Integrated 3D Scene Flow and Structure Recovery from Multiview Image Sequences

Ye Zhang and Chandra Kambhamettu
Video / Image Modeling and Synthesis Lab
Department of Computer & Information Sciences
University of Delaware
Newark, Delaware 19716
zhangye/chandra@cis.udel.edu

Abstract

Scene flow is the 3D motion field of points in the world. Given N (N > 1) image sequences gathered with a N-eye stereo camera or N calibrated cameras, we present a novel system which integrates 3D scene flow and structure recovery in order to complement each other's performance. We do not assume rigidity of the scene motion, thus allowing for non-rigid motion in the scene. In our work, images are segmented into small regions. We assume that each small region is undergoing similar motion, represented by a 3D affine model. Non-linear motion model fitting based on both optical flow constraints and stereo constraints is then carried over each image region in order to simultaneously estimate 3D motion correspondences and structure. To ensure the robustness, several regularization constraints are also introduced. A recursive algorithm is designed to incorporate the local and regularization constraints. Experimental results on both synthetic and real data demonstrate the effectiveness of our integrated 3D motion and structure analysis scheme.

1 Introduction

Optical flow is a 2D motion field in the image plane. By analogy, Vedula *et al.* [28] used the term *scene flow* to represent a dense 3D motion vector field defined for every point on every surface in the scene. 3D Scene flow has numerous potential applications such as scene structure prediction, dynamic rendering, automatic navigation, and interpretation tasks. In this paper, given N ($N > 1$) image sequences from N different viewpoints, we present a system which simultaneously computes 3D scene flow and structure in a mutually beneficial way. We do not assume any *a priori* knowledge of the dynamic scene, nor do we assume that the scene motion is rigid.

1.1 Related Work

Motion and structure recovery are fundamental problems in computational vision. There has been considerable interest in recovering 3D motion and structure from monocular view image sequences (*e.g.* [24, 7, 31, 23]). 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 image sequences (e.g. [10, 27]). However, 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 solving non-linear equations. Thus the computation is very sensitive to noise. Moreover, the 3D 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[26] described the defects in stereo and motion parallax (i.e., structure from motion) respectively and integrated them to recover 3D rigid shape. In his method the only goal was to recover 3D structure. Motion analysis didn't benefit from stereo analysis. Ballard *et al.* [4], Huang *et al.* [13], Mutch *et al.* [22], and Balasubramanyam *et al.* [3] simply computed the rigid motion parameters assuming the depth was known, or had been computed by stereo vision. Waxman *et al.* [29] used the difference between the flow fields of the left and right cameras to analyze

1063-6919/00 $10.00 © 2000 IEEE

the cases with unknown motion and structure. But their method assumed that the viewed surfaces were planar. Barron *et al.* [5] developed a relation between binocular velocity fields and the motion/structure parameters. A non-linear method was then presented to simultaneously compute the motion and structure. Aloimonos *et al.* [1] used two cameras to recover surface structure, then utilized the positions of feature points in the stereo image pairs to decide the direction of translation. More recently, Dornaika *et al.* [11] recovered the stereo correspondence using one motion of a stereo rig. Like in [26], their approach did not refine and improve motion analysis through coupling stereo and motion. Li *et al.* [19] 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 that was present in [29]. However, the 3D motion was still restricted to translational motion. Weng *et al.* [30] designed another two-step approach. A linear algorithm was first used for a preliminary estimate of rigid motion parameters, then an optimal objective function was 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 non-rigid motion is imperative and needs special attention in motion analysis. There has been very limited research on integration of non-rigid motion and stereo. Liao *et al.* [20] used a relation-based algorithm to cooperatively match features in both temporal and spatial domains. It therefore does not provide dense motion. Malassiotis *et al.* [21] used a grid deformable model to generalize the monocular approaches. However, model-based approach requires *a priori* knowledge of the scene. Kambhamettu *et al.* [17] coupled stereo and non-rigid 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.* [28] defined and computed the non-rigid 3D 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 matching measurements were not fully utilized in the scene structure recovery.

Another 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, 3D motion and stereo analyses were carried in different modules. The integration was not essentially coupled and was more like post-processing.

