Temporally Consistent Reconstruction from Multiple Video Streams Using Enhanced Belief Propagation

E. Scott Larsen
Google, Inc
Pittsburgh, Pennsylvania, USA
esl@google.com

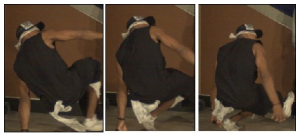
Philippos Mordohai, Marc Polleyfeys, and Henry Fuchs
University of North Carolina at Chapel Hill
Chapel Hill, North Carolina, USA
[mordohai, marc, fuchs]@cs.unc.edu

Abstract

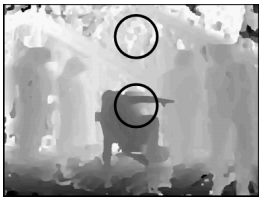
We present an approach for 3D reconstruction from multiple video streams taken by static, synchronized and calibrated cameras that is capable of enforcing temporal consistency on the reconstruction of successive frames. Our goal is to improve the quality of the reconstruction by finding corresponding pixels in subsequent frames of the same camera using optical flow, and also to at least maintain the quality of the single time-frame reconstruction when these correspondences are wrong or cannot be found. This allows us to process scenes with fast motion, occlusions and self-occlusions where optical flow fails for large numbers of pixels. To this end, we modify the belief propagation algorithm to operate on a 3D graph that includes both spatial and temporal neighbors and to be able to discard messages from outlying neighbors. We also propose methods for introducing a bias and for suppressing noise typically observed in uniform regions. The bias encapsulates information about the background and aids in achieving a temporally consistent reconstruction and in the mitigation of occlusion related errors. We present results on publicly available real video sequences. We also present quantitative comparisons with results obtained by other researchers.

1. Introduction

Multiple-view reconstruction has been one of the most active areas in computer vision. Considerable progress has been made and the reconstructed models are within millimeters from the ground truth. Examples of such accurate reconstructions and a comparison among them can be found at the Multi-View Stereo Evaluation webpage[4]. While 3D static reconstruction is widely useful, the domain of dynamic reconstructions is much larger. It includes applications such as visualization for training, consultation, entertainment, free-viewpoint video and 3D TV. Besides temporal consistency, we aim at achieving two other goals: geometric accuracy and rendering from viewpoints beyond the set of the input camera viewpoints. While good visualization results can be achieved using video-based rendering approaches that process each time-frame[1] independently, such as the one by Zitnick et al. [23], they do not always satisfy the other two requirements, namely geometric accuracy and the ability to visualize the reconstruction from a viewpoint that is very different than the input cameras.



(a) Consecutive video frames



(b) Depthmap using standard BP-based stereo

Figure 1. Parts of three consecutive frames from the MSR breakdancing dataset [23] and a depthmap reconstructed using standard BP. Notice the large motions from frame to frame and the highlighted limitations of single time-frame reconstructions, such as the "webbing" under the breakdancer's arm and the noise on the wall.

Methods for estimating what is termed *3D scene flow* have been reported in the literature [17, 18, 5, 9, 22, 6, 7] but they typically require reasonable estimates of the optical flow in the images for most of the pixels including the ones on moving parts. Our approach can utilize optical flow measurements when they are correct, but is also robust to large errors that typically occur when there is large motion in the scene (Fig. 1(a)). The consequences of large motion are that pixel correspondences cannot be detected due to large displacements and rotations, as well as that many pixels become occluded and un-occluded and thus have no optical

[1]The term "time-frame" is used in this paper to refer to a set of images captured at the same time

flow. A fundamental goal of our approach is that it should not achieve inferior performance compared to single-frame reconstruction when optical flow fails.

