3D Time-Lapse Reconstruction from Internet Photos

Ricardo Martin-Brualla[1] David Gallup[2] Steven M. Seitz[1,2]
University of Washington[1] Google Inc.[2]
{rmartin,seitz}@cs.washington.edu dgallup@google.com

Abstract

Given an Internet photo collection of a landmark, we compute a 3D time-lapse video sequence where a virtual camera moves continuously in time and space. While previous work assumed a static camera, the addition of camera motion during the time-lapse creates a very compelling impression of parallax. Achieving this goal, however, requires addressing multiple technical challenges, including solving for time-varying depth maps, regularizing 3D point color profiles over time, and reconstructing high quality, hole-free images at every frame from the projected profiles. Our results show photorealistic time-lapses of skylines and natural scenes over many years, with dramatic parallax effects.

1. Introduction

Time-lapses make it possible to see events that are otherwise impossible to observe, like the motion of stars in the night sky or the rolling of clouds. By placing fixed cameras, events over even longer time spans can be imaged, like the construction of skyscrapers or the retreat of glaciers [4]. Recent work [12, 13] has shown the exciting possibility of computing time-lapses from large Internet photo collections. In this work, we seek to compute 3D time-lapse video sequences from Internet photos where a virtual camera moves continuously in both time and space.

Professional photographers exploit small camera motions to capture more engaging time-lapse sequences [10]. By placing the camera on a controlled slider platform the captured sequences show compelling parallax effects. Our technique allows us to recreate such cinematographic effects by simulating virtual camera paths, but with Internet photo collections.

We build on our previous work [12] and introduce key new generalizations that account for time-varying geometry and enable virtual camera motions. Given a user-defined camera path through space and over time, we first compute time-varying depthmaps for the frames of the output sequence. Using the depthmaps, we compute correspondences across the image sequence (aka. "3D tracks").

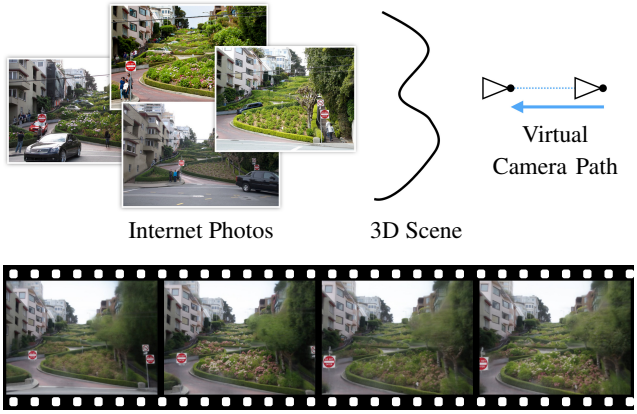


Figure 1. In this paper we introduce a technique to produce high quality 3D time-lapse movies from Internet photos, where a virtual camera moves continuously in space during a time span of several years. Top-left: Sample input photos of the gardens in Lombard Street, San Francisco. Top-right: Schematic of the 3D scene and the virtual camera path. Bottom: Example frames of the synthesized 3D time-lapse video. Please see the supplementary video available at the project website [15]. Credits: Creative Commons photos from Flickr users Eric Astrauskas, Francisco Antunes, Florian Plag and Dan Dickinson.

We then regularize the appearance of each track over time (its "color profile"). Finally, we reconstruct the time-lapse video frames from the projected color profiles.

Our technique works for any landmark that is widely photographed, where, over time, thousands of people have taken photographs of roughly the same view. Previous work [12] identified more than 10,000 such landmarks around the world.

The key contributions of this paper are the following: 1) recovering time-varying, temporally consistent depthmaps from Internet photos via a more robust adaption of [23], 2) a 3D time-lapse reconstruction method that solves for the temporal color profiles of 3D tracks, and 3) an image reconstruction method that computes hole-free output frames from projected 3D color profiles. Together, these contributions allow our system to correctly handle changes in geom-

etry and camera position, yielding time-lapse results superior to those of [12].

2. Related Work

Our recent work [12] introduced a method to synthesize time-lapse videos from Internet Photos spanning several years. The approach assumes a static scene and recovers one depthmap that is used to warp the input images into a static virtual camera. A temporal regularization over individual pixels of the output volume recovers a smooth appearance for the whole sequence. The static scene assumption proved to be a failure mode of that approach resulting in blurring artifacts when scene geometry changes. We address this problem by solving for time-varying geometry, and extend the appearance regularization to 3D tracks and moving camera paths.

