A probabilistic framework for geometry and motion reconstruction using prior information

Wende Zhang[1,2] and Tsuhan Chen[1]

[1] Electrical and Computer Engineering Department, Carnegie Mellon University, PA 15213, USA
[2] Electrical and Control Integration Lab, General Motors Corporation, MI 48090, USA

Abstract

In this paper, we propose a probabilistic framework for reconstructing scene geometry and object motion utilizing prior knowledge of a class of scenes, for example, scenes captured by a camera mounted on a vehicle driving through city streets. In this framework, we assume the video camera is calibrated, i.e., the intrinsic and extrinsic parameters are known all the time. While we assume a single camera moving during the capturing, the framework can be generalized to multiple stationary or moving cameras as well. Traditional approaches match the points, lines or patches in multiple images to reconstruct scene geometry and object motion. The proposed framework also takes advantage of each patch's appearance and location to infer its orientation and motion direction using prior information based on statistical learning from training data. The prior hence enhances the performance of geometry and motion reconstruction. We show that the prior-based 3D reconstruction outperformed traditional 3D reconstruction with synthetic data and real data, especially in textureless areas for geometry estimation and faraway areas for motion estimation.

Key Words: Visual Learning; Geometry; Stereo; Motion

Address for correspondence: W. Zhang, Electrical and Control Integration Lab, General Motors Corporation, MI 48090, USA. E-mail: wende.zhang@gm.com

1. Introduction

We categorize scenes by motion: stationary scenes and dynamic scenes. Stationary scenes contain no object motion, while dynamic scenes have at least one moving object.

Image-based geometry reconstruction for stationary scenes has been intensively studied recently (Seitz *et al.*, 2006; Shum *et al.*, 2003).

Pollefeys *et al.* (2004) presented geometry reconstruction systems to automatically extract 3D models from a sequence of images. They matched features and computed the relations between images. From this, both the structure of the stationary scene and the motion of the camera were reconstructed. Collins (1996) proposed an efficient multi-image matching technique using plane-sweeping for geometry reconstruction. Recently, Akbarzadeh *et al.* (2006) extended the plane-sweeping algorithm by sweeping planes in multiple directions for urban geometry reconstruction. Zitnick *et al.* (2004) used the modified plane-sweeping algorithm to estimate the current scene's geometry with a smoothness constraint between patches and a spatial consistency constraint between images. Other methods for geometry estimation include Voxel coloring (Seitz and Dyer 1999) or stereo (Scharstein and Szeliski 2002) methods, etc.

Some previous work requires multiple synchronized cameras to reconstruct dynamic scenes. These algorithms (Akbarzadeh *et al.* 2006; Zitnick *et al.* 2004), which were suitable for the stationary scene reconstruction, could be applied to identify the current scene geometry based on the synchronized multiple cameras without considering temporal consistency.

Reconstruction quality can be further improved if object motion is also estimated. Some researchers assumed a model for a specific class of objects, e.g., the human body,

and estimate its motion. Gavrilla (1999) introduced the marker-free motion capture algorithms that employed a prior model of human body. Fitzgibbon *et al.* (2000) reconstructed the independently moving objects, but only for a sparse set of feature points.

Most existing reconstruction approaches match points, lines or patches among multiple images for scene reconstruction. Considering that humans can easily understand the scene geometry structure based on prior knowledge, Hoiem *et al.* (2005) proposed prior-based geometry estimation from a single image using statistical learning. They could reconstruct a coarse geometry model by classifying each patch into ground, vertical or sky. Saxena *et al.* (2005) also applied supervised learning to predict the depth map of an outdoor scene from a single image. Their depth-map estimation model used a Markov Random Field that contained multi-scale local and global image features, and modeled both the depth at each individual point and the spatial relationship between depths at neighboring points. Zhang and Chen (2007) reconstructed stationary scene geometry from video using geometric prior information. They represented the scenes by small patches with different orientations: horizontal (e.g., ground), vertical (e.g., building facets towards the street and parallel to the camera motion), and frontal (e.g., building facets towards the street and perpendicular to the camera motion).

In this paper, we reconstruct dynamic scenes from video captured by a calibrated camera on a moving vehicle as shown in Figure 2. The camera can be calibrated based on vehicle's GPS sensor, speed sensor, and gyro/yaw-rate sensor. We represent the scenes by small patches with different orientations (Zhang and Chen, 2007); and objects (e.g., vehicles) move on the flat ground. We approximate the patch's motion directions

to be either parallel or perpendicular to the camera motion (e.g., left, right, forward or backward) in a short time period. Since humans can also identify object motion directions from a single image as shown in Figure 3, we also take advantage of each image patch's appearance and location to infer its motion direction from prior information. Our prior-based geometry and motion reconstruction algorithm extends Hoiem *et al.* (2005)'s approach and Zhang and Chen (2007)'s approach to reconstruct dense depth maps and motion maps from video as shown in Figure 4.

The paper is organized as follows. In the next section, we describe the prior-based geometry and motion reconstruction. In Section 3, we show experimental results and compare the geometry and motion reconstruction quality between the approaches with and without prior information as shown in Figure 1. Conclusions and future work are given in Section 4.

2. Prior-Based Geometry and Motion Reconstruction

In this section, we first provide an overview of prior learning methods and prior-based geometry and motion reconstruction approach, and then introduce each component in detail.

2.1 Overview

As shown in Figure 5, for the prior learning, we first segment training images into patches. We then train an orientation estimator and a motion direction estimator using labeled patches.

For the prior-based geometry and motion reconstruction, input images are first segmented into patches S_j. We then infer each patch's orientation distribution $P_j(o)$ and

motion direction distribution $P_j(m_{dir})$ based on image patch's appearance and location using the orientation estimator and the motion direction estimator, respectively. Motion magnitude distribution $P_j(m_{mag})$ is application-dependent. Meanwhile, we calculate the color consistency of every patch among multiple images at an assumed depth *d* with a given orientation *o* and a motion vector **m** to estimate the conditional probability $P_j(d \mid o, \mathbf{m})$, where $\mathbf{m} = [m_{dir}, m_{mag}]$. The initial likelihood of patch's geometry and motion $P_j^0(d, o, \mathbf{m})$ is approximated by the product of the prior probabilities $P_j(o)$, $P_j(m_{dir})$ and $P_j(m_{mag})$, and the conditional probability $P_j(d \mid o, \mathbf{m})$. A coarse patch-based smoothing algorithm is then applied to refine the initial geometry and motion likelihood $P_j^0(d, o, \mathbf{m})$ between its neighboring patches and between its corresponding regions at multiple times/viewpoints iteratively. The maximum likelihood estimates of patch's depth $\hat{d}$, orientation $\hat{o}$ and motion vector $\hat{\mathbf{m}}$, based on the resulting $P_j^t(d, o, \mathbf{m})$, determine initial depth map $d^0(x)$, orientation map, and initial motion map $\mathbf{m}^0(x)$ at each pixel position *x*. The initial depth map $d^0(x)$ and the initial motion map $\mathbf{m}^0(x)$ are further smoothed iteratively per pixel between images to have a refined depth map $d^t(x)$ and a refined motion map $\mathbf{m}^t(x)$.