Processing is performed on video streams captured by static, synchronized and calibrated cameras. The objective is a sequence of depthmaps for the reference camera that are geometrically accurate and temporally consistent. We begin by constructing the initial observation for the pixels of the reference camera using a multi-camera implementation of the locally adaptive support-weight method of Yoon and Kweon [21]. The final depth estimates are obtained using a stereo algorithm based on enhanced belief propagation (BP) that operates on the observations for multiple frames. In Section 3 we describe in more detail the following enhancements to address temporal reconstruction and errors in the estimated depthmaps such as those in Fig. 1(b):

- *Robust BP* which dynamically adapts the graph by removing outliers among the neighbors of each node.
- *Biased BP* which uses an additional *PDF* similar to the observation to influence pixels, those similar in color to the background and having low confidence, towards a bias from a higher level algorithm. This bias can be obtained many ways, including by segmentation and plane-fitting or by constructing a colored model of the background depth after observing all the frames. The strength of the bias is adjusted locally based on color similarity and confidence.
- *Quiet BP* which addresses noise propagation in uniform areas by not re-using the initial observation during message updates after initialization. Its objective is to mitigate the negative effects of loops in belief propagation.

In Section 5, we present results on the binocular stereo datasets of the Middlebury Stereo Evaluation webpage [14, 3], one of the 3D TV Network of Excellence datasets available at [1] and the Microsoft Research breakdancing sequence available at [2]. We also present quantitative comparisons with reconstructions of the latter dataset obtained by Zitnick et al. [23]. We are grateful to the authors for making their results publicly available. We acknowledge that the focus of their research was not the geometric accuracy of the reconstruction, but rather multi-camera video-based view interpolation that enables free-viewpoint video playback. There are, however, useful conclusions that can be drawn from this comparison.

2. Related Work

This section is a brief overview of multiple-view reconstruction algorithms that explicitly address temporal consistency for non-rigid scenes. See the Middlebury Multi-View Stereo Evaluation webpage [4] and references therein for work on the static case.

Vedula et al. [17, 18] introduced scene flow which is the 3D equivalent of optical flow: a dense motion field for all surface points from frame to frame. A key choice is to perform regularization in the images instead of the scene surfaces. A different approach was taken in [19] in which space carving [8] is used to carve away voxels that violate photoconsistency constraints in the 6D space of the coordinates of two temporally corresponding voxels. This results in tighter shape approximation than applying space carving to the two frames separately.

Carceroni and Kutulakos [5] use dynamic surfels as primitives that encode local shape, reflectance and motion of a small surface patch. In contrast to [18], regularization takes place on the surface. The approach is limited to a small number of surfels due to the high complexity of the proposed surfel sampling algorithm and requires known illumination of the scene. Neumann and Aloimonos [9] employ a time-varying multi-resolution surface that is fitted to the data using spatio-temporal multiple-view information and silhouette constraints. The subdivision, unlike [5], can be updated from frame to frame.

Pons et al. [12] propose a common variational framework for multiple-view reconstruction and scene flow estimation. The solution is obtained via a level set that evolves the shape in order to minimize image prediction error. Scene flow is estimated in the 3D domain where a vector field is evolved according to the same image prediction criterion. Goldluecke and Magnor [6] extend the iso-surface extraction approach to temporal reconstructions and seek a 3D iso-surface in 4D using a variational method. The desired iso-surface represents the evolution of the scene surfaces in time.

Tao et al. [16] present a method based on image segmentation that models the scene as a set of planes. It is well suited for scenes with uniform surfaces where optical flow is correct for segments but not pixels. Temporal depth hypotheses for each plane are verified in the following frame. The method is able to produce sharp surface boundaries due to image segmentation and the fitting of a plane to each segment. We achieve similar results by introducing a bias in belief propagation.

Zhang and Kambhamettu [22] present two viewpoint-based approaches for the estimation of 3D scene flow. One is based on a piecewise affine motion model and operates on N frames assuming constant velocity for each patch. The second approach computes structure and motion simultaneously taking into account image segmentation information and using validation between two depthmaps to detect reliable matches. Gong [7] proposed an algorithm for the estimation of disparity flow, which is a motion field that maps pixels and disparities to the corresponding pixels and dis-

parity values in the next frame. The disparity flow fields are used to predict the new disparity maps by biasing the cost volume in favor of the predicted values.

A limitation of methods such as [18, 19, 6] is that they require accurate spatial and temporal correspondences for most points of the scene. This makes them inapplicable to scenes with very fast motions or with significant occlusions and self-occlusions. The surfel sampling algorithm [5] may be more effective when exact correspondences cannot be found but it is limited by the requirement for known illumination, its high computational complexity and the independent optimization of adjacent surfels. Our approach is more similar to [22] and [7] since it is viewpoint-based and most importantly able to recover from failures of optical flow.