Very related to our work, Matzen and Snavely [13] model the appearance of a scene over time from Internet photos by discovering space-time cuboids, corresponding to rectangular surfaces in the scene visible for a limited amount of time, like billboards or graffiti art. Similarly, the 4D Cities project [17, 18] models the changes in a city over several decades using historical imagery. By tracking the visibility of 3D features over time, they are able to reason about missing and inaccurate timestamps. In contrast, we synthesize photorealistic time-lapses of the scene, instead of sparse 4D representations composed of textured rectangular patches or 3D points.

Photobios [7] are visualizations computed from personal photo collections that show how people age through time. The photos are displayed one by one, while fixing the location of the subject's face over the whole sequence. These visualizations are limited to faces and do not create the illusion of time flowing continuously, like our time-lapse sequences do.

Parallax Photography, by Zheng *et al*. [25], creates content-aware camera paths that optimize for parallax effects in carefully collected datasets. Additionally, Snavely *et al*. [22] discover orbit paths that are used to navigate Internet photo collections more efficiently. In our work, the user specifies the camera path as input.

Modeling the appearance of a scene from Internet photos is challenging, as the images are taken with different illumination, at different times of day and present many occluders. Laffont *et al*. [9] regularize the appearance of a photo collection by computing coherent intrinsic images across the collection. Shan *et al*. [20] detect cloudy images in a photo collection, to initialize a factored lighting model for a 3D model recovered from Internet photos.

Generating time-lapse videos from static webcams has also been studied in prior work. Bennett and McMillan [3] propose several objective functions to synthesize time-lapse videos, that showcase different aspects of the changes in the scene. Rubinstein *et al*. [16] reduce flicker caused by small motions in time-lapse sequences.

Kopf *et al*. [8] generate smooth hyper-lapse videos from first-person footage. Their technique recovers scene geometry to stabilize the video sequence, synthesizing views along a smoothed virtual camera path that allows for faster playback.

Although multi-view stereo has been an active topic of research for many years [19], few works have looked into time-varying reconstruction outside of carefully calibrated datasets. Zhang *et al*. [24] reconstruct time-varying depthmaps of moving objects with a spacetime matching term. Larsen *et al*. [11] compute temporally consistent depthmaps given calibrated cameras using optical flow to enforce depth consistency across frames. Zhang *et al*. [23] introduce a method to recover depthmaps of a static scene from handheld captured video sequences. Their method first computes a 3D pose for every frame, and then jointly optimizes the depthmaps for every frame, using a temporal consistency term. We extend their approach to handle dynamic scenes, and adapting it to Internet photo collections.

3. Overview

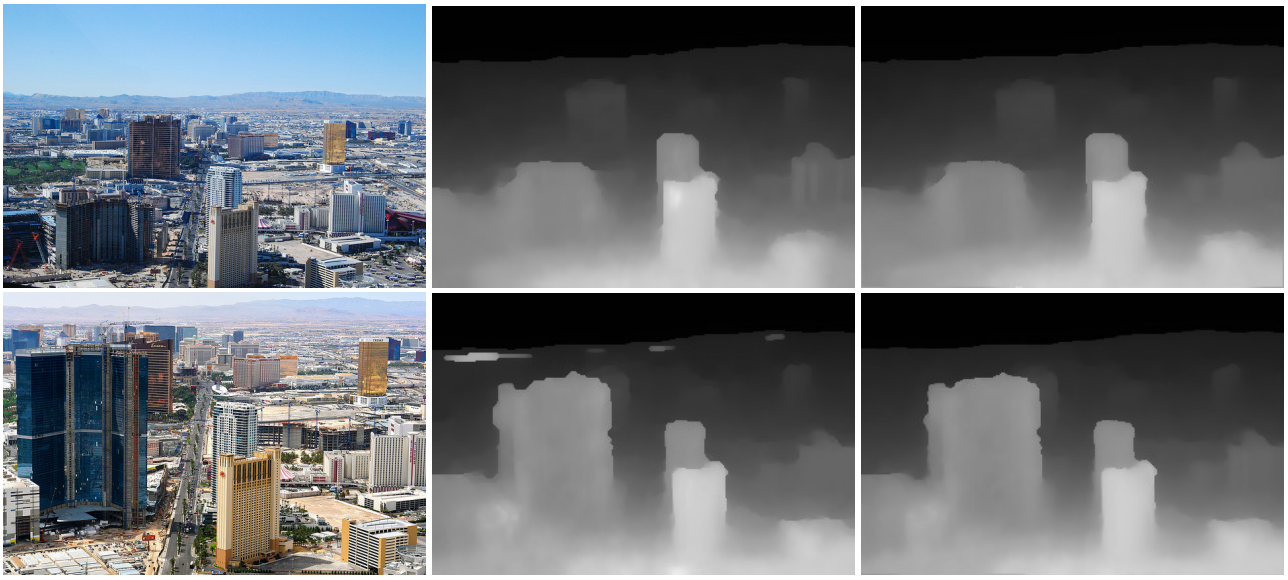
Given an Internet photo collection of a landmark, we seek to compute time-lapse video sequences where a virtual camera moves continuously in time and space. As a preprocessing step, we compute the 3D pose of the input photo collection using Structure-from-Motion (SfM) techniques [1].