1.2 Our Approach

In this paper, our goal is similar to that of [28], i.e, recovering 3D scene flow and dense scene structure from multiview image sequences. No *a priori* knowledge of the scene is assumed. Also, we do not assume that the scene is rigid. Our formulation is different from [28] in that we solve the problem using non-linear model fitting. Contributions of our work include:

1. formulation to simultaneously recover 3D motion and structure;
2. seamless integration of 2D motion and stereo constraints;
3. complete and automatic system computing 3D scene flow and dense scene structure;

In our approach, the images are first segmented into small regions. We assume that each region is undergoing similar motion which can be represented by a 3D affine model. Non-linear least square method is used to fit the motion model for each small region. Several regularization constraints are also used to ensure the robustness. A recursive algorithm is then presented to incorporate all the constraints. The rest of this paper is organized as follows: Section 2 describes our system that integrates the dual problem of 3D motion and stereo analyses. Multiple camera geometry is briefly discussed in Section 2.1. 3D affine motion model and local model fitting are presented in Section 2.2 and Section 2.3, respectively. Regularization constraints are introduced in Section 2.5. Our complete recursive algorithm is presented in Section 2.6. Section 3 shows the experimental results and the validation of our system. Section 4 concludes and addresses our future work.

2 Integrated System

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

2.1 Multiple Camera Geometry

Several multiple camera algorithms for stereo analysis have been proposed in the past[27, 9, 29, 16]. In our system, we utilize multiple cameras in a manner similar to [9, 27]. A pair of cameras are used as a *reference* or *basic stereo pair*. Other cameras provide extra information, thus contributing additional constraints.

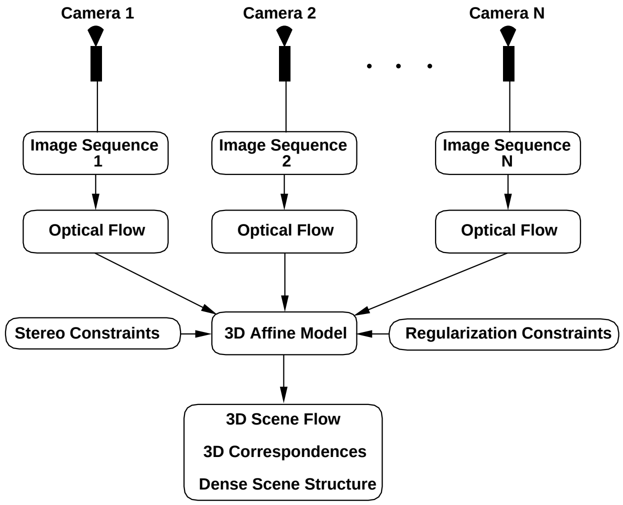

Figure 1: Block diagram of the system

At a given instance, a set of N cameras $C_0, C_1, ...C_{n-1}$ provide N images $I_0, I_1, ...I_{n-1}$, 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 3D scene flow and disparity map for each image point. A 3D 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 back-project $\mathbf{m}$ to a 3D 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 $C_1, C_2, ...C_{n-1}$ which is represented by $\mathbf{R}$,

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

2.2 Local Motion Model Selection

In order to describe 3D motion without rigidity assumption, it is important to choose a motion model powerful enough to describe different kinds of non-rigid motion [16]. There have been many works which use 2D affine motion model for image matching [6]. More recently, Ju *et al.* [15] and Bergen [6] have used 2D affine model to estimate image motion. Li *et al.* [18] and Zhou *et al.* [31] have used 3D affine model to analyze face and cloud non-rigid motion respectively, indicating the use of affine model in describing complex non-rigid motion. In our work, we utilize 3D affine model to describe the underlying non-rigid motion in the scene.

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

$$\mathbf{P_m}^{t+1} = \mathbf{M}^t \mathbf{P_m}^t \tag{4}$$

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 3D affine model is that it provides a simple way to combine non-rigid 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 Eq. 4 on the whole image in order to estimate 3D motion. This is because $\mathbf{M}^t$ is, in general, point dependent during non-rigid motion. However, in practice, motion field is spatially smooth. Thus if we apply affine model (Eq. 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 it is still possible to segment the images according to optical flow information (*e.g.* [14]). Non-rigid 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 non-linear model fitting. Zhou *et al.* [31] 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,

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

Eq. 6 reduces the number of unknowns in successive frames for each small region, thus improving the robustness of non-linear fitting.

