Measuring the Effects of Temporal Coherence in Depth Estimation for Dynamic Scenes

Iraklis Tsekourakis
Stevens Institute of Technology
Hoboken, NJ, 07030, USA
iteskour@stevens.edu

Philippos Mordohai
Stevens Institute of Technology
Hoboken, NJ, 07030, USA
Philippos.Mordohai@stevens.edu

Abstract

This paper presents a new algorithm for enforcing temporal coherence on depth estimation from multi-view videos of dynamic scenes as well as the first substantial quantitative evaluation of the improvement in depth estimation accuracy due to temporal coherence. The proposed algorithm is generally applicable and practical since it bypasses explicit scene flow estimation, which has a very large state space, and relies only on optical flow which is used to impose soft constraints on depth estimation for the next frame. As a result, our algorithm is applicable to scenes with large depth and motion ranges. The output is a sequence of depth maps that can be used for novel view synthesis among other applications. While it is intuitive that enforcing temporal coherence should improve the accuracy of depth estimation, this improvement has never been assessed quantitatively due to the lack of data with ground truth. To overcome this limitation we use the image prediction error as the criterion and show that the benefits of temporal coherence are significant on a diverse set of multi-view video sequences.

1. Introduction

3D reconstruction is one of the most studied problems in computer vision. Impressive 3D models can now be obtained from large collections of images of a static scene. The work of Shan et al. [1] in particular has issued a challenge to the research community to generate photo-realistic models that can appear indistinguishable from actual photographs. Even though the current top-performing methods have not yet passed this Turing test, one could foresee this day coming soon. The next frontier is achieving the same level of quality for dynamic scenes, which pose significant new challenges to 3D reconstruction algorithms.

The goal of this paper is to make progress towards free-viewpoint video generation for dynamic scenes. Free-viewpoint video is a technology that allows the viewer of a multi-camera video to control a virtual camera and generate videos from novel viewpoints by combining all available images. Interactive performance is often required, but in this paper we are concerned with accuracy first. One path that leads to free-viewpoint video is to ensure that the 3D reconstruction at each time instant, using only images taken simultaneously, is perfect. The work of Collet et al [2] is a step in that direction that demonstrates qualitatively outstanding results. However, each frame is reconstructed independently. A more realistic approach is to accept that perfection is hard to achieve and attempt to improve the 3D reconstruction of a given frame by leveraging information from previous frames. We, thus, aim to estimate temporally coherent depth given synchronized videos captured by multiple calibrated cameras.

This problem is related to *scene flow* estimation [3], which is the 3D equivalent of optical flow. As long as resampling in the temporal dimension is not required, i.e. novel viewpoint synthesis is restricted to the spatial domain, estimating depth only is sufficient. By not having to assign specific 3D velocities to pixels, we can impose soft constraints on depth optimization favoring temporal coherence with previous frames without having to make hard decisions on pixel-to-pixel correspondences in time. On the other hand, we assign a specific depth to each pixel, making hard decisions about correspondences in space.

Our approach uses Semi-Global Matching (SGM) [4] to optimize depth assignments according to a data term that comprises two parts: one due to spatial correspondence cost (photoconsistency) and one due to temporal constraints. The former is computed using the plane-sweeping algorithm [5], while optical flow is computed according to the method of Sun et el. [6]. The two parts of the data term are blended to form a single cost volume that favors the matches proposed by optical flow and to favor smoothness in disparity.

While it is intuitive that leveraging information from multiple frames should lead to improvements in accuracy,

Figure 1. First column: depth maps and corresponding renderings computed without temporal constraints. Second column: depth maps and corresponding renderings with temporal constraints for the ballet [10] and book arrival [11] datasets.

only qualitative results on dynamic scenes have been shown in the literature, with the exception of [7, 8]. Not surprisingly, the lack of appropriate datasets with ground truth is the cause for this shortcoming. Even though structured light or time of flight sensors registered with the cameras can be used to generate ground truth depth maps, but not scene flow, such datasets are not publicly available. The use of synthetic data [9] has recently emerged as a popular last resort. While there is some evidence that findings on synthetic data generalize to problems in the real world, we leave these experiments for future work.

In this paper we present the first comprehensive, quantitative evaluation of the improvement in multi-view 3D reconstruction of dynamic scenes due to temporal coherence constraints. The evaluation is conducted on a diverse set of multi-view videos [10, 12, 11] by excluding the frames of one camera from all computations and then rendering the colored depth map of the reference view to that camera to "predict" the actual image. The error metric we use is the average difference in RGB values between the predicted and actual image. While this metric may not accurately capture errors in textureless regions, it can clearly demonstrate when one depth map is superior to another. As advocated by [13, 14, 15], the ability to predict new views closely matches the requirements of many applications and does not require ground truth depth. Figure 1 shows the estimated depths computed by SGM, with and without temporal constraints and the corresponding renderings on the novel view used for evaluation. Both the depth maps and renderings are qualitatively and quantitatively better when temporal constraints are applied (see Section 7). Our algorithm improves reconstruction accuracy and reduces flickering artifacts in the videos.