First, a user specifies a desired virtual camera path through the reconstructed scene. This can be defined by specifying a reference camera and a parameterized motion path, such as an orbit around a 3D point or a "push" or "pull" motion path [10]. Good reference cameras are obtained using the scene summarization approach of [21].

Our system starts by computing time-varying, temporally consistent depthmaps for all output frames in the sequence, as described in Section 4. Section 5 introduces our novel 3D time-lapse reconstruction, that computes time-varying, regularized color profiles for 3D tracks in the scene. We then present a method to reconstruct output video frames from the projected color profiles. Finally, implementation details are described in Section 6 and results are shown in Section 7.

For the rest of the paper, we only consider images whose cameras in the 3D reconstruction are close to the reference camera. We use the same criteria for image selection as [12], that selects cameras by comparing their optical axis and camera center to those of the reference camera.

Throughout this paper, we will use the following terminology: each photo in the input collection consists of an image I_i, a registered camera C_i and a timestamp t_i. We also define the sequence $\mathcal{I} = (I_1, \ldots, I_N)$ as the chrono-



(a) Sample input photos (b) Initialized depthmap (c) After joint optimization

Figure 2. Results of our time-varying depthmap reconstruction. a) Sample input photos at different times from the Las Vegas skyline scene (not aligned to virtual camera). b) Initialized depthmap for the corresponding time of the photos on the left. c) Jointly optimized depthmaps. Note that artifacts near the top in the second depthmap are fixed after the joint optimization. The improvements to temporal consistency are dramatic and better seen in the supplementary video [15]. Credits: Creative Commons photos from Flickr users Butterbean and Alex Proimos.

logically sorted input image sequence. The output 3D time-lapse sequence is composed of M output frames whose views V_j are equally spaced along the virtual camera path and span the temporal extent of the input sequence, from earliest to the latest photo.

4. Time-Varying Depthmap Computation

In this section we describe how to compute a temporally consistent depthmap for every view in the output sequence. The world changes in different ways over time spans of years compared to time spans of seconds. In multi-year time scales, geometry changes by adding or substracting surfaces, like buildings being constructed or plants growing taller, and we design our algorithm to account for such changes.

Recovering geometry from Internet photos is challenging, as these photos are captured with different cameras, different lighting conditions, and with many occluders. A further complication is that included timestamps are often wrong, as noted in previous work [5, 13]. Finally, most interesting scenes undergo changes in both texture and geometry, further complicating depthmap reconstruction.

4.1. Problem Formulation

Our depth estimation formulation is similar to that of [23], except that we 1) use a Huber norm for the temporal consistency term to make it robust to abrupt changes in geometry, and 2) replace the photo-consistency term with that of [12] which is also robust to temporally varying geometry and appearance changes which abound in Internet photo collections.

We pose the problem as solving for a depthmap D_j for each synthesized view V_j, by minimizing the following energy function:

$$\sum_j \left[E^d(D_j) + \alpha E^s(D_j)\right] + \sum_{j,j'} \beta_{j,j'} E^t(D_j, D_{j'}) \quad (1)$$

where E^d is a data term based on a matching cost volume, E^s is a spatial regularization term between neighboring pixels, and E^t is a binary temporal consistency term that enforces the projection of a neighboring depthmap $D_{j'}$ into the view V_j to be consistent with D_j. The binary weight $\beta_{j,j'}$ is non-zero only for close values of j and j'.

Given the projected depthmap $D_{j'\to j}$ of the depthmap $D_{j'}$ into view V_j, we define the temporal regularization term for a pixel p in V_j as:

$$E^t(D_j, D_{j'})(p) = \delta\left(D_j(p) - D_{j'\to j}(p)\right) \quad (2)$$

if there is a valid projection of $D_{j'}$ in view V_j at p and 0 otherwise, and where δ is a regularization loss. We use z-buffering to project the depthmap so that the constraint is enforced only on the visible pixels of view V_j. Zhang *et al.* [23] assume a Gaussian prior on the depth of the rendered depthmap, equivalent to δ being the L_2 norm. In contrast, our scenes are not static and present abrupt changes in

depth, that we account for by using a robust loss, the Huber norm.