2.2 Image Segmentation

We use the efficient graph-based image segmentation technique proposed by Felzenszwalb and Huttenlocher (2004). Each image is represented by RGB pixels. The image pixels are grouped into small patches based on their intensities. The use of patches improves the computational efficiency of our algorithm to estimate the depth and motion

for each pixel, and allows complex statistics to be extracted for prior knowledge (orientation and motion direction) estimation.

2.3 Prior Estimation

Prior estimation contains two stages: learning and inference. In the prior learning stage, we first extract features from training patches, and then train the prior estimators using the extracted features with their labels.

Similar to human vision system, texture, color and location features are extracted from each patch as shown in Figure 6. The texture feature is the 15 mean values of the absolute responses of the Leung-Malik (LM) filter bank (Leung and Malik 2001). The color feature is the 6 mean values of RGB and HSV. And the location feature is the 2D mean location in the image coordinates.

In the training stage, we label image patch's orientation into frontal, vertical, horizontal, or sky as shown in Figure 7. Sky is treated as a special category, which is farthest-away frontal patch from the camera. We also label patch's motion direction into stationary, forward, backward, or left/right. We train the orientation estimator and motion direction estimator individually using Support Vector Machines (SVM) probability estimation (Wu *et al.* 2004) based on the labeled patch features. Compared to Hoiem *et al.* (2005)'s approach, we apply a weaker statistic learning approach only based on patch features.

In the inference stage, we calculate the prior distributions of patch's orientation and motion direction. We first extract patch S_j's features, and then infer its orientation distribution $P_j(o)$ and motion direction distribution $P_j(m_{dir})$ using the orientation

estimator and motion direction estimator, respectively. The SVM-based estimators provide the probabilities of all possible labels as shown in Figure 8.

The prior distribution $P_j(m_{mag})$ is approximated by a heuristic distribution $P_j(m_{mag}) \propto \exp(-a \cdot m_{mag})$ with a pre-determined value *a* in our experiment.

2.4 Initial Geometry and Motion Estimation

The initial distribution of patch geometry and motion $P_j^0(d, o, \mathbf{m})$ is evaluated by the product of the prior probabilities $P_j(o)$, $P_j(m_{dir})$, $P_j(m_{mag})$, and the conditional probability $P_j(d \mid o, \mathbf{m})$ of patch's depth *d* given the orientation *o* and motion **m,** where $\mathbf{m} = [m_{dir}, m_{mag}]$.

$$P_j^0(d, o, \mathbf{m}) = P_j(d \mid o, \mathbf{m}) P_j(o, \mathbf{m}), \tag{1}$$

where $P_j^0(o, \mathbf{m}) = P_j^0(o, m_{dir}, m_{mag}) \approx P_j(o) P_j(m_{dir}) P_j(m_{mag})$, since we assume that orientation *o*, motion direction m_{dir}, and motion magnitude m_{mag} are statistically independent of each other.

The conditional probability $P_j(d \mid o, \mathbf{m})$ is determined based on color consistency between images using the extended plane-sweeping algorithm with the given orientation and motion. We assume that scene patches follow a constant motion in a short time period. Patch S_j's depth with the given orientation and motion is evaluated by its color consistency between multiple images at the reference viewpoint (Camera 2) as illustrated in Figure 9. (In our scenario, the camera would be moving towards the scene, which is different from the illustration.) We compare the RGB color difference between every pixel in patch S_j at the reference viewpoint and its corresponding pixels with motion shift

at the other times/viewpoints k, k = 1...N, (in Camera 1 and Camera 3) to measure the color consistency $e_{diff}(S_j)$ using the following robust function with a parameter th:

$$e_{diff}(S_j) = \frac{1}{\text{num}_{S_j}} \sum_k \sum_{pixel \in S_j} \frac{\gamma_k^2}{\gamma_k^2 + th^2}, \quad (2)$$

where $\gamma_k = \left|r_{ref} - r_{cor,k}\right| + \left|g_{ref} - g_{cor,k}\right| + \left|b_{ref} - b_{cor,k}\right|$ is the RGB color difference, and num_{S_j} is the number of the pixels in S_j. The extended plane-sweeping algorithm is an extension to the oriented plane-sweeping algorithm (Akbarzadeh *et al.* 2006) with additional motion estimation.

The conditional probability $P_j(d \mid o, \mathbf{m})$ is determined by the color consistency measures $e_{diff}(S_j)$:

$$P_j(d \mid o, \mathbf{m}) = \frac{g(d, o, \mathbf{m})}{\sum_{d'} g(d', o, \mathbf{m})}, \text{ where } g(d, o, \mathbf{m}) = 1 - e_{diff}(S_j). \quad (3)$$

2.5 Patch-Based Smoothing

We refine patch's initial geometry and motion distribution $P_j^0(d, o, \mathbf{m})$ between its neighboring patches and between its corresponding regions at multiple times/viewpoints iteratively, which is similar to approach in Zitnick *et al.* 2004, with the extension of smoothing for additional orientation and motion estimation. We enforce a smoothness constraint that the neighboring patches (S_j and s_l in Frame 1) with similar colors (blue) should have similar depths ($d \approx d_i$), orientation (o = vertical) and motion vectors as shown in Figure 10. We also ensure scene's geometry and motion consistency constraint between images. We assume that a scene patch (S_j) follow a constant motion ($\mathbf{m}$) in a short time period. If we project S_j with its ground-truth depth d, orientation o, and motion

vector **m** into a neighboring image, the projected region (in Frame 2) should have the similar depth, orientation and motion if not occluded.

The likelihood distribution of patch's geometry and motion $P_j^t(d,o,\mathbf{m})$ is updated iteratively as follows:

$$P_j^{t+1}(d,o,\mathbf{m}) = \frac{n_j(d,o)n_j(\mathbf{m})\prod_{k\in N} c_{j,k}(d,o,\mathbf{m})}{\sum_{d',o',\mathbf{m}'} n_j(d',o')n_j(\mathbf{m}')\prod_{k\in N} c_{j,k}(d',o',\mathbf{m}')} \tag{4}$$

where $n_j(d,o)$ enforces patch's geometry smoothness constraint, and $n_j(\mathbf{m})$ enforces patch's motion smoothness constraint. $c_{j,k}(d,o,\mathbf{m})$ ensures patch's consistency constraint in each projected region at multiple times/viewpoints k, $k = 1\ldots, N$.

