Multi-layer, locally affine, optical flow and regularization with transparency," *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition*, pp. 307–314, 1996.

[35] H. Li, P. Roivainen, and R. Forchheimer, "3-D motion estimation in model-based facial image coding," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 15, no. 6, pp. 545–555, June 1993.

[36] L. Zhou, C. Kambhamettu, D. Goldgof, K. Palaniappan, and A. Hasler, "Tracking nonrigid motion and structure from 2-D satellite cloud images without correspondences," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 23, pp. 1330–1336, Nov. 2001.

[37] L. Zhou and C. Kambhamettu, "Hierarchical structure and nonrigid motion recovery from 2-D monocular views," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. II:752–II:759, 2000.

[38] Y. Huang, K. Palaniappan, X. Zhuang, and J. Cavanaugh, "Optic flow field segmentation and motion estimation using a robust genetic partitioning algorithm," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 17, pp. 1177–1190, Dec. 1995.

[39] R. Balasubramanian, D. Goldgof, and C. Kambhamettu, "Tracking of nonrigid motion and 3–D structure from 2-D image sequences without correspondences," in *Proc. Int. Conf. Image Processing*, 1998, pp. I:933–I:937.

[40] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling, *Numerical Recipes in C*. Cambridge, U.K.: Cambridge Univ. Press, 1988.

[41] S. Ghosal and P. Vanek, "A fast scalable algorithm for discontinuous optical-flow estimation," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 181–194, Feb. 1996.

[42] P. Felzenszwalb and D. Huttenlocher, "Image segmentation using local variation," *Proc. IEEE Conf. Computer Vision Pattern Recognition*, pp. 98–104, 1998.

[43] P. Fua, "Combining stereo and monocular information to compute dense depth maps that preserve depth discontinuities," in *Proc. Int. Joint Conf. Artificial Intelligence*, 1991, pp. 1292–1298.

[44] H. Tao, H. Sawhney, and R. Kumar, "A global matching framework for stereo computation," *Proc. IEEE Computer Soc. Int. Conf. Computer Vision*, pp. I:532–I:539, 2001.

[45] M. Black and A. Jepson, "Estimating optical-flow in segmented images using variable-order parametric models with local deformations," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 18, pp. 972–986, Oct. 1996.

[46] R. Szeliski, "Fast surface interpolation using hierarchical basis functions," *IEEE Trans. Pattern Anal. Machine Intell.*, vol. 12, pp. 513–528, June 1990.

Ye Zhang received the B.Eng. and the M.Eng. degrees in electrical and electronics engineering from Sichuan University, Sichuan, China, in 1995 and 1998, respectively, He also received the M.S. and the Ph.D. degrees in computer and information sciences from University of Delaware, Newark, in 2000 and 2002, respectively.

He is currently Chief Scientist and Vice President of Development at Ensuredmail, Inc., Wilmington, DE. His research interests include computer vision (especially nonrigid motion analysis, stereo matching, and facial gesture analysis), computer graphics and image processing.

Dr. Zhang is the winner of the "2002 Frank Pearson Achievement Award," the "2001 Quantum Leap Innovations Excellence Award in Artificial Intelligence," and the "2000–2001 University Competitive Fellowship Award" from the University of Delaware.

Chandra Kambhamettu (M'01) received the B.E. degree in computer science and engineering from Osmania University, Osmania, India, in 1989, the M.S. and Ph.D. degrees in computer science and engineering from the University of South Florida, Tampa, in 1991 and 1994 respectively.

He is an Associate Professor in the Department of Computer and Information Sciences, University of Delaware, Newark where he leads the Video/Image Modeling and Synthesis (VIMS) group. From 1997 to 2003, he was an Assistant Professor in the same department. From 1994 to 1996, he was a Research Scientist at NASA Goddard, Greenbelt, MD. He is an Associate Editor for the Journal of Pattern Recognition. His research interests include computer vision, biomedical image analysis, bioinformatics, computer graphics, and multimedia systems.

Dr. Kambhamettu received the "Excellence in Research Award" from Universities Space Research Association (USRA), as well as the NSF CAREER award in 2000.