The data term $E^d(D_j)$ is defined as the matching cost computed from a subset of input photos $\mathcal{S}_j \subset \mathcal{I}$ for each view V_j. We choose the subset as the subsequence of length $l = 15\% \cdot N$ centered at the corresponding view timestamp.

Using the images in subset $\mathcal{S}_j$, we compute aggregate matching costs following [12]. First, we generate a set of fronto-parallel planes to the view V_j using the computed 3D SfM reconstruction. We set the range to cover all but the 0.5% nearest and farthest SfM 3D points from the camera. In scenes with little parallax this approach might still fail, so we further discard SfM points that have a triangulation angle of less than 2 degrees.

For each pixel p in view V_j and depth d, we compute the pairwise matching cost $C^j_{a,b}(p,d)$ for images $I_a, I_b \in \mathcal{S}_j$, by projecting both images onto the fronto-parallel plane at depth d and computing normalized cross correlation with filter size 3×3. We adapt the best-k strategy described in [6] to the pairwise matchings costs and define the aggregated cost as:

$$C^j(p,d) = \textbf{median}_{a\in\mathcal{S}_j}\left(\textbf{median}_{b\in\mathcal{S}_j} C^j_{a,b}(p,d)\right) \quad (3)$$

Finally, the spatial regularization E^s consists of the differences of depth between 4 pixel neighborhoods, using the Huber norm, with a small scale parameter to avoid the staircasing effects observed by [14].

4.2. Optimization

The problem formulation of Equation 1 is hard to solve directly, as the binary temporal regularization term ties the depth of pixels across epipolar lines. We optimize this formulation similarly to Zhang *et al.* [23], by first computing each depthmap D_j independently, *i.e.*, without the consistency term E^t, and then performing coordinate descent, where the depthmap D_j is optimized while the others are held constant. We iterate the coordinate descent through all depthmaps for two iterations, as the solution converges quickly.

We solve the problem in the continuous domain with non-linear optimization [2], adapting the data term to the continuous case by interpolating the cost values for a pixel at different depths using cubic splines. We initialize each individual depthmap D_j by solving the MRF formulation of [12] for its corresponding support image set $\mathcal{S}_j$.

The joint optimization produces more stable depthmaps that exhibit fewer artifacts than the initialized ones without the temporal consistency term. Figure 2 shows examples of recovered time-varying depthmaps. The improvements in temporal consistency for the jointly optimized sequence are best seen in the supplementary video [15].

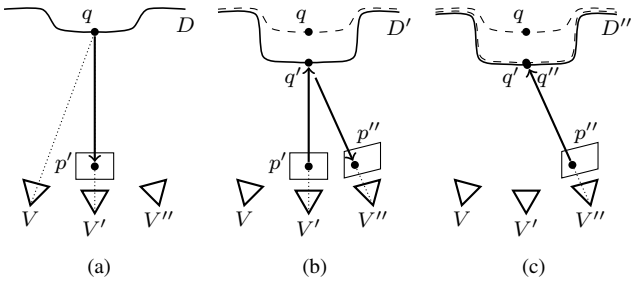


Figure 3. Diagram of how a 3D track is generated in three consecutive views. a) A 3D point q visible in view V is projected to view V' at pixel p'. b) Pixel p' is backprojected onto the depthmap D', creating the 3D point q'. Then, the 3D point q' is projected into view V'' at pixel p''. c) Finally, pixel p'' is backprojected onto the depthmap D'', creating the last point in the track q''. The computed track is $t = (q, q', q'')$. Note that because the geometry remains unchanged between V' and V'', the points q' and q'' are the same.

5. 3D Time-Lapse Reconstruction

Our goal is to produce photorealistic time-lapse videos that visualize the changes in the scene while moving along a virtual camera path. We pose the 3D time-lapse reconstruction problem as recovering time-varying, regularized color profiles for 3D tracks in the scene. A 3D track is a generalization of an image-to-image feature correspondence, which accounts for changes in 3D scene structure, and occlusions between views (See Fig. 3). First, we generate 3D tracks by following correspondences induced by the depthmap and the camera motion. We then solve for the temporal appearance of each 3D track, by projecting them onto the corresponding input images and solving for time-varying, regularized color profiles. Finally, we reconstruct the output time-lapse video from the projected color profiles of the 3D tracks.