In summary, the contributions of the paper are:

- a novel, generic and practical algorithm for temporally consistent depth estimation and
- the most extensive, to date, quantitative evaluation of the effects of temporal coherence on depth estimation accuracy.

2. Related Work

In this section, we focus on viewpoint-based methods that estimate depth for all pixels of the reference image. We then summarize the evaluation efforts presented in these publications. Approaches [16, 17] that reconstruct a single object, which has been segmented from the background, are not applicable to our inputs and are not covered here. Space-time stereo methods that operate in spatio-temporal volumes [18, 19, 20, 7, 8] require small frame-to-frame motion to be applicable. Our method, on the other hand, uses optical flow to detect long-range temporal matches.

Related to this paper is prior work that uses optical flow to improve disparity estimation by modifying the cost volume based on optical flow [21, 22, 23]. Larsen et al. [21] favor disparities that remain constant from frame to frame, while Bartczak et al. [22] allow the disparity to vary by one level between two consecutive frames. Yang et al. [23] explicitly segment the dynamic foreground from the static background and optimize them separately.

To reduce the number of parameters to be estimated and to regularize the solution, a common approach is to represent the scene as a collection of, typically planar, segments following parametric motion models. Vogel et al. [24] represent the scene with a set of rigidly moving planar patches and minimize an energy function that encompasses photoconsistency over multiple stereo pairs, image segmentation consistency and smoothness in 3D shape and motion.

Mustafa et al. [25] integrate a sparse-to-dense temporal correspondence technique with joint multi-view segmentation and reconstruction to obtain complete 4-D representations of static and dynamic objects.

Other authors do not rely on image segmentation and allow each pixel to have its own depth and 3D motion. Cech et al. [26] compute quasi-dense scene flow by growing spatial and temporal correspondence seeds. Variational approaches for joint estimation of all degrees of freedom have been published by [27, 28, 29]. Typically, the shape at time t_0 is initialized by stereo matching and then shape and motion are jointly estimated resulting in convergence to the nearest local minimum of an appropriate energy functional. To reduce computational complexity, some authors decouple disparity and motion estimation [30, 31].

Especially relevant to our work are methods that impose constraints on the next disparity map based on the current disparity and flow estimates [32, 33, 34]. Gong [32] computes the photoconsistency of all possible disparity flows per pixel, under a small motion assumption. The disparity values predicted by disparity flow are favored in the next frame by penalizing all other disparities. Liu and Philomin [33] employ a variational scene flow estimator [27] and use its output to predict the next disparity map and to impose soft constraints on disparity estimation for that frame. Min et al. [34] modify the cost function to enforce temporal coherence and use a frame similarity function to determine the influence of the modification. These methods, however, incur the high computational cost of scene flow estimation.

Evaluation The majority of the above publications do not include quantitative evaluation except on synthetic inputs, or they present results after applying the algorithms on static scenes and grouping the images in sets that are assumed to be acquired at different times. Overall, the effects of enforcing temporal coherence on the 3D reconstruction on *real* dynamic scenes compared to reconstructing sequences of 3D models ignoring temporal information have not been measured. Qualitative results on real data are presented by [27, 21, 17, 22, 33, 28, 26, 29, 30, 31, 23], while [34] also includes qualitative results on novel view synthesis. Even recent methods for scene flow on RGB-D sequences captured by depth cameras only present qualitative evaluation [35, 36, 37, 38, 39, 40].

Several publications [27, 33, 28, 38, 41] present results on the multi-baseline Middlebury data [42]. The algorithms still attempt to estimate flow in the vertical direction and in disparity even though they are always zero. The most extensive evaluation on rigid scenes is presented by [24] on the KITTI benchmark [43], which, however, does not contain independently moving objects. Menze and Geiger [44] proposed a joint depth and scene flow estimation method and a new dataset for evaluation. It assumes a finite number of rigidly moving objects in the scene, while our method can handle non-rigidity.

In the absence of ground truth data, Furukawa and Ponce [45] concatenated forward and reverse videos around a common frame, creating sequences such as $f_1 f_2 f_3 f_2 f_1$, and then measured the consistency of scene flow estimates between the same pairs of frames that appear in reverse order, such as $f_1 f_2$ and $f_2 f_1$. Ideally, shape estimates should be identical and scene flow vectors should have the same magnitude but opposite orientation. We did not adopt this technique here since it can fail to detect errors such as those due to excessive smoothness of the estimated flow.

