RECOVERING DEPTH OF A DYNAMIC SCENE USING REAL WORLD MOTION PRIOR

Adarsh Kowdle, Noah Snavely, Tsuhan Chen

Cornell University, NY, USA.

ABSTRACT

Given a video of a dynamic scene captured using a dynamic camera, we present a method to recover a dense depth map of the scene with a focus on estimating the depth of the dynamic objects. We assume that the static portions of the scene help estimate the pose of the cameras. We recover a dense depth map of the scene via a plane sweep stereo approach. The relative motion of the dynamic object in the scene however, results in an inaccurate depth estimate. Estimating the accurate depth of the dynamic object is an ambiguous problem since both the depth and the real world speed of the object are unknown. In this work, we show that by using occlusions and putting constraints on the speed of the object we can bound the depth of the object. We can then incorporate this real world motion into the plane sweep stereo framework to obtain a more accurate depth for the dynamic object. We focus on videos with people walking in the scene and show the effectiveness of our approach through quantitative and qualitative results.

Index Terms— Computer vision, Image sequences, Image sequence analysis, Depth from video

1. INTRODUCTION

Recovering a dense depth map of a scene from a video is a fundamental yet, hard problem that finds itself a number of applications in video analysis such as video editing, image based rendering, 3D modeling, scene understanding, etc. Given a sequence of images captured using a dynamic camera, a number of variants of stereo matching techniques have been proposed to recover a dense depth map of the scene.

A common feature of the prior approaches is that they are designed for videos of a *static* scene. In this work, we consider the task of recovering a dense depth map of the scene in the presence of *dynamic* objects. The depth of the dynamic objects in the scene would be over-estimated or under-estimated based on whether the object is moving in the same direction or the opposite direction of the camera respectively. We wish to address this issue to infer a more accurate depth estimate for the dynamic regions of the scene.

An overview of the approach is as follows. We assume that we see enough static regions in the scene to recover the camera parameters for the frames of the video. Given the camera poses, we first treat the video as that of a static scene and use a plane sweep stereo algorithm to obtain a dense depth map of the scene. Clearly, the depth estimates of the moving objects would be in error. We identify the regions corresponding to the moving object across the frames using a simple bounding box from the user. Given calibrated cameras there is a well defined relationship between the displacement of objects in the real world and the resulting image space displacement, as a function of the depth of the object. Now, using the depth of regions occluded by the moving object and by constraining the real world speed of the object, we obtain bounds on the depth of the object. We note that, in this work we focus on videos where the moving object is a person walking. We now incorporate these into the original plane sweep stereo framework to estimate a more accurate depth map. We show some results of our algorithm in Fig. 1.

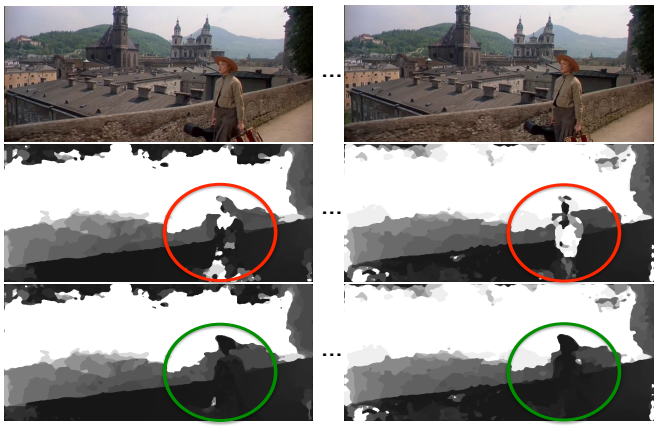


Fig. 1: **ROW 1** shows two frames from a video sequence from the movie *Sound of Music* where, the camera is translating to the left and the person is walking in the same direction. **ROW 2** shows the initial depth maps estimated using plane sweep stereo (white is far, black is close). The depth of the moving object is over-estimated as shown in the red circles. **ROW 3** shows the final depth maps inferred using the proposed approach after identifying and modeling the motion of the moving object. Note that more accurate depth map for the moving object shown in the green circles.