5.1. Generating 3D Tracks

We generate 3D tracks that follow the flow induced in the output sequence by the time-varying depthmap and the camera motion. Ideally, a track represents a single 3D point in the scene, whose appearance we want to estimate. However, occlusions and geometry changes may cause a track to cover multiple 3D points. Since the appearance regularization described in the next subsection is robust to abrupt changes in appearance, our approach works well even with occlusions.

A 3D track is defined by a sequence of 3D points $t = (q_{j_1}, \ldots, q_{j_n})$ for corresponding output views $j_1, \ldots, j_n$. To generate a 3D track, we define first a 3D point q for a view V that lies on the corresponding depthmap D. Let p' be the projection of the 3D point q onto the next view V'. We then define the track's next 3D point q' as the backpro-

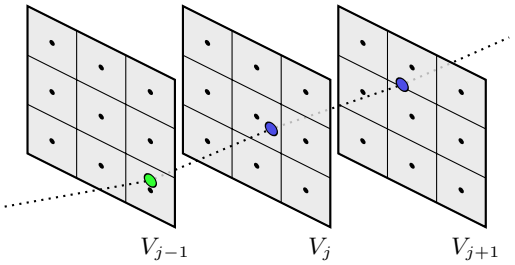
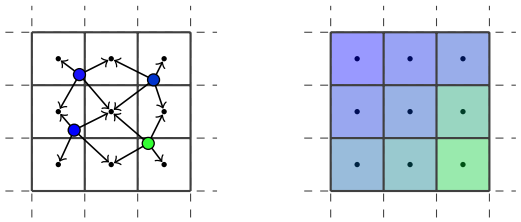


Figure 4. Projected temporal color profiles of a 3D track t into three views. The views are represented by a pixel grid, with the pixel centers marked as black dots. The projected temporal color profiles are defined by a real-valued projected position p_j^t into view j and a time-varying, regularized color y_j^t. The projected profile is shown as a sequence of colored circles, projected on each view, linked by a dashed line.



(a) Projected color profiles (b) Reconstructed image

Figure 5. Visualization of the output frame reconstruction algorithm from projected color profiles. Left: Projected color profiles at a given view shown as colored dots in the output frame with their bilinear interpolation weights shown as arrows from the projected sample to pixel centers. Right: We reconstruct an image that minimizes the bilinear interpolation residuals of the projected color profiles.

jection of pixel p' onto the corresponding depthmap D'. We compute the next 3D point q'' by repeating this process from q'. We define a whole track by iterating forwards and backwards in the sequence, and we stop the track if the projection falls outside the current view. 3D tracks are generated so that the output views are covered with sufficient density as described in Section 5.3.

Figure 3 shows the 3D track generation process. Note that when the geometry is static, points in a 3D track remain constant thanks to the robust norm used in the temporal consistency term, that promotes depthmap projections to match between frames. While drift can occur through this chaining process, in practice this does not affect the quality of the final visualizations.

5.2. Regularizing Color Profiles

We want to recover a time-varying, regularized color profile for each 3D track t. This is challenging as Internet photos display a lot of variation in appearance and often contain outliers, as noted in Section 4. We make the observation that the albedo of most surfaces in the real world does not change rapidly, and its variability in appearance stems mostly from illumination effects. Intuitively, we would like our time-lapse sequences to reveal the infrequent texture changes (the signal) while hiding the variability and outliers of the input photo collection (the noise).

To solve for time-varying color profiles, [12] used a temporal regularization term with a robust norm, that recovers piecewise continuous appearances of pixels in an output image sequence. The approach is restricted to a static virtual camera, as it works on the 2D domain by regularizing each pixel in the output sequence independently. Our approach uses the same temporal term to regularize the color profile of each 3D track.

Given a 3D track $t = (q_{j_1}, \ldots, q_{j_n})$, we define its appearance in view V_j as the RGB value $y_j^t \in [0,1]^3$. To compute y_j^t, we first assign input images to their closest view in time and denote these images assigned to view V_j by the support set $\mathcal{S}'_j \subset \mathcal{I}$. Note that the sets $\mathcal{S}'_j$ are not overlapping, whereas the support sets $\mathcal{S}_j$ used for depthmap computation are. We then project the 3D point q_j to camera C_i using a z-buffer with the depthmap D_j to check for occlusions and define x_i^t as the RGB value of image i at the projection of q_j.