3. Temporal Belief Propagation

In this section, we describe our modifications to the belief propagation framework. We begin with an overview of belief propagation, then discuss a simple extension to temporal belief propagation. This leaves us with unsatisfactory results though, so we provide two enhancements which improve the results significantly. These modifications address fundamental problems in temporal multiple-view stereo and are general: we expect them to be applicable to a variety of problems.

3.1. Belief Propagation Background

Before describing our research, we briefly present an overview of belief propagation [10, 20]. It is an optimization algorithm based on a message passing system that stores at each node a separate message for each of that node's neighbors. Belief propagation algorithms can be used for, among other things, optimization with an energy function for a pairwise Markov Random Field (MRF) as:

$$E(f) = -\sum_{i \in \mathcal{P}} \ln \phi_i(x_i) - \sum_{(i,j) \in \mathcal{N}} \ln \psi_{ij}(x_i, x_j) \quad (1)$$

This energy function is defined on a graph, with $\mathcal{P}$ the set of all nodes and $\mathcal{N}$ the set of node pairs, *i.e.* the set of edges. The set of nodes $j \in \mathcal{N}(i)$ with $(i,j) \in \mathcal{N}$ is termed the *neighborhood* of node i. Each node takes a labeling x_i, from some finite state space (*e.g.* disparity values). In these equations, $\phi_i(x_i)$ will be referred to as the *observation* - the data consistency term, and $\psi_{ij}(x_i, x_j)$ as the *compatibility* term - the smoothness term.

Keeping the observation at each node constant, the *belief* at a node i can be defined as:

$$b_i(x_i) = k\phi_i(x_i) \prod_{j \in \mathcal{N}(i)} m_{ji}(x_i), \quad (2)$$

where m_{ji} is the message from node j to node i about the state of i. The message m_{ji} is a vector of the dimensionality of the state space and its components are proportional to how likely the corresponding labeling (state) is for i according to j. Each message can be viewed as a discrete *PDF*. The messages are updated according to:

$$m_{ij}(x_j) \leftarrow \max_{x_i} \phi_i(x_i)\psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}(i) \setminus j} m_{ki}(x_i) \quad (3)$$

The product is over all messages coming to node i except the one coming from node j. Noted that BP offers no guarantees of optimality for graphs with loops, such as the ones used in computer vision problems. In the remainder of this section we discuss modifications to the BP framework that make it more effective for the problem of temporal multiple-view reconstruction.

3.2. Temporal Belief Propagation

Traditionally, the graph BP is performed on is the 2D image lattice with the neighborhood of a pixel being its 4-neighbors. This enforces spatial smoothness and has achieved outstanding results in binocular stereo with specific modifications to account for occlusion and other effects [15]. In order to integrate temporal smoothness, we need to define a graph in which pixels are also connected to their temporal neighbors. In this paper, we propose to use a graph consisting of three image lattices, one each for the previous, current and next frame of the reference camera. Each pixel is connected to its four image neighbors, one pixel in the previous frame and one in the next frame. This results in a graph with nodes of cardinality six on which the basic BP algorithm can be used without modifications.

Note that constructing the graph by simply connecting pixels with the same image coordinates across frames is incorrect when dealing with dynamic scenes. Therefore, we use optical flow to detect pixel correspondences in time. See Section 4 for more details. The challenge here is that optical flow often fails, especially for the parts of the scene with the fastest motion (Fig. 1(a)). Including these erroneous connections in the graph causes problems to the reconstruction.

Connecting nodes across frames implies that temporally corresponding pixels have nearly the same depth. Depending on the formulation of ϕ, "nearly the same" commonly includes cases where linearly interpolating the motion is correct - those approximated by constant first derivative. This assumption does not apply generally, *e.g.* when an object is accelerating toward the camera, but it tends to apply well in many common cases.