Contributions. The main contributions of this work are: to the best of our knowledge this is the first work that tries to address *dense* depth estimation of the dynamic scene captured by a single monocular camera. We show using quantitative and qualitative results that by obtaining depth bounds via the real world motion, we can obtain a more accurate depth estimate of the dynamic object.

2. RELATED WORK

The task of estimating the depth of a scene given a sequence of images captured from multiple viewpoints has been very well studied. A number of approaches have been proposed to tackle this well defined yet hard task. While enumerating all these is a mammoth task, we refer to some relevant works here that also include exhaustive summaries of the related works.

While one approach is to use the stereo matching algorithms on pairs of rectified frames of the video [1, 2], multiview stereo approaches [3–5] try to estimate the best depth estimate for each pixel using unstructured images. More recent large scale multiview stereo techniques allow for obtaining a dense point cloud or voxelized representation of the scene [6–8] however, our goal in this work is to obtain a dense per view depth map from the video sequence.

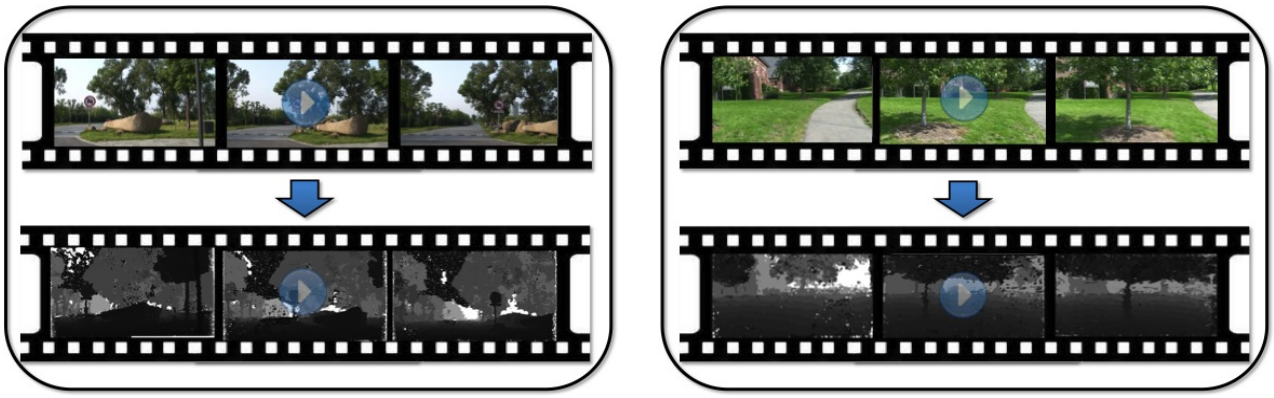


Fig. 2: Resulting depth maps for two static scenes using our plane sweep stereo implementation (White is far, black is close).

A line of work on depth from video by Pollefeys *et al*. have explored fast, real time depth estimation from monocular videos [9]. A recent line of work by Zhang *et al*. have shown some of the best results on depth from video [10]. We note that some prior work model the scene by breaking it down into piecewise planar regions by hypothesizing global planes in the scene, which also provides a dense depth map of the scene [11–13]. While these works have considered the task of depth from video, they focus on the regime of static scenes.

In this work, we wish to estimate a dense depth map of a dynamic scene. A few works have studied dynamic scenes captured simultaneously by multiple cameras to obtain a dynamic point cloud of the scene [14, 15]. Works in non-rigid structure-from-motion have not been cited here but the main focus of this line of work has been to reconstruct key-points on non-rigid objects such as human face, often using images from a static camera. More recently, Zhang *et al*. explored dynamic scenes to segment the moving object with some user interaction [16]. However, to the best of our knowledge we are the first work to consider the task of dense depth estimation of a dynamic scene captured by a monocular camera.