We obtain a time-varying, regularized color profile for each 3D track t by minimizing the following energy function:

$$\sum_j \sum_{i \in \mathcal{S}'_j} \delta_d \left(\left\| x_i^t - y_j^t \right\| \right) + \lambda \sum_j \delta_t \left(\left\| y_{j+1}^t - y_j^t \right\| \right) \quad (4)$$

where the weight λ controls the amount of regularization, and both δ_d and δ_t are the Huber norm, to reduce the effects of outliers in x_j^t and promote sparse temporal changes in the color profile.

In contrast to [12], the color profiles of the 3D tracks do not correspond to pixels in the output frames. We thus save the color profile y^t, together with the 2D projections p_j^t of the track 3D points q_j^t into the view j, as *projected profiles* that are used to reconstruct the output frames. Figure 4 shows a diagram of a projected color profile.

5.3. Reconstructing Video from Projected Profiles

Given regularized projected color profiles computed for a set of 3D tracks $\mathcal{T}$, we seek to reconstruct output frames of the time-lapse video that best fit the recovered color profiles.

We cast the problem of reconstructing each individual frame as solving for the image that best matches the color values of the projected color profiles when applying bilinear interpolation at the profiles' 2D projections. Figure 5 visualizes the reconstruction process, where the output pixels' color values are related to the projected profiles' samples by bilinear interpolation weights.

Figure 6. Comparison of different values of the 3D track sampling threshold ϵ for the Wall Street Bull scene. Left: Artifacts are visible when $\epsilon = 1$ pixel, with alternating black and white pixels, as the reconstruction problem is badly conditioned. Right: Using $\epsilon = 0.4$ pixel, the artifacts are not present.

For a given output view V_j, let $Y_{u,v} \in [0,1]^3$ be the RGB value of the pixel $(u,v) \in \mathbb{N}^2$ in the synthesized output frame Y. Let the regularized projected profile for a track t at view V_j have an RGB value y^t and a 2D projection $p^t \in \mathbb{R}^2$. We solve for the image Y that minimizes

$$\sum_{t \in \mathcal{T}} \left\| y^t - \sum_{s=1}^{4} w_s^t Y_{u_s^t, v_s^t} \right\|_2 \qquad (5)$$

where u_s^t, v_s^t are the integer coordinates of the 4 neighboring pixels to p^t and w_s^t their corresponding bilinear interpolation weights.

The reconstruction problem requires the set of 3D tracks $\mathcal{T}$ to be dense enough that every pixel $Y_{u,v}$ has a non-zero weight in the optimization, *i.e.*, each pixel center is within 1 pixel distance of a projected profile sample. To ensure this, we generate 3D tracks using the following heuristic: we compute 3D tracks for all pixels p in the middle view j of the sequence, so that the 3D track point q_j^t projects to the center of pixel p in V_j. Then, we do the same for all pixels in the first and last frame. Finally, we iterate through all pixels in the output frames Y and generate new 3D tracks if there is no sample within $\epsilon \leq 1$ pixels from the pixel center coordinates.

The reconstruction problem can be badly conditioned, producing artifacts in the reconstructions, such as contiguous pixels with alternating black and white colors. This happens in the border regions of the image that have lower sample density. We avoid such artifacts by using a low threshold value $\epsilon = 0.4$ pixels, so that for each pixel there is a projected profile whose bilinear interpolation weight is > 0.5. Figure 6 shows an example frame reconstruction using two different threshold values for ϵ.

6. Implementation

The photo collections used in our system consist of publicly available Picasa and Panoramio images. For a single landmark, the 3D reconstructions contain up to 25K photos, and the input sequences filtered with the camera selection criteria of [12] contain between 500 and 2200 photos. We generate virtual camera paths containing between 100 and 200 frames.

Figure 7. Comparison of two methods for output frame reconstruction from projected profiles for the Musée D'Orsay scene. Left: baseline method based on Gaussian kernel splatting, with kernel radius $\sigma = 1$. Right: our reconstruction approach. The baseline method produces a blurred reconstruction, whereas the proposed approach recovers high frequency details in the output frame.

The weights for the depthmap computation are $\alpha = 0.4$ and the temporal binary weight is defined as $\beta_{j,j'} = k_1 \max(1 - |j' - j|/k_2, 0)$ with $k_1 = 30$ and $k_2 = 8$. The scale parameter of the Huber loss used for E^s and E^t is 0.1 disparity values. For appearance regularization, we use the Huber loss for δ_d and δ_t with scale parameter of 1^{-4}, *i.e.*, about $1/4$ of a pixel value. Finally, the temporal regularization weight is $\lambda = 25$. We use Ceres Solver [2] to solve for the optimized color profiles, that we solve per color channel independently.