3.3. Robust Belief Propagation

To address the problems caused by propagating messages between nodes that do not belong to the same surface, we propose a scheme for dynamically adjusting the structure of the graph. It detects and removes outliers among the incoming messages based on variance reduction [11]. If a

message is inconsistent with the others, removing it results in a reduction of total per state variance that can be computed using all incoming messages.

$$R(m_{ij}) = \sum_k \sigma^k_{all} - \sigma^k_{all \setminus m_{ij}} \quad (4)$$

If the reduction $R(m_{ij})$ is positive, the edge between nodes i and j is removed from the graph and the node under consideration (j) remains connected with fewer neighbors. It is worth pointing out here that this process applies to all edges, not only the ones connecting temporal neighbors. Due to the small size of the neighborhood, we never remove more than two edges. Figure 2 shows the six messages for a point on the elbow of the breakdancer (Fig. 1(a)), where optical flow has failed. The two temporal neighbors are detected as outliers and removed.

Notice that it is common, *e.g.* in [15], to use pixel color similarities to weaken propagation across color edges. Our method correctly makes the implicit "same surface" decision in the presence of high frequencies in the image. In low-frequency portions of the image where discontinuities exist, *e.g.* a foreground and background object of the same color, our approach can again correctly detect discontinuities due to the differing PDF shapes. Finally, our approach updates the outlier decisions after every iteration since the beliefs change after propagation. Thus, edge decisions are refined as correct depths are converged upon.

3.4. Biased Belief Propagation

To address errors caused by occlusion and lack of texture, we propose an enhancement to belief propagation called Biased BP. The bias *PDF*, θ, acts as an additional observation and its purpose is to correct corrupted observations, which typically occur at occluded regions, while having minimal impact on nodes where the observations are reliable. Since the occluded regions cannot be easily handled by stereo, the bias acts as a higher-level prior. The construction of the bias is not critical for the algorithm. Two different options are presented in Section 4. The strength of the bias is modulated by a weight ω, which is set on a per-node basis. Its purpose is two-fold: to detect whether the color of the input image is similar to the background model and to decide whether the current confidence in the belief is low. The weights are defined according to:

$$\omega_i = e^{-\frac{\|I_i - B_i\|}{\gamma_c}} \times e^{-2 \max_{j \in \mathcal{N} \cup \phi_i} \{m_{ji}\}} \quad (5)$$

The first exponential term modulates the weight according to color similarity between the current image I_i and the background color $\mathcal{B}_i$. The second term examines the observation of node i as well as all incoming messages to it and finds the maximum element among them. If this is small, indicating no strong likelihood for any depth, the weight is large. No normalization is necessary here since beliefs and messages are normalized *PDF*s. The normalization parameter for the color, γ_c can be estimated from the bias model as discussed in Section 4.

The resulting energy equation and update rule become:

$$\begin{aligned} E(f) = & - \sum_{i \in \mathcal{P}} (\ln \phi_i(x_i) + \omega_i \ln \theta_i(x_i)) \\ & - \sum_{(i,j) \in \mathcal{N}} \ln \psi_{ij}(x_i, x_j) \end{aligned} \quad (6)$$

$$m_{ij}(x_j) \leftarrow \max_{x_i} \omega_i \theta_i(x_i) \phi_i(x_i) \psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}(i) \setminus j} m_{ki}(x_i) \quad (7)$$

3.5. Quiet Belief Propagation

As mentioned in the introduction of this section, BP is not optimal in the presence of loops in the graph. The loops reinforce the observation ϕ_i since it is a component in all outgoing messages, which, having traversed loops, return and reinforce the observation in future updates. We have shown in synthetic simulations, not included here due to lack of space, that strong signals do not propagate far in fields with uniform observations perturbed by a small amount of noise. This occurs because noisy nodes whose most likely states happen to agree reinforce each other and produce "alternative" strong beliefs. This effect is more pronounced as the "loopiness" of the graph, the number of cycles per node, increases. We have managed to address this problem by using the observation to initialize all messages, but then removing it from the update equation. The message initialization and update rules for quiet BP are:

$$m^0_{ij} x_j \leftarrow \phi_i(x_i), \quad (8)$$

$$m_{ij}(x_j) \leftarrow \max_{x_i} \psi_{ij}(x_i, x_j) \prod_{k \in \mathcal{N}(i) \setminus j} m_{ki}(x_i). \quad (9)$$