The geometry smoothness coefficient $n_j(d,o)$ enforces that the neighboring patches with similar colors should have similar depths and the same orientation.

Let s_l denote one of patch S_j's neighboring patches as shown in Figure 10. $\hat{d}_l$, $\hat{o}_l$ and $\hat{\mathbf{m}}_l$ are the maximum likelihood estimates of its depth, orientation, and motion based on $P_l^t(d,o,\mathbf{m})$, respectively.

$$\left\langle \hat{d}_l \quad \hat{o}_l \quad \hat{\mathbf{m}}_l \right\rangle = \arg\max_{d,o,\mathbf{m}} P_l^t(d,o,\mathbf{m}) \tag{5}$$

We assume that if patches S_j and s_l have the same orientation, the depth d of patch S_j is modeled by a contaminated Gaussian distribution with mean $\hat{d}_l$ and variance σ_l^2. We define $n_j(d,o)$ to be:

$$n_j(d,o) = \begin{cases} \prod_{s_l} N(d;\hat{d}_l,\sigma_l^2) + \varepsilon & o = \hat{o}_l \\ c & o \neq \hat{o}_l \end{cases} \tag{6}$$

where $N(x; mean, \sigma^2)$ is the Gaussian distribution, and ε and c are small constants (e.g.10^{-10}). We evaluate the variance σ_l^2 using color similarity, neighboring similarity, and s_l's geometry and motion maximum likelihood:

- Color similarity of the patches $\Delta_{j,l}$, which measures the color difference between patches S_j and s_l.
- Neighboring similarity $b_{j,l}$, which is the percentage of patch S_j's border between patches S_j and s_l.
- Geometry and motion maximum likelihood for patch s_l: $P_l^t(\hat{d}_l, \hat{o}_l, \hat{\mathbf{m}}_l)$, which represents the accuracy of the maximum likelihood estimates for patch s_l's geometry and motion.

$$\sigma_l^2 \text{ is defined as } \sigma_l^2 = \frac{v}{P_l^t(\hat{d}_l, \hat{o}_l, \hat{\mathbf{m}}_l)^2 b_{j,l} N(\Delta_{j,l}; 0, \sigma_\Delta^2)} \tag{7}$$

where v and σ_Δ^2 are constants. Therefore, if patch S_j and its neighboring patch s_l have similar colors, and patch S_j's depth and orientation are consistent with its neighbor's depth and orientation maximum likelihood estimates, we expect $n_j(d, o)$ to be large.

The motion smoothness coefficient $n_j(\mathbf{m})$ enforces that the neighboring patches with similar colors should have similar motion. We assume that patch S_j's motion vector $\mathbf{m}$ is also modeled by a contaminated Gaussian distribution with mean $\hat{\mathbf{m}}_l$ and variance Σ_l. Therefore, we define $n_j(\mathbf{m})$ as follows:

$$n_j(\mathbf{m}) = \prod_{s_l} N(\mathbf{m}; \hat{\mathbf{m}}_l, \Sigma_l) + \varepsilon \text{, where } \Sigma_l = \sigma_l^2 I_{2\times 2}. \tag{8}$$

If patch S_j and its neighboring patch s_l have similar colors, and patch S_j's motion is consistent with its neighbor's motion maximum likelihood estimate, we expect $n_j(\mathbf{m})$ to be large.

The spatial consistency coefficient $c_{j,k}(d,o,\mathbf{m})$ ensures that the patch S_j's depth, orientation and motion estimates are consistent with the depth, orientation and motion estimates at time/viewpoint k. We compute $c_{j,k}(d,o,\mathbf{m})$ based on spatial consistency, visibility, and patch S_j's initial geometry and motion likelihood:

1. Spatial consistency without occlusion. We first project patch S_j with the depth d, orientation o and motion vector $\mathbf{m}$ onto a neighboring image. We then calculate patch S_j's projecting distribution $b^t_{j,k}(d,o,\mathbf{m})$ based on the geometry and motion distribution at the projected time/viewpoint k to estimate the spatial consistency without occlusion.

$$b^t_{j,k}(d,o,\mathbf{m}) = \frac{1}{num_{S_j}} \sum_{x \in S_j} P^t_{r(k,x)}(d,o,\mathbf{m}) \tag{9}$$

where $r(k,x)$ is the patch index at the time/viewpoint k, on which the corresponding pixel of the pixel position x on patch S_j is. And num_{S_j} is the number of the pixels on patch S_j. If the projected region's depth, orientation and motion maximum likelihood estimates are consistent with patch S_j's estimates, we expect $b^t_{j,k}(d,o,\mathbf{m})$ to be large when patch S_j is visible at the time/viewpoint k (Frame 2) as shown in Figure 10.

2. Visibility. Due to the possible occlusions, a patch might not have the corresponding pixels at another time/viewpoint as shown in Figure 11. We estimate the overall visibility likelihood $v_{j,k}$, that the patch is visible:

$$v_{j,k} = \min\left(1.0, \sum_{d',o',\mathbf{m}'} b^t_{j,k}(d',o',\mathbf{m}')\right) \tag{10}$$

If the patch S_j is visible at the time/viewpoint k (Frame 2) as shown in Figure 10, we can find its corresponding region when we search the space of the depth d, orientation o, and motion $\mathbf{m}$. The ground-truth solution and its neighboring solutions offer large $b^t_{j,k}(d',o',\mathbf{m}')$ values. If the patch S_j is occluded at the time/viewpoint k (Frame 2) as shown in Figure 11, we can not find its corresponding region when we search the space of depth d, orientation o, and motion $\mathbf{m}$. No solution provides large $b^t_{j,k}(d',o',\mathbf{m}')$ value. Therefore, we use $v_{j,k}$ as a robust and computational efficient measure of patch's visibility.

We also estimate the specific visible likelihood $vc_{j,k}(d,o,\mathbf{m})$ that given depth d, orientation o and motion $\mathbf{m}$, patch S_j is visible at time/viewpoint k as follows:

$$vc_{j,k}(d,o,\mathbf{m}) = \frac{1}{num_{S_j}} \sum_{x \in S_j} P^t_{r(k,x)}(d,o,\mathbf{m}) h(\hat{d}'_l - d), \tag{11}$$

where $h(x)$ is the Heaviside step function.