Our multi-threaded CPU implementation runs on a single workstation with 12 cores and 48Gb memory in 4 hours and 10 minutes for a 100 frame sequence. The breakdown is the following: 151 minutes for depthmap initialization, 30 minutes for joint depthmap optimization, 55 minutes for 3D track generation and regularization, and 25 minutes for video reconstruction. We compute the output sequences at an image resolution of 800×600, with a depthmap resolution of 400×300. Our execution time is dominated by the cost volume computation for all the views, and we subsample the support sets $\mathcal{S}_j$ to contain at most 100 images without noticeable detrimental effects.

7. Results

We generated high-quality 3D time-lapse videos for 14 scenes, spanning time periods between 4 and 10 years. Figure 10 shows sample frames from four different scenes. We refer the reader to the supplementary video [15] to better appreciate the changes in the scenes and the parallax effects in our 3D time-lapses.

Figure 8 shows a comparison of our 3D time-lapse for the Las Vegas sequence with the result of previous work [12], that was noted as a failure case due to changing scene geometry. Our 3D time-lapse result eliminates the blurry artifacts, as the time-varying depthmap recovers the building construction process accurately.

(a) Static depthmap (b) Time-varying depthmap

Figure 8. Comparison of output time-lapse frames for two different timestamps for the Las Vegas sequence. a) Using a static depthmap solved with a discrete MRF as in [12]. b) Using our time-varying, temporally consistent depthmaps. The static depthmap is not able to stabilize the input images for the whole time-lapse, creating blurry artifacts where the geometry changes significantly. Thanks to the time-varying depthmap, our 3D time-lapses are sharp over the whole sequence.

(a) Missing thin structures (b) Extrapolation artifacts

Figure 9. Examples of failure cases in our system. a) The street sign is not fully reconstructed in the Lombard Street sequence. b) An extended camera orbit contains a virtual camera far from the set of input cameras causing blurry artifacts in the Flatiron Building dataset.

We also compare our output frame reconstruction approach with a baseline method that uses splatting of the projected color profiles with Gaussian weights. Each projected profile sample contributes its color to nearby pixels with a weight based on the distance to the pixel center. Figure 7 shows that the baseline produces blurred results whereas our approach recovers high frequency details in the output frame.

7.1. Limitations

We observed a few failure cases in our system. Inaccurate depthmaps create blurring or shearing artifacts, especially if close objects are present. For example, in the Lombard Street sequence shown in Figure 9(a), the system fails to reconstruct thin structures, blurring them away. Recovering more accurate, time-varying geometry from Internet photo collections is an area of future work.

Our system also generates artifacts when extrapolating the input photo collection. This happens when a camera looks at a surface not visible in any input photo. For example, in Figure 9(b) a view is synthesized for a camera outside the convex hull of reconstructed cameras, showing a face of a building that is not visible from any photo. Future work could consider using visibility information to constrain the virtual camera paths like in [25].

Our technique is limited to reconstructing 3D time-lapses given pre-specified camera paths. Future work includes enabling interactive visualizations of these photorealistic 3D time-lapses.

8. Conclusion

In this paper we introduce a method to reconstruct 3D time-lapse videos from Internet photos where a virtual camera moves continuously in time and space. Our method involves solving for time-varying depthmaps, regularizing 3D point color profiles over time, and reconstructing high quality, hole-free output frames. By using cinematographic camera paths, we generate time-lapse videos with compelling parallax effects.

(a) Flatiron Building, New York

(b) Lombard Street, San Francisco

(c) Ta Prohm, Cambodia

(d) Palette Springs, Yellowstone

Figure 10. Frames from example 3D time-lapses, with time spans of several years and subtle camera motions. Sequences a), c) and d) contain an orbit camera path, while b) contains a camera "push". Parallax effects are best seen in the video available at the project website [15]. Limitations of our system include blurry artifacts in the foreground, like in c) and d).

Acknowledgements

The research was supported in part by the National Science Foundation (IIS-1250793), the Animation Research Labs, and Google.