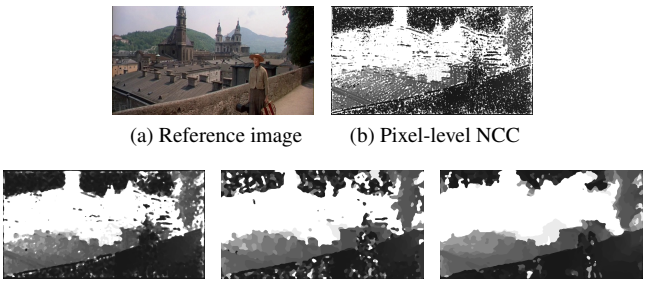
3. ALGORITHM

We describe our algorithm in detail in this section. Given a video of a dynamic scene captured using a dynamic camera, we extract the frames of the video sampled at 30fps. We assume that enough static regions of the scene are observed and recover the camera parameters for the frames of the video using structure-from-motion (SFM) [6].

3.1. Plane sweep stereo

Motivated by the success of the plane sweep stereo algorithm [4, 9, 10], we base our depth from video algorithm on the same framework. We use fronto-parallel planes discretizing the 3D space in inverse depth. In particular, we obtain the minimum ($\frac{1}{D_{max}}$) and maximum inverse depth ($\frac{1}{D_{min}}$) by projecting the 3D points recovered from SFM onto the optical axis. We divide this range into equally spaced bins thus obtaining the 3D planes to perform the plane sweep stereo.

We use a sliding window of 10 frames and normalized cross correlation (NCC) between the reference image and the image obtained by warping the neighboring view onto the 3D plane, as a metric to find the best depth. The result on videos of static scenes are shown in Fig 2. However, scene irregularities such as homogenous surfaces, thin structures and specular surfaces result in a very noisy cost cube. We qualitatively compare algorithmic choices to filter this noisy data in Fig 3. As we observe, the best depth map was obtained by filtering the cost cube by using guided filtering [17]. Using the original image to guide the filtering of the cost cube, we observe that the dominant edges are preserved resulting in a clean output by using an $argmax$ [1] operation on the filtered cost cube.



(a) Reference image (b) Pixel-level NCC

(c) Median filtering (d) Pixel graphcut (e) Guided filtering

Fig. 3: Comparison of algorithmic choices (White is far, black is close for the depth maps). (a) Sample image from a video; (b) Depth map using pixel-level NCC score; (c) Cleaner depth map by median filtering result (a); (d) A pixel-level labeling using graph cuts produces a better result but, noisy; (e) The best result was obtained by guided filtering the cost cube, guided by edges in the original image.

3.2. Dynamic scene

We now consider a video of a dynamic scene. We start off by running the plane sweep stereo algorithm. Note that the depth of the region corresponding to the dynamic object would be incorrectly estimated as observed in ROW 2 of Fig 1. Intuitively, one can obtain a better depth estimate by identifying the spatial region corresponding to the dynamic object and factoring the real world motion of the dynamic object into the plane sweep stereo framework. While some works attempt to segment out the dynamic object (with supervision), estimating the real world motion of the object is non-trivial.

In this work, we segment out the dynamic object via a simple user input in the form of a bounding box around the moving object. We track the bounding box over the successive frames using the optical flow of the spatial region within the bounding box. We note that we can also use other interactive approaches to perform co-segmentation across the frames [16, 18]. We however focus on the second non-trivial task of modeling real world object motion in the next section.

[1] Note that the larger the value, the better since we are using NCC.

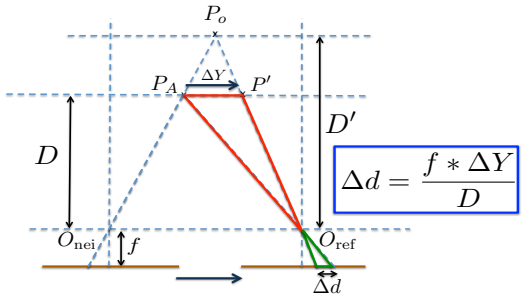


Fig. 4: Modeling the object motion: The relationship in the blue box results from the similarity between the red and green triangles. Please refer to Section 3.2.1 for more details.