2.3 Motion Model Fitting

Balasubramanian *et al.* [2] and Zhou *et al.* [31] 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) [25] 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*, i.e., 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 Eq 4 for each region. The unknown vector is represented by,

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

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

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

where $\mathbf{U}_t^*$ is the optimal unknown vector and $EOF(\mathbf{U}_t)$ 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.

2.3.1 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 [8] 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 back-projected to a 3D point ${\mathbf{P}_\mathbf{m}}^t$ in the scene by Eq. 2. Clearly, from frame t to frame $t+1$, the 2D motion of the projective point of ${\mathbf{P}_\mathbf{m}}^t$ on the image plane of camera j can be computed as,

$$\mathbf{V}_j(\mathbf{m}, t) = \mathbf{H}(\mathbf{T}_j(\mathbf{M}^t{\mathbf{P}_\mathbf{m}}^t - {\mathbf{P}_\mathbf{m}}^t)) \tag{9}$$

where,

$$\mathbf{H}(\begin{pmatrix} x \\ y \\ w \end{pmatrix}) = \begin{pmatrix} \frac{x}{w} \\ \frac{y}{w} \end{pmatrix} \tag{10}$$

is homogenizing function.

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

$$\mathbf{F}_j(\mathbf{m}, t) = \mathbf{U}_j(\mathbf{T}_j{\mathbf{P}_\mathbf{m}}^t). \tag{11}$$

It is evident that the optical flow and the projected 2D motion of ${\mathbf{P}_\mathbf{m}}^t$ should be compatible. Thus from Eq. 9 and 11, the optical flow constraint can be represented as,

$$\|\mathbf{V}_j(\mathbf{m}, t) - \mathbf{F}_j(\mathbf{m}, t)\| \to 0. \tag{12}$$

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 $\underset{i,j}{Corel}(\mathbf{m}_1, \mathbf{m}_2)$. The range of $Corel$ is $[0, 1]$ and "1" means well correlated. Thus if $\mathbf{m}$ is assigned a good disparity, we have,

$$\underset{i,j}{Corel}(\mathbf{T}_i\mathbf{M}^t{\mathbf{P}_\mathbf{m}}^t, \mathbf{T}_j\mathbf{M}^t{\mathbf{P}_\mathbf{m}}^t) \to 1, \quad i, j \in [0, N). \tag{13}$$

Optical flow and stereo constraints are illustrated in Figure 2. In a local region A, the *error-of-fitting* for successive S frames can be defined as Eq. 8, 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

$$EOF = \sum_{t=\tau}^{\tau+S-1} \sum_{\mathbf{m}\in A} \sum_{j=0}^{N-1} \|\mathbf{V}_j(\mathbf{m},t) - \mathbf{F}_j(\mathbf{m},t)\| - w \sum_{t=\tau}^{\tau+S-1} \sum_{\mathbf{m}\in A} \sum_{i,j=0,\ i\neq j}^{N-1} \underset{i,j}{Corel}(\mathbf{T}_i\mathbf{M}^t\mathbf{P_m}^t, \mathbf{T}_j\mathbf{M}^t\mathbf{P_m}^t) \quad (8)$$

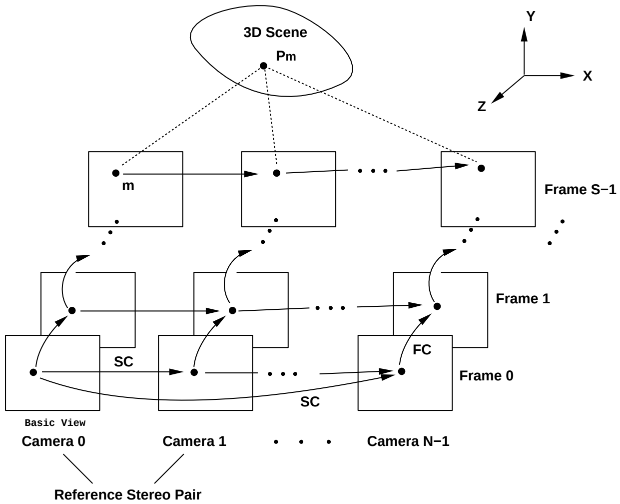

Figure 2: Constraints for local fitting: SC denotes stereo constraints; FC denotes optical flow constraints

from the optical flow constraint is too large, w should be increased. This prevents the motion field from overwhelming the similarity measurement.