This suggests that if patch S_j is visible at the time/viewpoint k, its corresponding depth d should not be under the surface of the depth map $\hat{d}'_l$ estimated at the time/viewpoint k.

3. Initial S_j's geometry and motion likelihood $P^0_j(d,o,\mathbf{m})$.

Now, we combine the visible and occluded cases. If the patch is visible, $c_{j,k}(d,o,\mathbf{m})$ is calculated from the visible consistency likelihood $b^t_{j,k}(d,o,\mathbf{m})P^0_j(d,o,\mathbf{m})$. Otherwise, its occluded consistency likelihood is $(1-vc_{j,k}(d,o,\mathbf{m}))P^0$ with uniform prior $P^0 = 1/\{size(d)size(o)size(\mathbf{m})\}$. $size(d)$, $size(o)$, and $size(\mathbf{m})$ are the sizes of depth, orientation and motion hypothesis spaces, respectively. Therefore,

$$c_{j,k}(d,o,\mathbf{m}) = v_{j,k} b^t_{j,k}(d,o,\mathbf{m}) P^0_j(d,o,\mathbf{m}) + (1 - v_{j,k})(1 - vc_{j,k}(d,o,\mathbf{m}))P^0. \quad (12)$$

2.6 Pixel-Based Motion and Depth Smoothing

The maximum likelihood estimates of each patch's depth $\hat{d}$, motion $\hat{\mathbf{m}}$ and orientation $\hat{o}$ based on $P^t_j(d,o,\mathbf{m})$ determine initial depth map $d^0(x)$, initial motion map $\mathbf{m}^0(x)$, and orientation map for each pixel *x*.

$$\left\langle \hat{d}_l \quad \hat{o}_l \quad \hat{\mathbf{m}}_l \right\rangle = \arg\max_{d,o,\mathbf{m}} P^t_l(d,o,\mathbf{m}) \quad (13)$$

We further refine the depth map $d^t(x)$ and motion map $\mathbf{m}^t(x)$ iteratively between images. For each pixel *x* at the current time/viewpoint, we find its corresponding pixel *y* at the neighboring time/viewpoint *k* with the constant motion assumption. If the corresponding pixel's depth $d^t_k(y)$ is similar to pixel *x*'s depth $d^t(x)$, the pixel *x*'s depth $d^{t+1}(x)$ is replaced by the average of $d^t(x)$ and $d^t_k(y)$. If the corresponding pixel's motion $\mathbf{m}^t_k(y)$ is similar to pixel *x*'s motion $\mathbf{m}^t(x)$, the pixel *x*'s motion $\mathbf{m}^{t+1}(x)$ is replaced by the average of $\mathbf{m}^t(x)$ and $\mathbf{m}^t_k(y)$. The iterative updating equations are

$$d^{t+1}(x) = \frac{1}{N}\sum_k \left(u_k \frac{d^t(x) + d^t_k(y)}{2} + (1 - u_k) d^t(x) \right) \quad (14)$$

$$\mathbf{m}^{t+1}(x) = \frac{1}{N}\sum_k \left(u_k \frac{\mathbf{m}^t(x) + \mathbf{m}^t_k(y)}{2} + (1 - u_k) \mathbf{m}^t(x) \right) \quad (15)$$

where $\mu_k = (\delta_{d,k}\ and\ \delta_{d,k})$, and $\delta_{d,k} = | d^t(x) - d^t_k(y) | < \lambda_d$, $\delta_{\mathbf{m},k} = | \mathbf{m}^t(x) - \mathbf{m}^t_k(y) | < \lambda_{\mathbf{m}}$. δ is the indicator variable (0,1) testing input similarity with the threshold parameter λ, and *N* is the number of the neighboring images.

3. Experiments

In this section, we first showed the experimental results of the prior-based estimation based on a single image.

We trained the SVM-based orientation estimator with 4756 labeled patches, extracted from 29 color images with the resolution of 320x240 pixel^2. The sample images for training the orientation estimator were shown in Figure 12.

We inferred the orientation distribution of each image patch using the orientation estimator. In Figure 13, we showed the classification results of the orientation estimator on a sample image with the maximum likelihood estimates represented by the shaded colors: red (horizontal), green (vertical), and blue (frontal).

The SVM-based motion direction estimator was trained based on 10109 patches, extracted from 59 color images with resolution 320x240 pixel^2. The sample images for training the motion direction estimator were shown in Figure 12.

We inferred the motion direction distribution of each patch using the motion direction estimator. In Figure 14, we showed the classification results of the motion direction estimator on a sample image with the maximum likelihood estimates represented by the shaded colors: red (forward), green (backward), yellow (left/right) and blue (stationary).

Next, we showed experimental results of the prior-based geometry and motion reconstruction from video.

We ran experiments on synthetic data of a dynamic street simulated by POV-ray (Ray tracking software at http://www.povray.org/). As illustrated in Figure 15, six images were captured by a forward moving camera with the known intrinsic and extrinsic camera

calibration parameters. Only the yellow school bus in the scene was moving towards right.

Each image was segmented into small patches, and patch's orientation and motion direction distributions were inferred based on patch's appearance and location. We applied the prior-based geometry and motion reconstruction algorithm to reconstruct a depth map and a motion map as illustrated in Figure 4.

We compared the results of the proposed prior-based algorithm with the estimated prior distributions ($P_j(o)$ and $P_j(m_{dir})$) and the results of previous work without using any prior, which is effectively our algorithm with uniform prior distributions ($P_j(o) = c_o$ and $P_j(m_{dir}) = c_m$) for geometry and motion reconstruction in Figure 16. The results are summarized in Table 1. Compared with the ground-truth depth map, the prior-based method provided the reconstructed depth map with 6.1 error/pixel on average, while the baseline approach without prior information offered 13.1 error/pixel. Compared with the ground-truth motion map, the prior-based method provided the reconstructed motion map with 0.0084 error/pixel on average, while the baseline approach offered 0.0497 error/pixel. Therefore, the prior-based method provided better depth map and motion map than the baseline approach, especially in the textureless (e.g. ground) areas.

We also ran experiments on real data. A forward moving camera captured 7 input images in a lab as shown in Figure 17. Only the grey vehicle in the scene was moving towards left.

We calibrated the camera's intrinsic parameters (camera's focal length and optical center) with checker board patterns offline and the extrinsic parameters (the translation vector and the rotation matrix) with markers on the ground using Zhang (1998)'s method.

We applied the prior-based geometry and motion reconstruction algorithm on these input images to reconstruct the depth map and motion map at each time instance as shown in Figure 4.

We compared the results of the proposed prior-based algorithm with the estimated prior distributions and the results of previous work without using any prior for geometry and motion reconstruction.