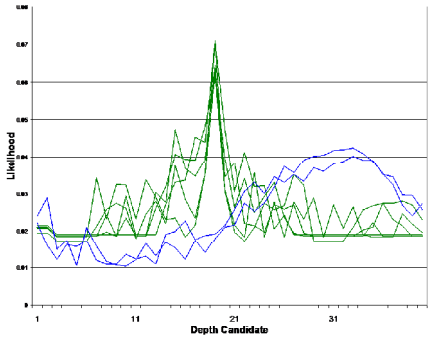


Figure 2. The six incoming messages for a node on the elbow of the breakdancer seen in Fig. 1(a). Since optical flow has failed, the two messages in blue come from background pixels and are inconsistent with the majority. They are detected using Eq. (4) and discarded.

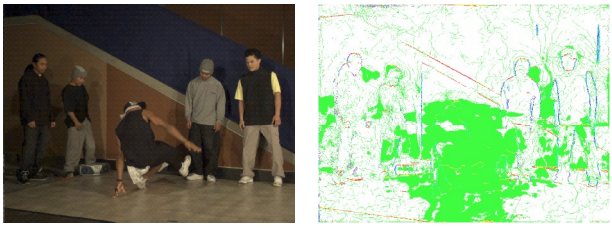
The effect of this approach is that noise in the observation does not act as a partial "filter" to future message updates. One could conceive that a drawback of this is that messages from far away that are different than the observation can come in and dominate if not filtered by the observation. In practice, a strong observation signal is still reinforced through the loops, while a weak observation blurs out. We were able to improve our results on both synthetic and real data.

4. Temporal Multiple-View Reconstruction using Enhanced Belief Propagation

In this section we describe the application of all enhancements we made to belief propagation applied to the problem of temporal reconstruction from multiple video streams. The observation is constructed using the locally adaptive support-weight method of Yoon and Kweon [21]. We generate a set of depth hypotheses for each ray of the reference camera and each hypothesis is projected on all other cameras. The pairwise similarity between the reference and target camera is computed in a window of radius 11 according to [21], taking into account the color similarity of pixels as well as the distance to the central pixel. The aggregated similarities are averaged over all target images, to account for potentially different numbers of cameras in which the hypothesis is visible. Due to memory limitations, we only use 40 hypotheses per pixel in 1024×768 images, which explains the blockiness in some of our results.

BP is performed on a sliding window containing three frames of the reference view, each with an observation computed as above. All pixels become nodes in the graph and are connected with their 4-neighbors and with their temporal neighbors in the previous and next frames. When these correspondences are correctly detected, reconstruction results improve both in dynamic parts of the scene due to the combination of spatial and temporal smoothness and in static parts where temporal consistency reduces the errors and improves the visual quality of the reconstruction by reducing jittering. When optical flow fails, messages from unrelated nodes corrupt the beliefs and degrade the quality or the reconstruction. Robust BP is used to remove this effect. We estimate optical flow using [13].

Robust BP (Section 3.3) is applied to remove edges connecting nodes with inconsistent *PDF*s from the graph. This applies to both temporal and spatial neighbors. Figure 3 shows the edges that are disconnected from each node for a frame of the breakdancing sequence. Temporal neighbors are typically disconnected in areas with large motion, while spatial neighbors are disconnected at discontinuities. In Fig. 5(b) we show that in the case of static binocular stereo, edges across depth discontinuities are removed using this approach. This step achieves one of the goals of our research, which is to ensure that quality does not degrade when the establishment of temporal correspondences fails.



(a) An input frame (b) Disconnected edges

Figure 3. Edges disconnected per pixel for a frame of the breakdancing sequence. Green: disconnected temporal neighbors, red: disconnected vertical neighbors and blue: disconnected horizontal neighbors. Observe that this errors on the conservative side: discarding more than necessary.