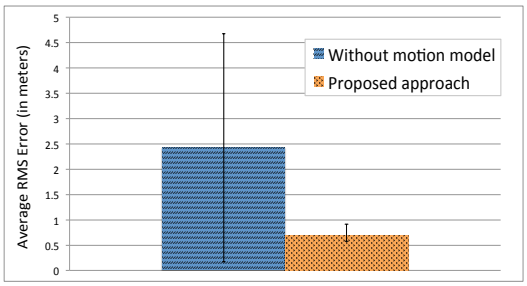


Fig. 5: Quantitative analysis: We show the average RMS error in estimated depth measure using kinect data (averaged over 5 videos with about 50 frames in each video). Note that the proposed approach gives significant improvement.

3.2.1. Modeling the motion of the dynamic object

Estimating the depth of a moving object is a hard and ambiguous task. We estimate the depth using plane sweep stereo by making some assumptions about the object motion in the real world.

Let O_{ref} and O_{nei} be the camera centers of the reference and neighboring frames respectively from a video sampled at r frames per second (Fig 4). Let P_A be the position of the dynamic object as seen from O_{nei} but moves to position P' when seen in O_{ref}. Thus, the depth of the object, D is incorrectly estimated as D'. Let ΔY be the real world distance travelled by the object between the frames *i.e.* the distance traveled by the object traveling at a speed v in time $\Delta t = \frac{1}{r}$. Given the focal length (f), the similarity between the red and green triangles gives the relationship between the disparity Δd in the image space and the real world motion as,

$$\Delta d = \frac{f * \Delta Y}{D} = \frac{f * v * \Delta t}{D} = \frac{f * v}{r * D} \tag{1}$$

While on one hand the depth of the moving object is unknown, the speed at which it is traveling is also unknown. A large object moving fast and far away from the camera, can appear very similar to a smaller object moving slower and located close to the camera. We see from Eqn (1) that this is an under-constrained problem since any pair of the depth (D) and speed (v) can result in the same image projection. In order to relax this ambiguity, we make some assumptions about the object motion in real world. We consider videos with people moving and bound the speed of dynamic object (v) to be within 2 meters per second (human walking speed). Since the video is sampled at r frames per second, this results in a bound on the image space displacement (Δd) as a function of the depth (D) as follows,

$$\Delta d \leq \frac{f * 2}{r} \frac{1}{D} \tag{2}$$

We add an additional bound on the actual depth of the object using the region of the scene occluded by the object. We consider the *minimum* depth of the 3D points from SFM that lies in the occluded region or projects onto the segmented out dynamic object region and use it to upper-bound the depth (D) since the object has to be in-front-of the occluded region.

We then incorporate this into the plane sweep algorithm. While evaluating the cost for a plane hypothesis for the reference frame (i.e. a plane that falls within the depth upper-bound), we use the depth of the hypothesized plane to obtain the bound on displacement (Δd) using Eqn (2). We allow for the segmented dynamic object region in the reference frame to undergo an in-plane shift of a maximum of Δd in either direction and evaluate the best score (NCC) for this region. Intuitively, at the *true* depth, the in-plane shift will allow the segmented region to obtain a better score than before, resulting in a better depth estimate. The result of incorporating the object motion into the depth estimation is evident in Fig 5 and Fig 6. We note that while the estimated depth is more accurate than before, the solution is not unique and is subject to how tight the bounds are.

4. RESULTS

Quantitative results. We capture five indoor scene videos using a Kinect by moving it on a dolly, and extract the aligned ground truth depth map for each frame. A person walked across the scene in two videos and an RC car was driven across the scene in three videos. On an average each video had about 50 frames. We show in Fig 5 the average RMS error in the estimated dense depth maps, computed over the spatial region corresponding to the dynamic object, averaged over all the frames. We note that the proposed approach significantly reduces the error.

Qualitative results. We qualitatively evaluate the performance of the algorithm on amateur video sequences captured by a user and video clips extracted from the movie *Sound of Music*. We show some of the results in Fig 6. Note the inaccurate depth estimate of the dynamic object in ROW 2 of each video, and the significantly improved depth map seen in ROW 3. We also use the depth map to synthesize a stereo pair using a baseline of 77cm. The resulting anaglyph images shown on ROW 4 gives the right perception of depth. We note here that the result using the proposed algorithm may not be the accurate depth due to ambiguous continuous space of possibilities mentioned in Section 3.2.1 however, the proposed algorithm that incorporates the object motion gives significant improvement (Fig 5).