A small scale quantitative evaluation was conducted by Popham et al. [46] who measured the accuracy of scene flow estimation over long sequences (90 frames) on a small number of points manually clicked on the images. Clearly, this approach does not scale well and also suffers from selection bias. The most thorough evaluations were published by Sizintsev and Wildes [7, 8]. Ground truth is acquired using structured light sensors on stop motion sequences in a motorized stage. While this study is unprecedented and valuable, the experimental setup is not ideal since several fiducial markers had to be placed on each independently moving surface to aid ground truth generation. This also improves the accuracy of the algorithms being evaluated, not only on the markers themselves, but also on nearby pixels that are affected via regularization.

3. Problem Statement

In this paper we address the estimation of temporally coherent depth maps from multiple synchronized and calibrated video sequences of a scene. To this end, we combine outputs of depth estimation for sets of images taken at the same time (*spatial correspondences*) with frame-to-frame optical flow computed for the reference camera (*temporal correspondences*). Depth estimation is carried out in two stages: plane-sweeping stereo for generating the cost or likelihood volume and SGM for extracting the final depth estimates from the cost volume.

Temporal smoothness constraints are based on optical flow computation. Since optical flow estimation is not perfect in practice, we do not implement temporal matches as hard constraints, but we blend them into the cost volume encoding our preference for depths at time $t + 1$ that are consistent with the optical flow results. The implementation is presented in Section 5.

Since no datasets with ground truth depth over entire video sequences are publicly available, we evaluate the accuracy of the estimated depth maps based on the quality of synthesized views generated from them. We perform these evaluations by excluding the frames of one camera from all computations and then rendering the colored depth map of the reference view to that camera to "predict" the actual im-

age. The error metric we use is the average difference in RGB values between the predicted and actual image. While this metric may not accurately capture errors in textureless regions, it is suitable for free viewpoint video. The evaluation methodology and experimental results are presented in Sections 6 and 7, respectively.

4. Multi-baseline Semi-Global Matching

In this section, we present depth map estimation for the single-frame case, before temporal constraints are applied. Our approach combines the plane-sweeping algorithm [5] with Semi-Global Matching (SGM) optimization [4]. The former is used for computing the likelihood of a number of possible depths for each pixel of the reference view. Since plane-sweeping does not require the images to be rectified, it is very convenient for multi-view matching. SGM is used for obtaining a depth map that approximately optimizes an energy function considering both fidelity to the matching likelihoods and smoothness. We use the rSGM implementation provided by Spangenberg et al. [47]. For each dataset, we select one camera as the reference view and compute depth for its pixels using frames from other cameras to compute the matching likelihood.

In plane sweeping stereo we define a family of planes parallel to the image plane of the *reference view*. For each pixel (x, y), depth hypotheses are formed by intersecting the corresponding ray with the set of planes. We then define a square window centered at (x, y) in the reference view and warp it to the *target views* using the homographies from the reference view to the target views through the current plane. We compute the normalized cross-correlation (NCC) between the window on the reference view and each warped window on the target views, and store the average as the likelihood of assigning to the pixel the depth corresponding to the current plane. Target images in which the matching window falls out of bounds are excluded. The output of this stage is a likelihood volume that contains the average NCC for assigning plane index d to pixel (x, y). It is converted to a cost volume $C(x, y, d)$ by negating the NCC scores.

SGM approximately optimizes a global two-dimensional energy function by combining 1D minimization problems in multiple directions. We use eight paths for dynamic programming and 256 discretized depths per pixel. The energy of a depth map D has the form of a summation of a data cost for assigning depth d_p to pixel p and smoothness costs that penalize depth discontinuities.

$$E(D) = \sum_p \{C(p, d_p) + \sum_{q \in N_p} P_1 T[|d_p - d_q| = 1] + \sum_{q \in N_p} P_2 T[|d_p - d_q| > 1]\}. \quad (1)$$

P_1 is the penalty added to the energy function of a pixel p for pixels q in the 1D neighborhood N_p of p, for which we observe a depth change equal to one discrete level, which may be due to slanted or curved surfaces. P_2 is the penalty for all depth changes greater than 1 ($P_2 > P_1$). $T[\cdot]$ is an indicator function which is 1 when its argument is true. In the implementation of [47], P_2 is defined adaptively based on the intensity values I of pixels p and q:

$$P_2(p) = max\{\gamma - \alpha \cdot |I(p) - I(q)|, P_{2,min}\} \quad (2)$$

where $P_{2,min}$ is the minimum acceptable penalty value. Minimization is performed in each direction separately and the final cost for assigning a depth value to a pixel is obtained by adding the costs all paths that go through the pixel at that depth. The depth with the smallest total cost is selected. We then apply subpixel refinement and a 3×3 median filter as in [47].