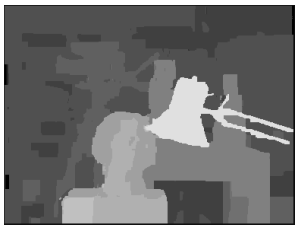
Occlusion is a major source of errors for stereo. Since the observation for occluded pixels is corrupted, we resort to a higher-level mechanism to correct these errors. It is based on the observation that occluded pixels typically belong to the background. If we could detect the occlusion, then we could assign the depth of the background (static part of the scene) to occluded pixels if their color is consistent with the background. Biased BP accomplishes that by using a prior for the background depth whose influence is stronger when the color of the pixel under consideration is similar to that of the background model and the confidence of the depth estimate for the pixel is low. See Section 3.4 for details. Here we briefly describe two methods for constructing the bias, but any suitable prior can be used instead. The first is similar to the work of [16], but applicable to more general types of scenes. We segment the image in regions of uniform color and robustly fit a plane to each segment using the current depth estimates. This method can also be used for single time-frame reconstructions. The second method uses the observations for all frames of the input video sequences to construct a colored depth map as the background model. For each pixel we select the most likely depth estimate in each frame and cluster all these estimates in a 4D HSV plus depth space. We select the median color and depth of the furthest significant cluster as the background model of each pixel. Figure 4 shows that the breakdancer has been successfully removed from the background. The parameter γ_c in Eq. (5) is estimated from the color distribution in either the segment or the background cluster depending on the method used.

Since the graph we operate on is 3D and each node is 6-connected, the number of cycles is very high. As a result noisy observations severely degrade the performance of BP in uniform ambiguous regions. Quiet BP addresses this problem by inputting the observation into the messages

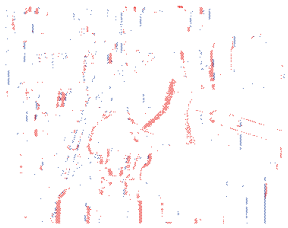
(a) Background color

(b) Background depth

Figure 4. The background model for the first 100 frames of the breakdancing sequence. Since the spectators move very little, their median positions are part of the background model.



(a) Depthmap for Tsukuba



(b) Disconnected edges

Figure 5. The depthmap we obtained for the Tsukuba dataset (using two images) and the edges disconnected by our robust scheme. Blue and red correspond to horizontal and vertical disconnected edges respectively.

only once, thus allowing smoothness in both space and time to play a larger role.

5. Experimental Results

In this section we present experimental results on a variety of datasets combining all the modifications we introduced in Section 3.

5.1. Binocular Stereo

We first tested our algorithm on the Middlebury Stereo Evaluation webpage datasets [3]. Since the data consists of image pairs, there are no temporal neighbors and we construct the bias by fitting planes to image segments. Robust BP successfully disconnects nodes across depth discontinuities (Fig. 5). Furthermore, QuietBP suppresses noise from uniform areas. Our results rank sixth overall in the evaluation as of April 2007. These results very promising since our method is not tailored for the binocular case.

5.2. 3D TV data

We also reconstructed data from the 3DTV data repository [1]. The "Janine" dataset consists of eight video streams and presents several challenges, including complex occluders (the plants), over-exposed uniform walls and a relatively wide baseline. Due to large inconsistencies in color calibration we were forced to drop one of the video streams. The plane-fitting method was used to construct the bias for this example. The depth map after initialization and the final depthmap we obtain can be seen in Fig. 6.

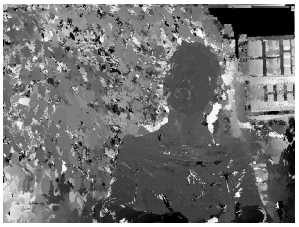
Table 1. Quantitative evaluation for the new Middlebury image pairs (acceptable error at 1.0 disparity level). Error rates and ranks as superscripts.