References

[1] S. Agarwal, Y. Furukawa, N. Snavely, I. Simon, B. Curless, S. M. Seitz, and R. Szeliski. Building rome in a day. *Communications of the ACM*, 54(10):105–112, 2011. 2

[2] S. Agarwal, K. Mierle, and Others. Ceres Solver. http://ceres-solver.org. 4, 6

[3] E. P. Bennett and L. McMillan. Computational time-lapse video. In *ACM SIGGRAPH 2007 Papers*, SIGGRAPH '07, New York, NY, USA, 2007. ACM. 2

[4] Earth Vision Institute. Extreme Ice Survey. http://extremeicesurvey.org/, 2007. 1

[5] D. Hauagge, S. Wehrwein, P. Upchurch, K. Bala, and N. Snavely. Reasoning about photo collections using models of outdoor illumination. In *Proceedings of BMVC*, 2014. 3

[6] S. B. Kang and R. Szeliski. Extracting view-dependent depth maps from a collection of images. *International Journal of Computer Vision*, 58:139–163, 2004. 4

[7] I. Kemelmacher-Shlizerman, E. Shechtman, R. Garg, and S. M. Seitz. Exploring photobios. In *ACM SIGGRAPH 2011 Papers*, SIGGRAPH '11, pages 61:1–61:10, New York, NY, USA, 2011. ACM. 2

[8] J. Kopf, M. F. Cohen, and R. Szeliski. First-person hyper-lapse videos. *ACM Trans. Graph.*, 33(4):78:1–78:10, July 2014. 2

[9] P.-Y. Laffont, A. Bousseau, S. Paris, F. Durand, and G. Drettakis. Coherent intrinsic images from photo collections. *ACM Transactions on Graphics (SIGGRAPH Asia Conference Proceedings)*, 31, 2012. 2

[10] V. Laforet. Time Lapse Intro: Part I. http://blog.vincentlaforet.com/2013/04/27/time-lapse-intro-part-i/. 1, 2

[11] E. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs. Temporally consistent reconstruction from multiple video streams using enhanced belief propagation. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007. 2

[12] R. Martin-Brualla, D. Gallup, and S. M. Seitz. Time-lapse mining from internet photos. *ACM Trans. Graph.*, 34(4):62:1–62:8, July 2015. 1, 2, 3, 4, 5, 6, 7

[13] K. Matzen and N. Snavely. Scene chronology. In *Proc. European Conf. on Computer Vision*, 2014. 1, 2, 3

[14] R. A. Newcombe, S. Lovegrove, and A. Davison. Dtam: Dense tracking and mapping in real-time. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2320–2327, Nov 2011. 4

[15] Project Website. http://grail.cs.washington.edu/projects/timelapse3d. 1, 3, 4, 6, 8

[16] M. Rubinstein, C. Liu, P. Sand, F. Durand, and W. T. Freeman. Motion denoising with application to time-lapse photography. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition*, CVPR '11, pages 313–320, Washington, DC, USA, 2011. IEEE Computer Society. 2

[17] G. Schindler and F. Dellaert. Probabilistic temporal inference on reconstructed 3d scenes. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 1410–1417. IEEE, 2010. 2

[18] G. Schindler, F. Dellaert, and S. B. Kang. Inferring temporal order of images from 3d structure. In *Computer Vision and Pattern Recognition, 2007. CVPR '07. IEEE Conference on*, pages 1–7, June 2007. 2

[19] S. Seitz, B. Curless, J. Diebel, D. Scharstein, and R. Szeliski. A comparison and evaluation of multi-view stereo reconstruction algorithms. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, volume 1, pages 519–528, June 2006. 2

[20] Q. Shan, R. Adams, B. Curless, Y. Furukawa, and S. Seitz. The visual turing test for scene reconstruction. In *3D Vision - 3DV 2013, 2013 International Conference on*, pages 25–32, June 2013. 2

[21] I. Simon, N. Snavely, and S. Seitz. Scene summarization for online image collections. In *Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on*, pages 1–8, Oct 2007. 2

[22] N. Snavely, R. Garg, S. M. Seitz, and R. Szeliski. Finding paths through the world's photos. *ACM Transactions on Graphics (Proceedings of SIGGRAPH 2008)*, 27(3):11–21, 2008. 2

[23] G. Zhang, J. Jia, T.-T. Wong, and H. Bao. Consistent depth maps recovery from a video sequence. *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, 31(6):974–988, June 2009. 1, 2, 3, 4

[24] L. Zhang, B. Curless, and S. Seitz. Spacetime stereo: shape recovery for dynamic scenes. In *Computer Vision and Pattern Recognition, 2003. Proceedings. 2003 IEEE Computer Society Conference on*, volume 2, pages II–367–74 vol.2, June 2003. 2

[25] K. C. Zheng, A. Colburn, A. Agarwala, M. Agrawala, D. Salesin, B. Curless, and M. F. Cohen. Parallax photography: Creating 3d cinematic effects from stills. In *Proceedings of Graphics Interface 2009*, GI '09, pages 111–118, Toronto, Ont., Canada, Canada, 2009. Canadian Information Processing Society. 2, 7