In Figure 18, we showed the comparison of resulting motion maps. The prior-based method also offered better motion maps, since it identified the grey vehicle's correct motion (median speed towards left). The method without prior information provided wrong motion at the ground and faraway background areas, since ground was textureless, and it was difficult to tell whether the faraway scene was stationary or moving slightly without any prior information. Therefore, the prior-based method had better reconstructed depth maps than the traditional approach in Figure 19.

We also showed experimental results of our prior-based geometry and motion reconstruction of a parking lot scene. A forward moving camera captured 6 images with a white car as shown in Figure 20. Only the white car was moving towards right. We have no similar environments in the training set as shown in Figure 12.

In Figure 21, we showed the comparison of resulting motion maps. The prior-based method offered better motion maps, since it recovered the white car's motion (median speed towards right). The method without prior information indicated wrong motions at the ground and sky areas, since the ground was textureless, and it was difficult to tell the motion of the sky without any prior information. Therefore, the prior-based method had better estimated depth maps than the traditional approach in Figure 22.

4. Conclusions and Future Work

In this paper, we proposed a novel framework for reconstructing scene geometry and object motion utilizing prior information. Traditional approaches match the points, lines or patches among multiple images to reconstruct scene geometry and object motion. Our framework also takes advantage of each image patch's appearance and location to infer its orientation and motion direction using statistical learning. We showed that the prior-based geometry and motion reconstruction method outperformed the traditional reconstruction methods, especially in the textureless areas for geometry estimation and faraway areas for motion estimation. Compared with ground-truth values of a synthetic scene, the reconstructed depth-map errors of the proposed method are 1/2 of the errors of the baseline method. The reconstructed motion-map errors of the proposed method are 1/6 of the errors of the baseline method.

In future, we will include rotation into the motion modeling and estimation.

References

Akbarzadeh A., Frahm J.M., Mordohai P., Clipp B., Engels C., Gallup D., Merrell P., Phelps M., Sinha S., Talton B., Wang L., Yang Q., Stewenius H., Yang R., Welch G., Towles H., Nistér D. and **Pollefeys M.** 2006: Towards urban 3D reconstruction from video. *3DPVT*.

Collins R.T. 1996: A space-wweep approach to true multi-image matching. *IEEE CVPR*, 358-363.

Fitzgibbon A.W. and **Zisserman A.** 2000: Multibody structure and motion: 3-D reconstruction of independently moving objects. *ECCV* 1, 891–906.

Felzenszwalb P.F. and **Huttenlocher D.P.** 2004: Efficient graph-based image segmentation. *International Journal of Computer Vision* 59(2), 167-181.

Gavrila D. 1999: The visual analysis of human movement: a survey. *Computer Vision and Image Understandin*g 73(1), 82-98.

Hoiem D., Efros A.A. and **Hebert M.** 2005: Automatic photo pop-up. *ACM Transactions on Graphics* 24(3), 577 - 584 .

Leung T. and **Malik J.** 2001: Representing and recognizing the visual appearance of materials using three-dimensional textons. *International Journal of Computer Vision* 43(1), 29-44.

Pollefeys M., Gool L.V., Vergauwen M., Verbiest F., Cornelis K., Tops J. and **Koch R.** 2004: Visual modeling with a hand-held camera. *International Journal of Computer Vision* 59(3), 207-232.

Saxena A., Chung S.H. and **Ng A.Y.** 2005: Learning depth from single monocular images. *NIPS* 18.

Scharstein D. and **Szeliski R.** 2002: A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *International Journal of Computer Vision* 47(1/2/3), 7-42.

Seitz S.M., Curless B., Diebel J., Scharstein D. and **Szeliski R.** 2006: A comparison and evaluation of multiview stereo reconstruction algorithms. *IEEE CVPR* 1, 519–526.

Seitz S.M. and **Dyer C.R.** 1999: Photorealistic scene reconstruction by Voxel coloring. *International Journal of Computer Vision* 35(2), pp. 151-173.

Shum H.Y., Kang S.B. and **Chan S.C.** 2003: Survey of image-based representations and compression techniques. *IEEE Transactions on Circuits and Systems for Video Technology* 13(11), 1020-1037.

Wu T.F., Lin C.J. and **Weng R.C.** 2004: Probability estimates for multi-class classification by pairwise coupling. *Journal of Machine Learning Research* 5, 975–1005.

Zhang W. and **Chen T.** 2007: A probabilistic framework for geometry reconstruction using prior information. *IEEE ICIP* 2, 529-532.

Zhang Z. 1998: A flexible new technique for camera calibration *Microsoft Technical Report*-98-71.

Zitnick L., Kang S.B., Uyttendaele M., Winder S. and **Szeliski R.** 2004: High-quality video view interpolation using a layered representation. *ACM Transactions on Graphics* 23(3), 600-608.

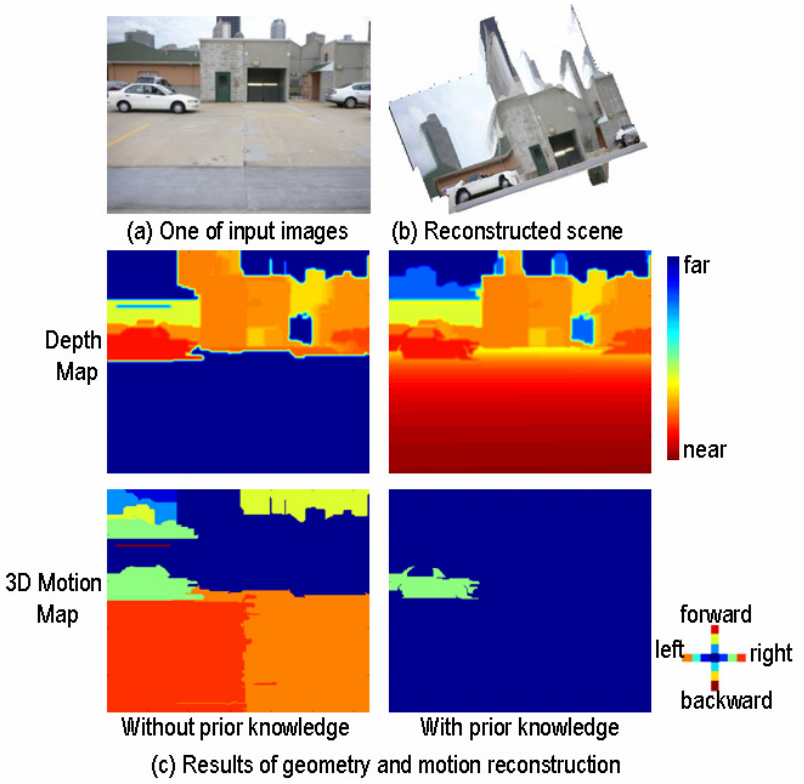

Figure 1. Experimental results. (a) shows one of the images captured in a parking lot by a forward moving camera with a white car moving towards right. (b) is the rendering

result of the prior-based geometry and motion reconstruction. (c) shows that the proposed prior-based method outperforms the traditional technique in terms of the reconstructed depth map and motion map. The depth is represented by a color map. Assuming all objects are moving either left/right or forward/backward (no vertical motion), the motion is quantized into fast, medium, slow and no motion, and represented by different colors in the motion map.