Eq. 8 is then used to solve Eq. 7 in our recursive algorithm. Clearly, in this EOF formulation, optical flow and stereo constraints are considered together in order to estimate 3D motion and structure simultaneously.

2.4 Initial Guesses

We use LM algorithm to solve Eq. 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 Eq. 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 $a_1^\tau = 1, b_1^\tau = 0, c_1^\tau = 0, a_2^\tau = 0, b_2^\tau = 1, c_2^\tau = 0, a_3^\tau = 0, b_3^\tau = 0, c_3^\tau = 1, \alpha^{\tau+1} = ... = \alpha^{\tau+S-2} = 1$. We also need the initial depth guess for frame 0. Zhou *et al.* in [31] 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.

2.5 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.* [8, 12]) addressed how to preserve discontinuity in optical flow computation. Experiments have shown [12] 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,

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

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

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

$$C_R = \frac{\lambda}{\|\nabla I\| + \|\nabla D\|} \|V_z - \overline{V}_z\| \quad (15)$$

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

Eq.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,

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

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.

2.6 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 3D 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**
      if $(flag = 1)$
        then
          Local model fitting: add Eq. 15, Eq 16 into Eq. 8, then solve Eq. 7 in region i;
        else
          Local model fitting: add Eq. 16 into Eq. 8, then solve Eq. 7 in region i;
    Compute $\overline{V}_z$ in adjacent Q regions.
    Set $flag := 1$.
end

---

3 Experiments

3.1 Synthetic Scene

In order to test the applicability of our system, we have performed experiments with synthetic multiview image sequences. Since the ground truth is available, we can quantitatively evaluate the system. Specially, we used OpenInventor to generate two-view image sequences (10 frames) consisting of a deformable sphere. Figure 3 shows the generated synthetic input and the recovered structure. The initial depth guess for frame 0 was computed by using the algorithm proposed in [27]. The mean error of structure (in pixels) for every frame is shown in Figure 4. We can see that the largest mean structure error in the sphere is within 1 pixel. Figure 5 shows the recovered 3D scene flow, where we can see that our system tracked the 3D motion successfully.

It is clear that in our system, motion analysis benefits from the stereo constraints (Eq.8). We also want to show how stereo analysis gets refined. Thus, we intentionally added Gaussian noise into the initial depth

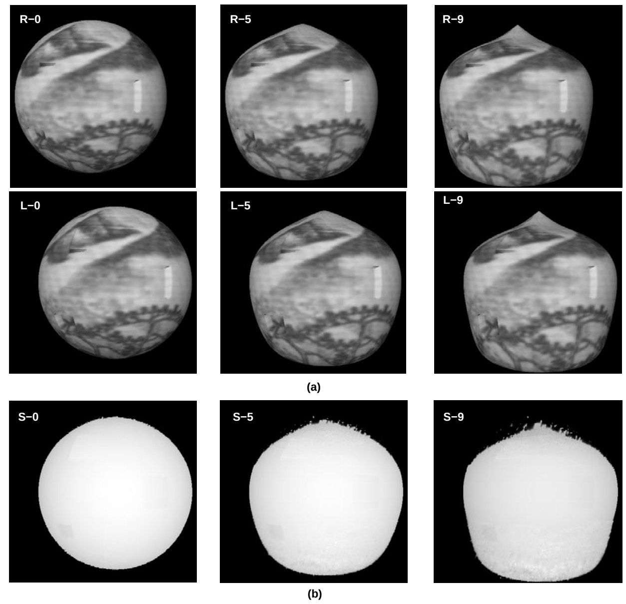

Figure 3: Results on synthetic image sequences: (a) 2-view synthetic image sequences; (b) S-0: refined structure for frame 0; S-5, S-9: recovered 3D structure for frames 5 and 9.

guess and increased the initial mean structure error to 0.76 pixel. As can be seen in Figure 4, the mean structure error for the first frame was decreased to 0.48 pixel by our algorithm, indicating that motion analysis does help the structure refinement in our system.