5. Imposing Temporal Constraints

We assume that if a pixel (x_t, y_t) with depth d_t moves to coordinates (x_{t+1}, y_{t+1}) in the next frame, then the depth of (x_{t+1}, y_{t+1}) should be close to d_t. We further assume that d_{t+1} is normally distributed around d_t, as shown below. To estimate the optical flow between frames t and $t + 1$ of the reference camera, we use the software of Sun et al. [6]. Figure 2 shows an example of two consecutive frames. A visualization of the estimated flow is presented on the second row, using the Middlebury color coding [48] on the left and the vector plot on the right.

The key assumption is that the depth of temporal matches in the next frame d_{t+1} follows a probability distribution P that is centered around the depth of the previous

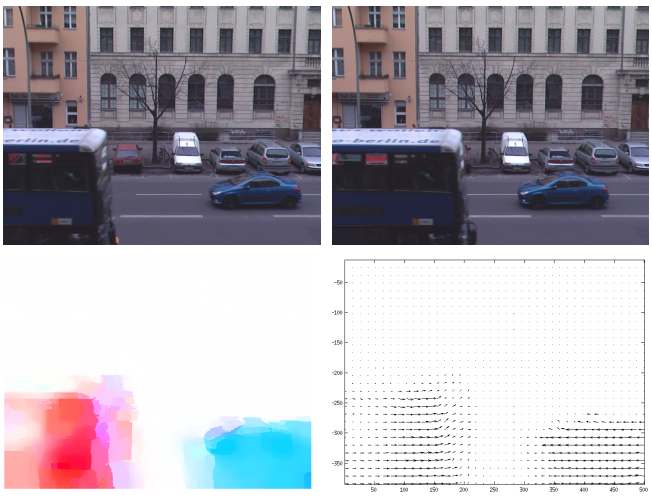


Figure 2. First row: Input images at time instance t (left) and $t+1$ (right). Second Row: Visualization of flow with Middlebury color coding [48], and vector plot.

frame d_t at pixel (x_t, y_t) which corresponds to (x_{t+1}, y_{t+1}) according to optical flow. This idea was inspired by the work of Unger et al. [49] who presented a probabilistic depth map fusion algorithm for static scenes by approximating the distribution of projected depths on the reference view from projection uncertainties. Here, we use optical flow to establish matches. We assume that the likelihood L of the depth at time $t+1$ is maximum at the plane $d_{t+1} = d_t$ and it decreases with increasing distance in depth.

$$L(d_{t+1}, x_{t+1}, y_{t+1} | d_t, x_t, y_t) = \frac{1}{A} exp\left(- \frac{(d_{t+1} - d_t)^2}{2\sigma_d^2} \right) \quad (3)$$

with σ_d set to one disparity value. Disparity is defined as $b_{max} f/d$, where b_{max} is the maximum baseline between the reference and a target view used in plane sweeping, and f is the focal length of the reference camera. A controls the relative weight of the temporal constraints compared to the data term. This formulation allows us to apply soft temporal constraints on depth estimation at time $t + 1$ by blending them into the cost volume C_{t+1}. This is accomplished by subtracting the likelihood L from the corresponding cost values, which is equivalent to adding the likelihood before the NCC is negated.

The cost volume C'_{t+1} is updated based on the optical flow OF_t from t to $t + 1$, the initial depth map of the previous frame D_t and the current cost volume C_{t+1} as inputs.

$$\begin{aligned} C'_{t+1}(x_{t+1}, y_{t+1}, d_{t+1}) &= C_{t+1}(x_{t+1}, y_{t+1}, d_{t+1}) \\ &- P(d_{t+1}, x_{t+1}, y_{t+1} | d_t, x_t, y_t), \forall (x_t, y_t) \end{aligned} \quad (4)$$

The updated depth map D'_{t+1} is computed from C'_{t+1} using SGM for all frames. We call this process temporal coherence constraint with a time horizon of one frame.

We also applied this method on longer time horizons by using the updated depths of the previous frame d'_t instead of d_t in Eq. 4. This led to the *unlimited horizon temporal coherence constraint*. As expected, the unlimited horizon constraint in some cases results in propagation of errors or blending of surfaces in the updated depth assignments. In Section 7 we evaluated different time horizons by not allowing temporal constraints to persist longer than a given number of frames.

6. Evaluation Methodology

In the absence of ground truth, we use view prediction errors [13, 15, 50, 51, 14, 52, 53] to evaluate the generated depth maps. In all cases, we use a completely separate *validation camera* for evaluation and entirely exclude its frames from depth estimation. We always choose an *extrapolating view* for validation so that errors are more pronounced in it. An interpolating view, according to Szeliski [13], is one that lies between views used in the computation, while an extrapolating view is beyond the set of target and reference views. Clearly, synthesizing extrapolating views is more challenging, since the sensitivity to depth errors increases as the viewpoint of the validation camera moves away from the reference camera.