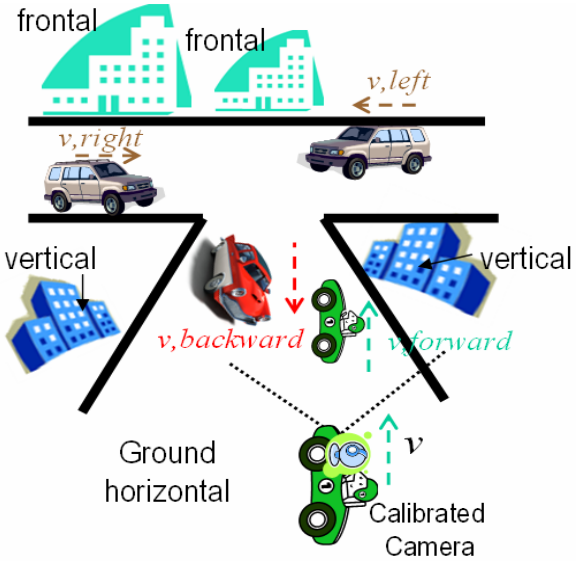

Figure 2. Illustration of scene capture scenario

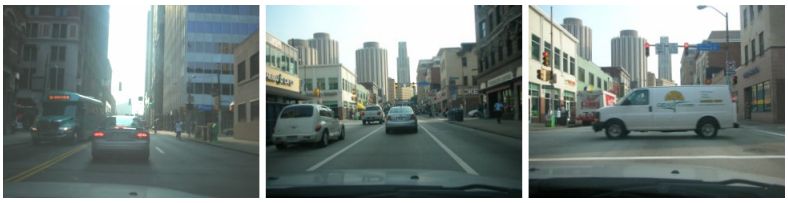

Figure 3. Sample images captured by a moving camera

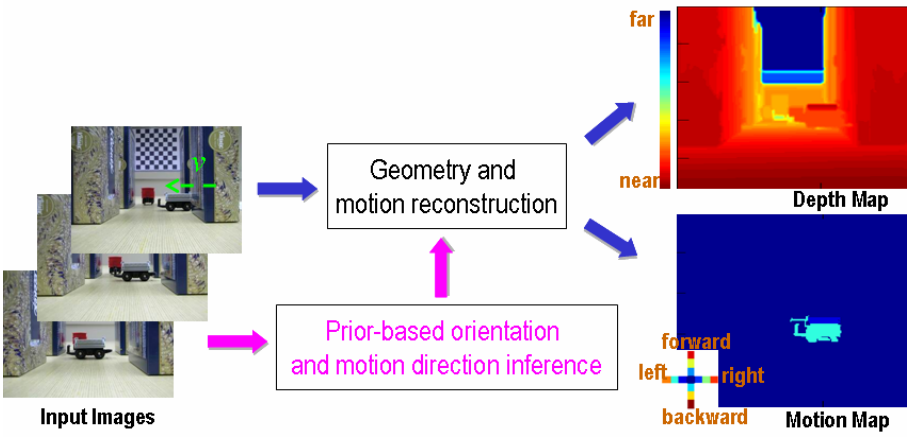

Figure 4. Prior-based geometry and motion reconstruction. The depth is represented by a color map. We assume that objects have no vertical motion. The motion vector (x, y) is quantized into fast, medium, slow or no motion in 4 directions, and represented by different colors for illustration.