5. CONCLUSIONS AND FUTURE WORK

We present an algorithm using the plane sweep stereo framework to estimate the depth of a *dynamic* scene captured using a dynamic camera. We develop the relationship between the real world object motion and the displacement induced by it in on the image plane. Adding bounds on the motion of the object and incorporating this into the same plane sweep stereo framework we showed that we can obtain a more accurate estimate of depth of the moving object. Future work would explore relaxing some of the assumptions, by intelligently incorporating the user into the loop by quantifying the ambiguity of the algorithm and accepting user input when needed.

6. REFERENCES

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

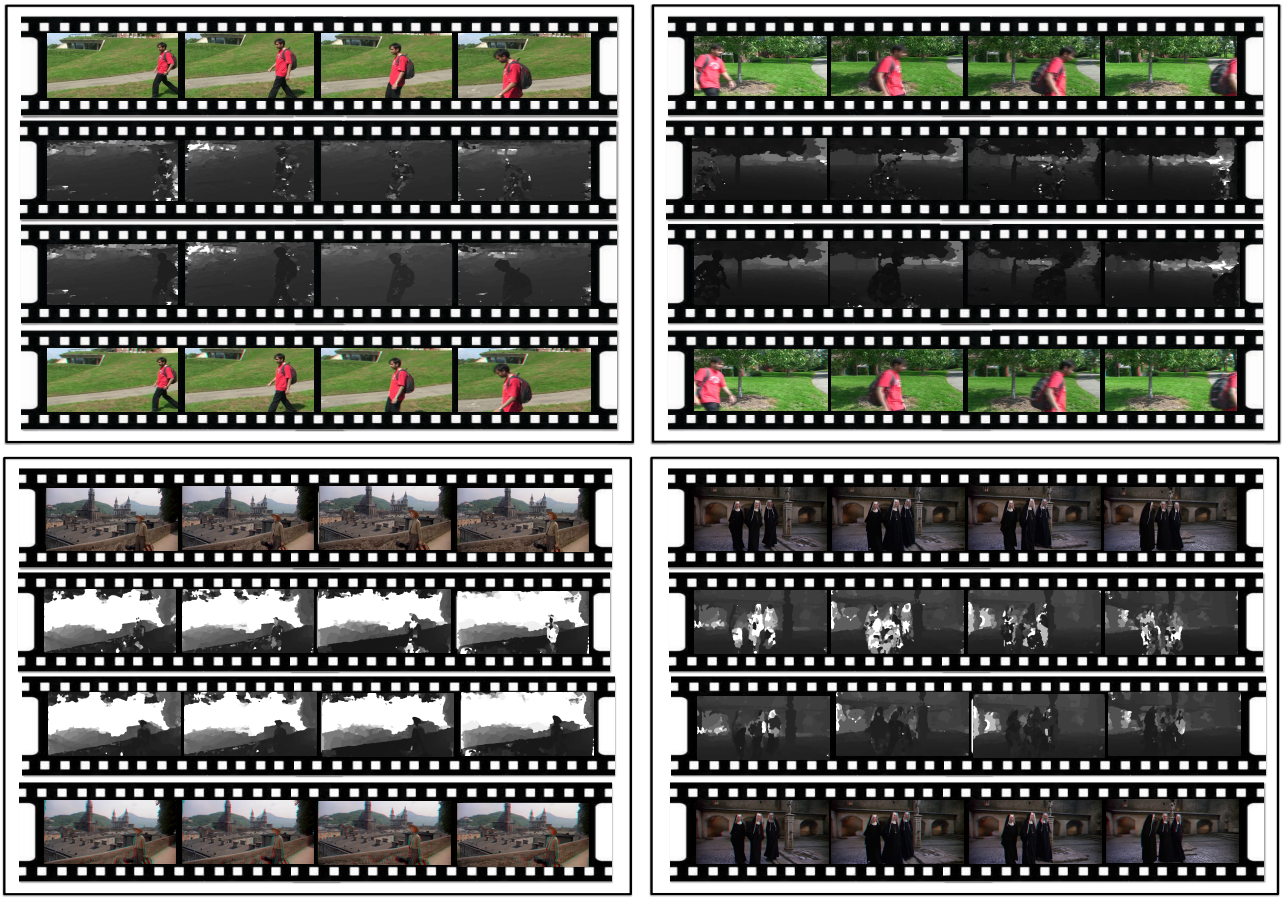


Fig. 6: Each black box shows results on a video sequence. **ROW 1** shows four frames from a video with a moving object. **ROW 2** shows the initial depth maps using plane sweep stereo (white is far, black is close). Note that the depth of the moving object is inaccurately estimated. **ROW 3** shows the final depth maps inferred using the proposed approach after identifying and modeling the motion of the moving object. Note the more accurate depth map for the moving object in each case. **ROW 4** shows anaglyphs obtained by synthesizing the left image of the stereo pair using the original image as the right image and the recovered depth map (Requires red - cyan glasses)

[2] S. M. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski, "A comparison and evaluation of multi-view stereo reconstruction algorithms," in *CVPR*, 2006.

[3] M. Okutomi and T. Kanade, "A multiple-baseline stereo," *PAMI*, vol. 15, pp. 353–363, April 1993.

[4] Robert T. Collins, "A Space-Sweep Approach to True Multi-Image Matching," in *CVPR*, 1996.

[5] Sing Bing Kang and Richard Szeliski, "Extracting view-dependent depth maps from a collection of images," *IJCV*, vol. 58, pp. 139–163, 2004.