In recent work, Waechter et al. [14] present an extensive analysis of novel view prediction error as a measurement of the accuracy of 3-D reconstructed models. In our context, it has two main advantages. First, it allows the use of datasets without ground truth depth for evaluation. This is critical since no *real* datasets with ground truth exist. Second, it makes the comparison of different methods that use various scene representations feasible.

We synthesize views by projecting the colored depth map of the reference view, after subpixel refinement, to the validation view. If multiple projections fall onto a pixel, we keep the one nearest to the camera. Two types of errors occur after this process: pixels of the synthesized image may differ in RGB from those of the actual image and there may be no synthesized RGB values for some pixels. To avoid unnecessarily penalizing algorithms for pixels that cannot be predicted, we detect pixels of the validation camera that cannot be the projection of any point inside the frustum of the reference camera bounded by the minimum and maximum depth. These pixels are excluded from the evaluation. We then use the Manhattan distance in RGB over pixels that receive projections and we set the error to 256 per color channel for uncovered pixels. We also tried the 1-NCC error according to [14], but it has limitations in textureless areas.

7. Experimental Results

We evaluate our method using four publicly available, multi-view datasets, which were captured in widely different configurations. *Cheongsam* [12] is captured in a dome with a 4.2m diameter by twenty cameras arranged in a ring. Each video is 30 frames long. The *ballet* data [10] are acquired by eight cameras forming a 30 ° arc, thus with narrower baselines. The depth range is 7.6m and the length of the video is 100 frames. The *book arrival* and *outdoor* videos are provided by the MOBILE3DTV project [11]. They are captured using the same sixteen-camera rig, with the cameras mounted side-by-side parallel to each other. The maximum depth of the book arrival video and the outdoor video is 3.2m and 32m respectively and the length is 100 frames for both.

All experiments are performed with constant parameters for all parts of our method except for the number of target views in the plane-sweeping algorithm. We used two neighboring target views on each side of the reference view

for all datasets except for the Cheongsam dataset, where the wide angle between neighboring views forced us to use one target view on each side. NCC is computed in 5×5 windows over 256 fronto-parallel planes with subpixel spacing. For SGM we use 8 paths, $P_1 = 11$, $\alpha = 0.5$, $\gamma = 35$ and $P_{2,min} = 17$. For the temporal constraint computation, the parameters are $A = 10$ and $L_{min} = 0.01$. The latter is a threshold on L, below which we do not perform the cost volume updates according to Eq. (4) because they are negligible. A is the most important parameter as it controls the blending of the two cues that are combined in the updated cost volume.

Given a video, we computed cost volumes for every frame using plane-sweeping. We consider as a *baseline* the initial depth maps extracted by applying SGM on these cost volumes. Then, we generated depth maps by applying temporal constraints over all possible time horizons. For example, time horizon equal to 5 means that for the current time *t* we start updating the depths using the temporal constraints at *t*-5 and use the updated depths until we reach time *t*. Starting from a time horizon equal to 1 and increasing the value, we observe improvements until the horizon becomes equal to 3, where peak performance is observed. Beyond that, accuracy decreases reaching a minimum in most cases for a horizon close to 10. Accuracy then plateaus and stays approximately constant as the horizon reaches the length of the video. Therefore, only results for time horizons equal to 1 frame, 3 frames and unlimited (all previous) frames are shown.

	base	hor = 1	unlim. hor	hor = 3
book arrival	85.9	71.5	65.5	61.5
outdoor	26.6	21.7	24.2	20.2
Cheongsam	406.8	401.9	402.3	401.3
ballet	321.4	308.9	299.1	301.5

Table 1. Average RGB L1 distance of synthesized novel views compared to actual images. The average is taken over all pixels of all frames of a single reference camera. *Base* denotes the standard SGM algorithm without temporal constraints.

	base	hor = 1	unlim. hor	hor = 3
book arrival	90.4%	92.3%	93.3%	93.7%
outdoor	98.2%	98.8%	98.9%	99.0%
Cheongsam	49.6%	50.3%	50.2%	50.4%
ballet	59.7%	61.4%	62.7%	62.4%