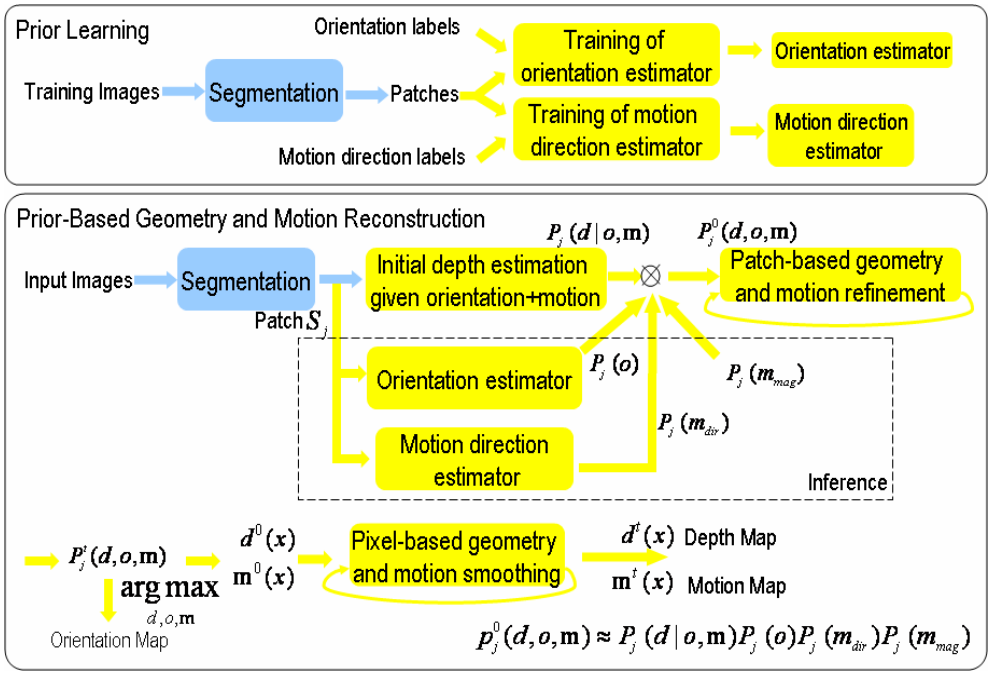

Figure 5. Prior learning and prior-based geometry and motion reconstruction

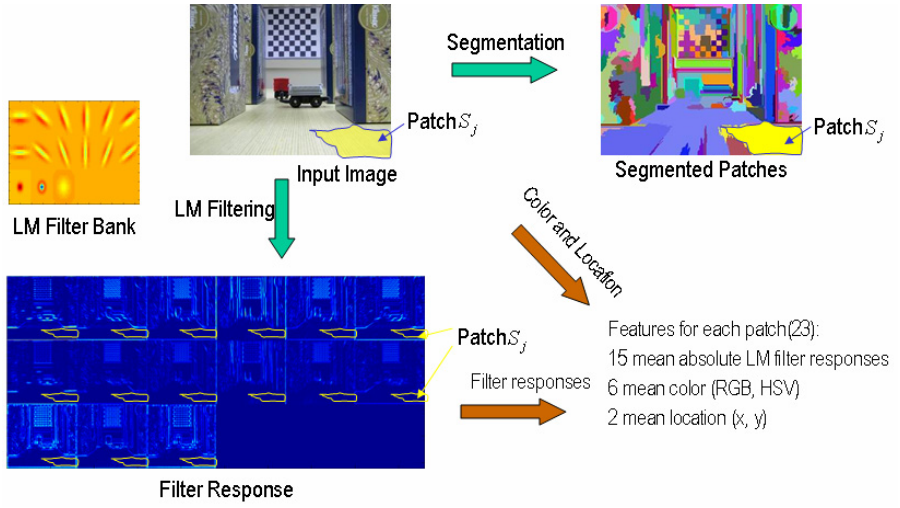

Figure 6. Feature extraction

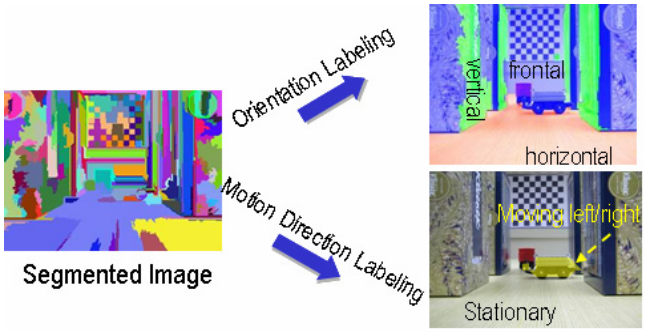

Figure 7. Labeled data for training the estimators

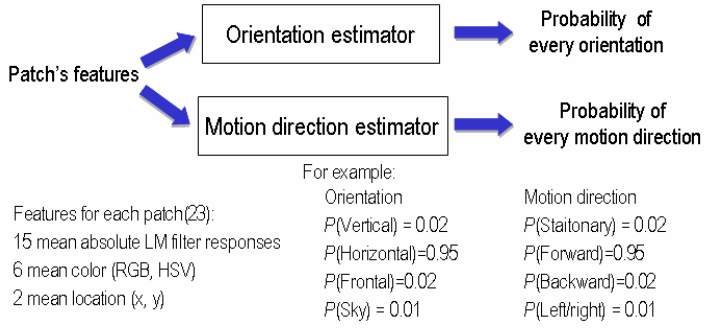

Figure 8. Patch's prior distribution inference

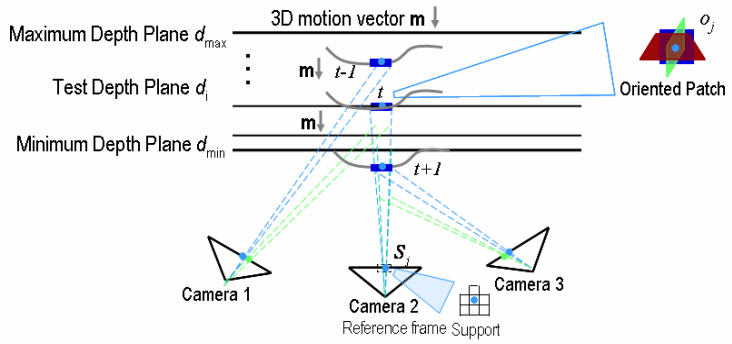

Figure 9. Extended plane-sweeping algorithm with orientation and motion hypotheses

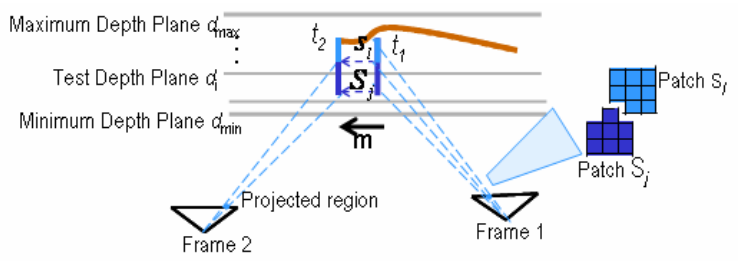

Figure 10. Patch-based smoothing

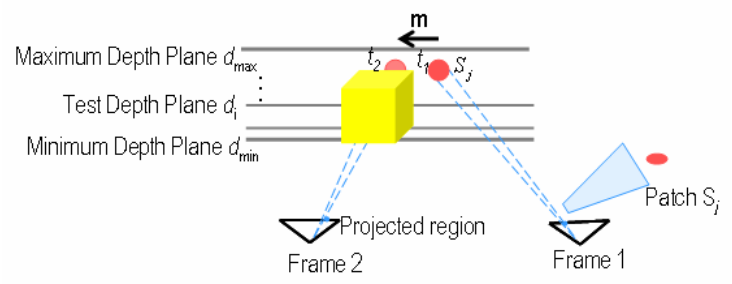

Figure 11. Red patch is occluded in Frame 2.

Figure 12. Sample images for training the orientation estimator and motion direction estimator

(a) Sample image (b) Classification results

Figure 13. Prior-based orientation estimation results

(a) Sample image (b) Classification results

Figure 14. Prior-based motion direction estimation results

Figure 15. Sample input images of a dynamic street

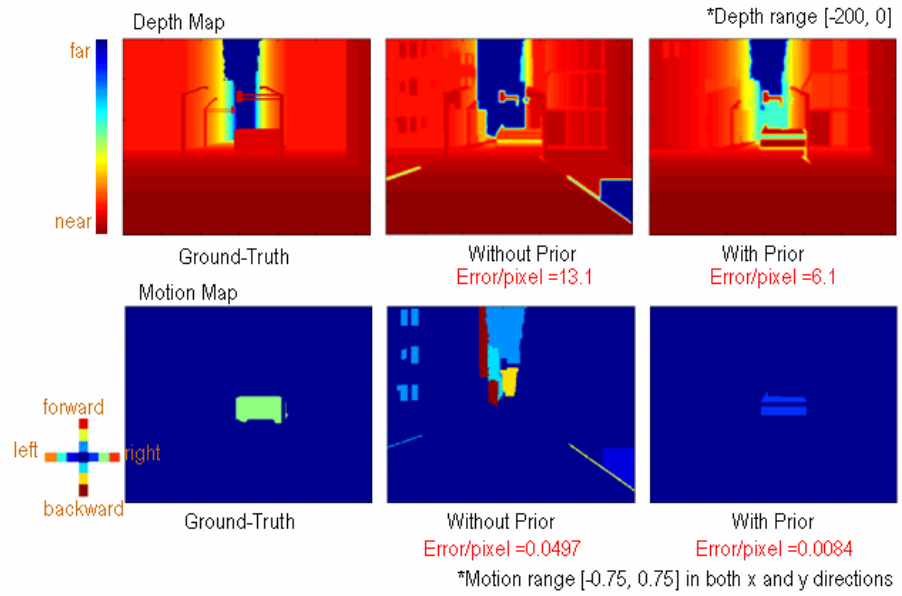

Figure 16. Depth map and motion map comparison of different algorithms for dynamic scene reconstruction.