3.2 Real Scene

In order to test and evaluate our approach in practice, we have performed experiments with real scene sequences. Figure 6 (a) illustrates three-view image sequences used in our system (only the image sequence captured from reference camera is shown). The sequences were acquired with Triclops system: a 3-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 system, the image sequences were captured under poor illumination conditions (thus our initial stereo correspondences and optical flow were noisy). Figure 6 (b), (c) show the recovered scene structure and 3D scene flow. As can be seen, the moving parts in the scene (such as the arms of the subject) were successfully tracked. Also, the recovered structure at the moving parts preserved the correct shape (*e.g.*, the arm can be distinguished).

4 Conclusion and Future Work

We have described a complete and automatic system for 3D scene flow and structure recovery. In our sys-

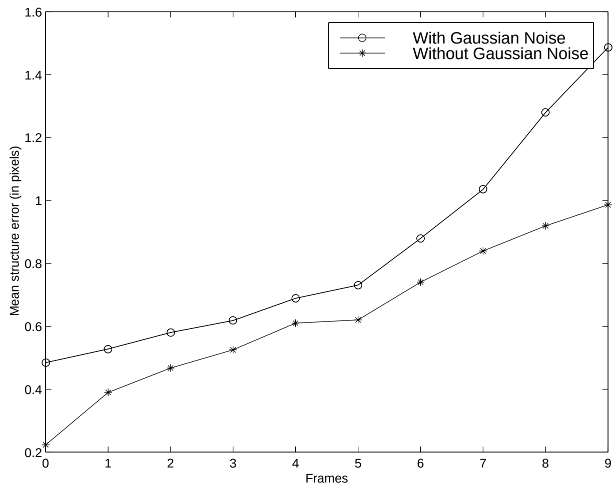

Figure 4: Mean errors of structure recovery for synthetic image sequences

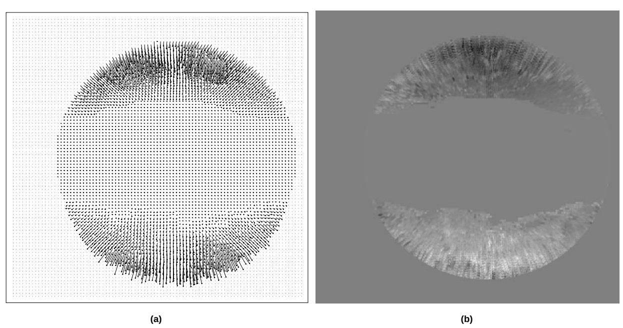

Figure 5: Recovered 3D scene flow of synthetic sequences: (a) Projected 3D scene flow; (b) 3D scene flow along z direction: darker means moving away from the camera; brighter means moving toward the camera.

tem, we integrated 2D motion and stereo constraints and simultaneously computed 3D motion and structure. Through experiments, we showed that motion and stereo benefit from each other. This makes 3D motion and structure analyses more stable. We also quantitatively evaluated the structure analysis of our system based on synthetic input. There are many potential applications for our system such as robust scene structure recovery, dynamic scene interpretation, dynamic rendering, *etc.* In our system, we use correlation measure as a stereo matching criterion. However, other matching measures can be used. We believe more such constraints make the system perform even better.

Our future work includes:

1. Incorporating more sophisticated stereo matching measures;
2. Incorporating more constraints when *a priori* knowledge (*e.g.* model for object of interest) of the scene is available.
3. Implementing robust parallel and multi-resolution methods to improve efficiency;
4. Explaining 3D scene flow events based on spatiotemporal information;

Acknowledgments

Research funding was provided by the National Science Foundation Grant NSF IRI-9619240.

References

[1] Y. Aloimonos and A. Basu. Shape and 3-d motion from contour without point to point correspondences: General principles. In *CVPR86*, pages 518–527, 1986.

[2] R. Balasubramanian, D. Goldgof, and C. Kambhamettu. Tracking of nonrigid motion and 3d structure from 2d image sequences without correspondences. In *ICIP98*, pages I:933–937, 1998.

[3] P. Balasubramanyam. Computation of motion in depth parameters: A first step in stereoscopic motion interpretation. In *DARPA88*, pages 907–920, 1988.