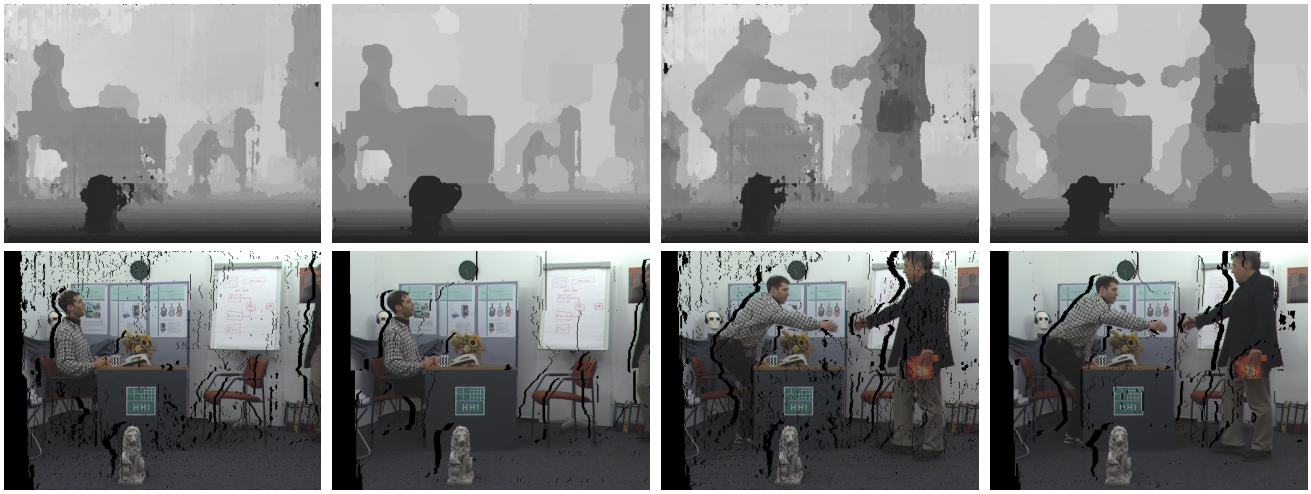
Table 2. Average percentage of pixels with valid projection, excluding the impossible ones. The average is taken over all frames of a single reference camera.

Tables 1, and 2 summarize the accuracy of all methods according to the criteria of Section 6. The best results were obtained with a time horizon of 3 frames in all videos except for the ballet video, where the unlimited horizon performed slightly better due to the textureless background. Cheongsam and ballet have less coverage due to their much larger baselines. The results for a time horizon of 1 frame and those with unlimited horizon show the range within which our solutions vary, but sensitivity is low. Once the time horizon reaches the upper single digits, the metrics become virtually constant. The optimal time horizon for a scene depends on factors such as the fraction of pixels that is occluded or unoccluded in each frame and the velocity of the surfaces. The improvement due to temporal constraints is smaller for the Cheongsam data. This is because the initial depth estimates are more accurate compared to the rest of the datasets. The variation of the average RGB differences presented in Table 1 is explained by the completely different configuration of the cameras used in each video (angle and distance between cameras as well as depth range). This determines the degree of occlusion from the reference view to the novel view. Figure 3 illustrates the qualitative improvement achieved by the temporal constraints. We observe improvements in the reconstruction of both the stationary background and moving foreground objects. It is worth mentioning here that in every single frame of all videos tested, the novel view generated with temporal constraints is superior to the baseline in terms of both L1 RGB distance from the actual image, and also includes a larger percentage of pixels with valid projection.

8. Conclusions

We have presented a general algorithm for improving the temporal coherence of depth estimation for dynamic scenes and demonstrated significant quantitative and qualitative improvements. While this finding is not unexpected, this type of study was missing from the literature. We are optimistic that our algorithm will be adopted by the research community because it is compatible with all discrete optimization methods and only has one parameter to be tuned (A in Eq. 3). Unlike [32, 33, 34], our approach does not require scene flow estimation to impose temporal coherence.

The quantitative results of Section 7 show significant overall improvements due to temporal coherence. Moreover, novel views generated with temporal constraints are *always* superior in terms of our metrics. Depending on data-specific factors, the average improvement can be as high as 30%, as in the book arrival sequence. These factors include camera configuration, i.e. the angle between the reference and validation view and the baseline, as well as the frequency content of the images, which determines the sensitivity in terms of the synthesis of novel RGB values. Analyzing these effects is an interesting future direction.



Book arrival

Ballet

3DV Outdoor

Figure 3. First and third columns: depths and corresponding novel view projections without temporal constraints. Second and fourth columns: depths and corresponding novel view projections with temporal constraints.

Acknowledgments This research has been supported in part by the National Science Foundation under awards IIS-1217797, IIS-1527294 and IIS-1637761.

References

[1] Q. Shan, R. Adams, B. Curless, Y. Furukawa, and S. M. Seitz, "The visual turing test for scene reconstruction," in *3DV*, 2013, pp. 25–32.

[2] A. Collet, M. Chuang, P. Sweeney, D. Gillett, D. Evseev, D. Calabrese, H. Hoppe, A. Kirk, and S. Sullivan, "High-quality streamable free-viewpoint video," *ACM Transactions on Graphics (TOG)*, vol. 34, no. 4, p. 69, 2015.