[6] N. Snavely, S. Seitz, and R Szeliski, "Photo tourism: Exploring photo collections in 3d," in *SIGGRAPH*, 2006.

[7] Michael Goesele, Noah Snavely, Brian Curless, Hugues Hoppe, and Steven M. Seitz, "Multi-view stereo for community photo collections," in *ICCV*, 2007.

[8] Yasutaka Furukawa, Brian Curless, Steven M. Seitz, and Richard Szeliski, "Towards internet-scale multi-view stereo," in *CVPR*, 2010.

[9] M. Pollefeys, D. Nistr, J. Frahm, A. Akbarzadeh, P. Mordohai, B. Clipp, C. Engels, D. Gallup, S. Kim, P. Merrell, C. Salmi, S Sinha, B. Talton, L. Wang, Q. Yang, H. Stewnius, R. Yang, G. Welch, and H. Towles, "Detailed real-time urban 3d reconstruction from video.," *IJCV*, vol. 78, no. 2-3, pp. 143–167, 2008.

[10] Guofeng Zhang, Jiaya Jia, Tien-Tsin Wong, and Hujun Bao, "Consistent depth maps recovery from a video sequence," *PAMI*, vol. 31, June 2009.

[11] S. Sinha, D. Steedly, and R. Szeliski, "Piecewise planar stereo for image-based rendering," in *ICCV*, 2009.

[12] David Gallup, Jan-Michael Frahm, and Marc Pollefeys, "Piecewise planar and non-planar stereo for urban scene reconstruction," in *CVPR*, 2010.

[13] A. Kowdle, Y. Chang, A. Gallagher, and T. Chen, "Active learning for piecewise planar multiview stereo," *CVPR*, 2011.

[14] Hyun Soo Park, Takaaki Shiratori, Iain Matthews, and Yaser Sheikh, "3d reconstruction of a moving point from a series of 2d projections," in *ECCV*, 2010.

[15] Li Guan, Jean-Sebastien Franco, and Marc Pollefeys, "Probabilistic multi-view dynamic scene reconstruction and occlusion reasoning from silhouette cues," in *IJCV*, 2010.

[16] Guofeng Zhang, Jiaya Jia, Wei Hua, and Hujun Bao, "Robust bilayer segmentation and motion/depth estimation with a handheld camera," *PAMI*, vol. 33, pp. 603–617, March 2011.

[17] Kaiming He, Jian Sun, and Xiaoou Tang, "Guided image filtering," in *ECCV*, 2010.

[18] D. Batra, A. Kowdle, D. Parikh, J. Luo, and T. Chen, *Interactive Co-segmentation of Objects in Image Collections*, SpringerBriefs in Computer Science, 2011.