[4] D.H. Ballard and O.A. Kimball. Rigid body motion from depth and optic flow. *CVGIP*, 22(1):95–115, April 1983.

[5] J.L. Barron, A.D. Jepson, and J.K. Tsotsos. Determination of egomotion and environmental layout from noisy time-varying velocity in binocular image sequences. In *IJCAI87*, pages 822–825, 1987.

[6] J.R. Bergen, P. Anandan, K.J. Hanna, and R. Hingorani. Hierarchical model-based motion estimation. In *ECCV92*, pages 237–252, 1992.

[7] M.J. Black. Explaining optical flow events with parameterized spatio-temporal models. In *CVPR99*, pages I:326–332, 1999.

[8] M.J. Black and P. Anandan. The robust estimation of multiple motions: Parametric and piecewise-smooth flow-fields. *CVIU*, 63(1):75–104, January 1996.

[9] I.J. Cox, S.L. Hingorani, S.B. Rao, and B.M. Maggs. A maximum-likelihood stereo algorithm. *CVIU*, 63(3):542–567, May 1996.

[10] U.R. Dhond and J.K. Aggarwal. Structure from stereo: A review. *SMC*, 19(6):1489–1510, November 1989.

[11] F. Dornaika and R. Chung. Stereo correspondence from motion correspondence. In *CVPR99*, pages I:70–75, 1999.

[12] S. Ghosal and P. Vanek. A fast scalable algorithm for discontinuous optical-flow estimation. *PAMI*, 18(2):181–194, February 1996.

[13] T.S. Huang and S.D. Blostein. Robust algorithms for motion estimation based on two sequential stereo image pairs. In *CVPR85*, pages 518–523, 1985.

[14] Y. Huang, K. Palaniappan, X. Zhuang, and J.E. Cavanaugh. Optic flow field segmentation and motion estimation using a robust genetic partitioning algorithm. *PAMI*, 17(12):1177–1190, December 1995.