[3] S. Vedula, S. Baker, P. Rander, R. T. Collins, and T. Kanade, "Three-dimensional scene flow," *PAMI*, vol. 27, no. 3, pp. 475–480, 2005.

[4] H. Hirschmüller, "Stereo processing by semiglobal matching and mutual information," *PAMI*, vol. 30, no. 2, pp. 328–341, 2008.

[5] D. Gallup, J. M. Frahm, P. Mordohai, Q. Yang, and M. Pollefeys, "Real-time plane-sweeping stereo with multiple sweeping directions," in *CVPR*, 2007.

[6] D. Sun, S. Roth, and M. Black., "Secrets of optical flow estimation and their principles," in *CVPR*, 2010.

[7] M. Sizintsev and R. Wildes, "Spatiotemporal stereo and scene flow via stequel matching," *PAMI*, vol. 34, no. 6, pp. 1206–1219, 2012.

[8] M. Sizintsev and R. P. Wildes, "Spacetime stereo and 3d flow via binocular spatiotemporal orientation analysis," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 36, pp. 2241–2254, 2014.

[9] N. Mayer, E. Ilg, P. Häusser, P. Fischer, D. Cremers, A. Dosovitskiy, and T. Brox, "A large dataset to train convolutional networks for disparity, optical flow, and scene flow estimation," in *CVPR*, 2016.

[10] C. L. Zitnick, S. B. Kang, M. Uyttendaele, S. Winder, and R. S. Szeliski, "High-quality video view interpolation using a layered representation," *ACM Trans. on Graphics*, vol. 23, no. 3, pp. 600–608, 2004.

[11] A. Smolic, G. Tech, and H. Brust, "Report on generation of stereo video data base," in *Technical Report D2.1 v3.0*, 2010.

[12] Y. Liu, Q. Dai, and W. Xu, "A point cloud based multi-view stereo algorithm for free-viewpoint video," *IEEE Trans. on Visualization and Computer Graphics*, vol. 16, no. 3, pp. 407–41, 2010.

[13] R. Szeliski, "Prediction error as a quality metric for motion and stereo," in *ICCV*, 1999, pp. 781–788.

[14] M. Waechter, M. Beljan, S. Fuhrmann, N. Moehrle, J. Kopf, and M. Goesele, "Virtual rephotography: Novel view prediction error for 3d reconstruction," *ACM Trans. on Graphics*, vol. 36, no. 1, pp. 8:1–8:11, 2017.

[15] J. Kilner, J. Starck, J. Y. Guillemaut, and A. Hilton, "Objective quality assessment in free-viewpoint video production," *Signal Processing: Image Communication*, vol. 24, no. 1-2, pp. 3 – 16, 2009.

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

[17] J. P. Pons, R. Keriven, and O. D. Faugeras, "Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score," *IJCV*, vol. 72, no. 2, pp. 179–193, 2007.

[18] L. Zhang, B. Curless, and S. M. Seitz, "Spacetime stereo: shape recovery for dynamic scenes," in *CVPR*, 2003.

[19] J. Davis, D. Nehab, R. Ramamoorthi, and S. Rusinkiewicz, "Spacetime stereo: A unifying framework for depth from triangulation," *PAMI*, vol. 27, no. 2, pp. 296–302, 2005.

[20] C. Richardt, D. Orr, I. Davies, A. Criminisi, and N. Dodgson, "Real-time spatiotemporal stereo matching using the dual-cross-bilateral grid," in *ECCV*, 2010, pp. 510–523.

[21] E. S. Larsen, P. Mordohai, M. Pollefeys, and H. Fuchs, "Temporally consistent reconstruction from multiple video streams using enhanced belief propagation," in *ICCV*, 2007.

[22] B. Bartczak, D. Jung, and R. Koch, "Real-time neighborhood based disparity estimation incorporating temporal evidence," in *DAGM*, 2008, pp. 153–162.

[23] W. Yang, G. Zhang, H. Bao, J. Kim, and H. Y. Lee, "Consistent depth maps recovery from a trinocular video sequence," in *CVPR*, 2012.

[24] C. Vogel, K. Schindler, and S. Roth, "3D scene flow estimation with a piecewise rigid scene model," *IJCV*, vol. 115, no. 1, pp. 1–28, 2015.

[25] A. Mustafa, H. Kim, J.-Y. Guillemaut, and A. Hilton, "Temporally coherent 4d reconstruction of complex dynamic scenes," in *CVPR*, 2016.

[26] J. Cech, J. Sanchez-Riera, and R. Horaud, "Scene flow estimation by growing correspondence seeds," in *CVPR*, 2011.

[27] F. Huguet and F. Devernay, "A variational method for scene flow estimation from stereo sequences," in *ICCV*, 2007.