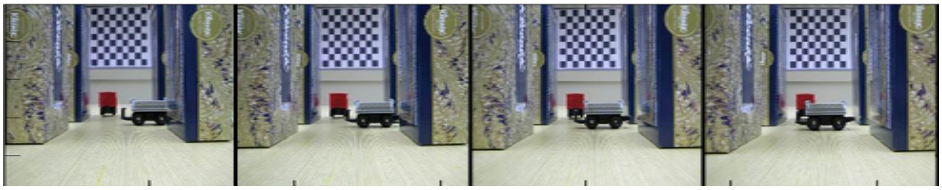

Figure 17. Sample input images of a dynamic lab scene

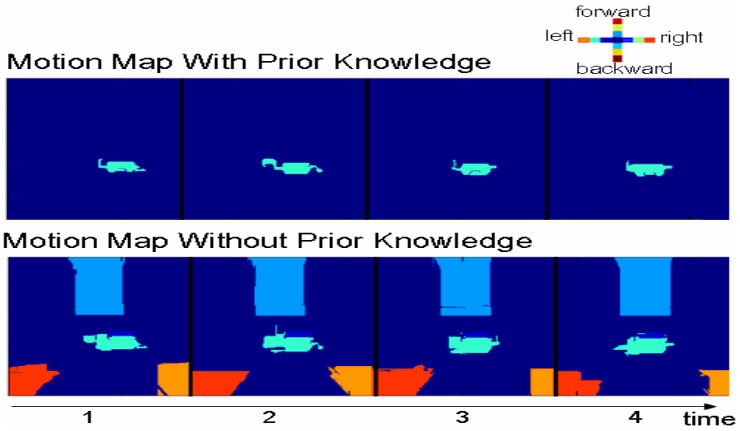

Figure 18. Motion map comparison on a lab scene

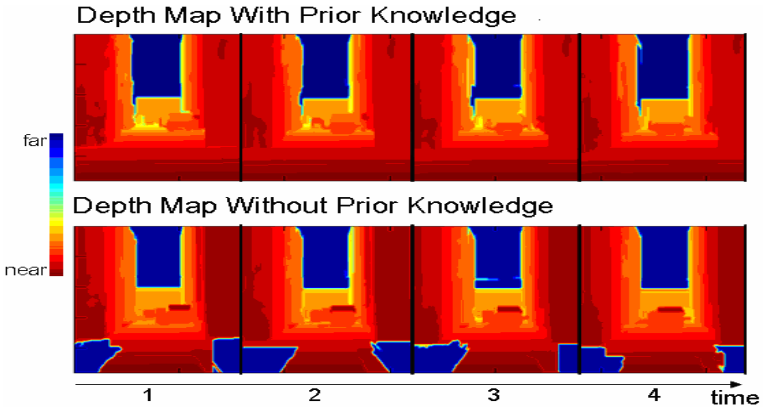

Figure 19. Depth map comparison on a lab scene

Figure 20. Sample input images of a parking lot

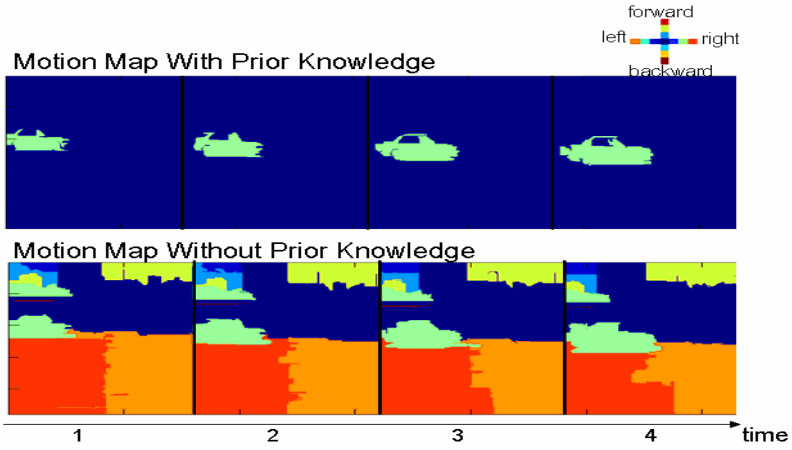

Figure 21. Motion map comparison on an outdoor scene

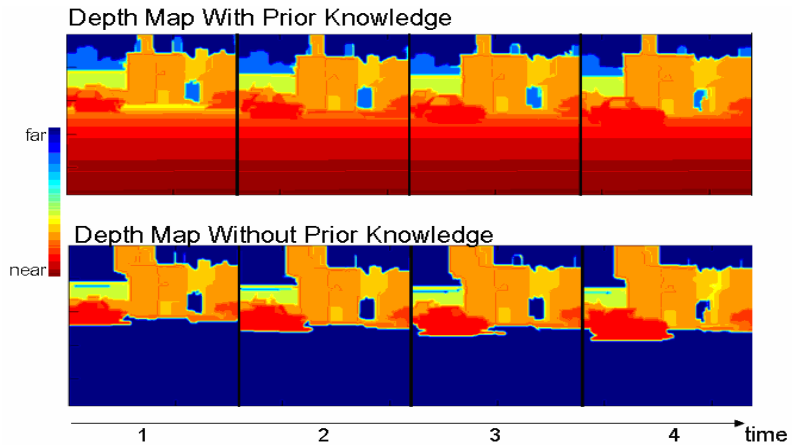

Figure 22. Depth map comparison on an outdoor scene

Reconstruction error (Unit/pixel)	Baseline Approach	Prior-Based Approach
Depth Map	13.1	6.1
Motion Map	0.0497	0.0084

Table 1. Experiment results on synthetic data