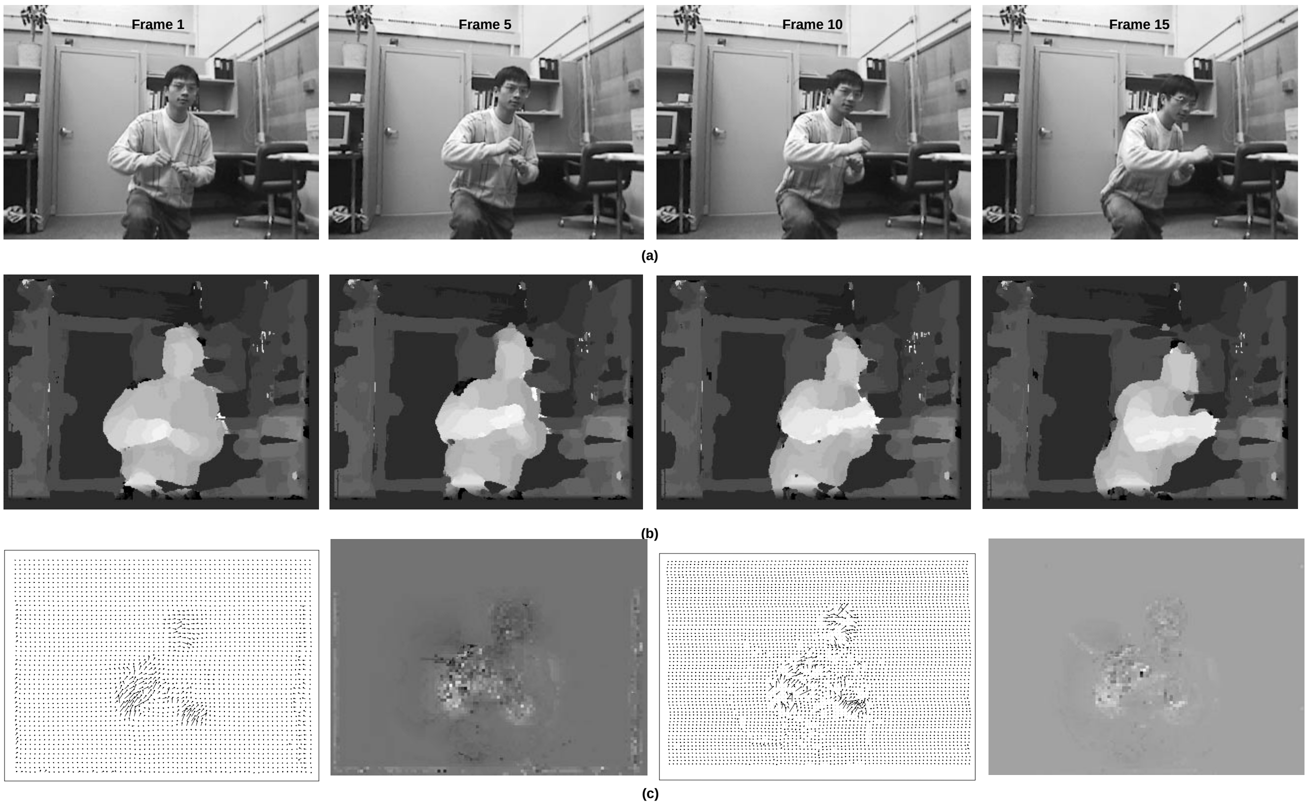

Figure 6: Results on real image sequences: (a) Image sequence gathered with reference camera (one of the 3-view sequences); (b) Recovered Scene Structure; (c) Recovered 3D scene flow at frames 1 and 10. Needle graphs represent the 2D projection, and intensity maps represent the z velocity.

[15] S. Ju, M.J. Black, and A.D. Jepson. Skin and bones: Multi-layer, locally affine, optical flow and regularization with transparency. In *CVPR96*, pages 307–314, 1996.

[16] C. Kambhamettu, D.B. Goldgof, D. Terzopoulos, and T.S. Huang. Nonrigid motion analysis. In *Handbook of PRIP-CV94*, pages 405–430, 1994.

[17] C. Kambhamettu, K. Palaniappan, and A.F. Hasler. Coupled, multi-resolution stereo and motion analysis. In *SCV95*, pages 43–48, 1995.

[18] H. Li, P. Roivainen, and R. Forchheimer. 3-d motion estimation in model-based facial image coding. *PAMI*, 15(6):545–555, June 1993.

[19] L. Li and J.H. Duncan. 3-d translational motion and structure from binocular image flows. *PAMI*, 15(7):657–667, July 1993.

[20] W.H. Liao, S.J. Aggarwal, and J.K. Aggarwal. The reconstruction of dynamic 3d structure of biological objects using stereo microscope images. *MVA*, 9(4):166–178, 1997.

[21] S. Malassiotis and M.G. Strintzis. Model-based joint motion and structure estimation from stereo images. *CVIU*, 65(1):79–94, January 1997.

[22] K.M. Mutch. Determining object translation information using stereoscopic motion. *PAMI*, 8(6):750–755, November 1986.

[23] M.A. Penna. The incremental approximation of nonrigid motion. *CVGIP*, 60(2):141–156, September 1994.

[24] A.P. Pentland and B. Horowitz. Recovery of nonrigid motion and structure. *PAMI*, 13(7):730–742, July 1991.

[25] W. Press, B. Flannery, S. Teukolsky, and W. Vetterling. In *Numerical Recipes in C*, Cambridge University Press, Cambridge, UK, 1988.

[26] W. Richards. Structure from stereo and motion. *J. Opt. Soc. Am. A*, 2(2):343–349, Feburary 1985.

[27] S. Roy and I.J. Cox. A maximum-flow formulation of the n-camera stereo correspondence problem. In *ICCV98*, pages 492–499, 1998.

[28] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *ICCV99*, pages 722–729, 1999.

[29] A.M. Waxman and J.H. Duncan. Binocular image flows: Steps toward stereo-motion fusion. *PAMI*, 8(6):715–729, November 1986.

[30] J. Weng, N. Ahuja, and T.S. Huang. Optimal motion and structure estimation. *PAMI*, 15(9):864–884, September 1993.

[31] L. Zhou, C. Kambhamettu, and D.B. Goldgof. Extracting nonrigid motion and 3d structure of hurricanes from satellite image sequences without correspondences. In *CVPR99*, pages II:280–285, 1999.