[28] T. Basha, Y. Moses, and N. Kiryati, "Multi-view scene flow estimation: A view centered variational approach," in *CVPR*, 2010.

[29] T. Müller, J. Rannacher, C. Rabe, and U. Franke, "Feature- and depth-supported modified total variation optical flow for 3d motion field estimation in real scenes," in *CVPR*, 2011.

[30] A. Wedel, T. Brox, T. Vaudrey, C. Rabe, U. Franke, and D. Cremers, "Stereoscopic scene flow computation for 3d motion understanding," *IJCV*, vol. 95, pp. 29–51, 2011.

[31] J. Park, T. H. Oh, J. Jung, Y.-W. Tai, and I. S. Kweon, "A tensor voting approach for multi-view 3d scene flow estimation and refinement," in *ECCV*, 2012, pp. 288–302.

[32] M. Gong, "Real-time joint disparity and disparity flow estimation on programmable graphics hardware," *CVIU*, vol. 113, no. 1, pp. 90 – 100, 2009.

[33] F. Liu and V. Philomin, "Disparity estimation in stereo sequences using scene flow," in *BMVC*, 2009.

[34] D. Min, S. Yea, and A. Vetro, "Temporally consistent stereo matching using coherence function," in *3DTV-Conference*, 2010.

[35] A. Letouzey, B. Petit, E. Boyer, and M. Team, "Scene flow from depth and color images." in *BMVC*, 2011.

[36] E. Herbst, X. Ren, and D. Fox, "RGB-D flow: Dense 3-D motion estimation using color and depth," in *ICRA*, 2013, pp. 2276–2282.

[37] M. Keller, D. Lefloch, M. Lambers, S. Izadi, T. Weyrich, and A. Kolb, "Real-time 3d reconstruction in dynamic scenes using point-based fusion," in *3DV*, 2013.

[38] D. Ferstl, C. Reinbacher, G. Riegler, M. Rüther, and H. Bischof, "aTGV-SF: Dense variational scene flow through projective warping and higher order regularization," in *3DV*, 2014.

[39] R. A. Newcombe, D. Fox, and S. M. Seitz, "Dynamicfusion: Reconstruction and tracking of non-rigid scenes in real-time," in *CVPR*, 2015, pp. 343–352.

[40] M. Dou, J. Taylor, H. Fuchs, A. Fitzgibbon, and S. Izadi, "3d scanning deformable objects with a single RGBD sensor," in *CVPR*, 2015, pp. 493–501.

[41] D. Sun, E. B. Sudderth, and H. Pfister, "Layered RGBD scene flow estimation," in *CVPR*, 2015, pp. 548–556.

[42] D. Scharstein and R. Szeliski, "High-accuracy stereo depth maps using structured light," in *CVPR*, 2003.

[43] A. Geiger, P. Lenz, C. Stiller, and R. Urtasun, "Vision meets robotics: The kitti dataset," *The International Journal of Robotics Research*, vol. 32, no. 11, pp. 1231–1237, 2013.

[44] M. Menze and A. Geiger, "Object scene flow for autonomous vehicles," in *CVPR*, 2015, pp. 3061–3070.

[45] Y. Furukawa and J. Ponce, "Dense 3D motion capture from synchronized video streams," in *CVPR*, 2008.

[46] T. Popham, A. Bhalerao, and R. Wilson, "Multi-frame scene-flow estimation using a patch model and smooth motion prior," in *BMVC Workshop*, 2010.

[47] R. Spangenberg, T. Langner, S. Adfeldt, and R. Rojas, "Large scale semi-global matching on the cpu," in *IEEE Intelligent Vehicles Symposium*, 2014, pp. 195–201.

[48] S. Baker, D. Scharstein, J. Lewis, S. Roth, M. Black, and R. Szeliski, "A database and evaluation methodology for optical flow," *IJCV*, vol. 92, no. 1, pp. 1–31, 2011.

[49] C. Unger, E. Wahl, P. Sturm, and S. Ilic, "Stereo fusion from multiple viewpoints," in *Joint 34th DAGM and 36th OAGM Symposium*, 2012, pp. 468–477.

[50] P. Mordohai, "On the evaluation of scene flow estimation," in *Workshop on Unsolved Problems in Optical Flow and Stereo Estimation*, 2012.

[51] J. Flynn, I. Neulander, J. Philbin, and N. Snavely, "Deepstereo: Learning to predict new views from the world's imagery," in *CVPR*, 2016.

[52] T. Taniai, S. N. Sinha, and Y. Sato, "Fast multi-frame stereo scene flow with motion segmentation," in *CVPR*, 2017.

[53] I. Tsekourakis and P. Mordohai, "A comparison of scene flow estimation paradigms," in *RFMI*, 2017.