Image pair	nonoccluded	all	discontinuities
Tsukuba	0.94^{2}	1.74^{5}	5.05^{2}
Venus	0.35^{6}	0.86^{6}	4.34^{7}
Teddy	8.11^{11}	13.3^{9}	18.5^{10}
Cones	5.09^{13}	11.1^{9}	11.8^{10}

5.3. Comparison with Video View Interpolation

We performed extensive qualitative and quantitative comparisons with the video view interpolation approach of Zitnick et al. [23]. The videos were collected using eight synchronized cameras forming a 30^{o} arc and are available at [2]. We emphasize at this point, that their approach does not aim at geometrically accurate reconstruction, but rather at the synthesis of realistic views using a small number of the nearest viewpoints. The comparisons are not meant to sug-

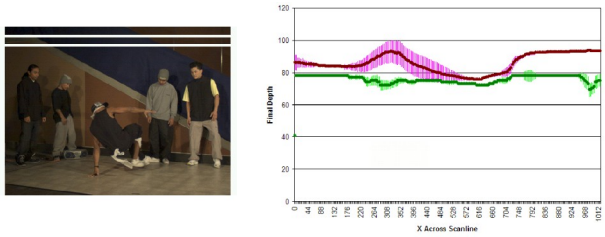
(a) Input frames from the two extreme and a central camera



(b) Observation

(c) Final depthmap

Figure 6. 3DTV data. Top row: three frames from the seven cameras. Notice the occlusion by the plants, the wide field of view and the over-exposed walls. Bottom row: initial depthmap and depthmap obtained by our method. Noise in the uniform areas is significantly reduced while details such as the pen and the folds on the clothes are preserved.



(a) A row on the wall

(b) Median depths and standard deviations

Figure 7. A frame of the breakdancing sequence with row 40 highlighted and a plot of the median depth and variance for all pixels of row 40 under the approach of [23] in purple (top curve) and our approach labeled in green (bottom curve). Note that our result is more planar and its standard deviation is smaller.

Figure 8. Zoomed in portions of the reprojections of the colored depthmap for camera 3 to camera 7. Left: using the depthmaps of [23]. Right: our results. Ghosting artifacts are reduced using our approach.

gest that our approach is better than theirs, especially since many of the aspects we focus on are out of the scope of video view interpolation. Their publicly available datasets and depthmaps have been invaluable to our research as inputs and baseline result respectively.

We first examine the time consistency of the reconstructions from frame to frame. A video in the supplemental material shows the depthmap for camera 3 of the breakdancing sequence computed by [23] who process each frame independently (on the left) and by our approach (on the right). Notice the "motion" in the walls and the floor which is reduced with our approach. We also performed a quantitative analysis on row 40 of the first 100 frames of camera 3. Since this row is entirely background the depth of all pixels should remain constant throughout the sequence. We computed the median depth for each pixel as well as the standard deviation of the estimates around the median value. Results can be seen in Fig. 7 for both methods. Note that the median values of the depth are not in agreement. This is because we use all cameras for the reconstruction thus increasing the accuracy due to the larger baseline. The fact that our depth estimates are more accurate is verified in the next experiment. What concerns us in this experiment is consistency across time which improves the quality of the visualization (see also supplemental video). The average standard deviation over all pixels was 13.94 for the method of [23] and 3.04 for our method. The minimum and maximum values of the standard deviation were 0.12 and 55.08 for [23] and 0 and 21.46 for our method.

We also computed the reprojection error of the reconstructions for camera 3 projected into camera 7, which is on one end of the arc. It should be noted here that the method of [23] does not use far away cameras to produce the interpolated views. Their reconstructions are used as baseline for the comparison with no claims that we outperform them in something that was out of their scope. Figure 8 shows zoomed in portions of the reprojection of the colored depthmap of camera 3 to camera 7. Ghosting artifacts, which are more obvious for the reconstructions of [23], are signs of inaccuracy in depth estimation. The RMS of the reprojection error in RGB, without compensating for occlusions, for the first 10 frames of the sequence was 29.36 for [23] and 20.84 for our method.

A depthmap and renderings of the reconstructed model from a novel viewpoint can be seen in Fig. 9. The boundaries of the breakdancer are sharp in our depthmap (Fig. 9(b)) even though we do not explicitly address boundary localization. This is an effect of Biased BP that correctly forces partially occluded background pixels towards the background depth. Notice that there is also no "webbing" between the arm and the body as in Fig. 9(a) or when using standard BP (Fig. 1(a)). The static, uniform parts of the scene are smooth in both cases, but they are also temporally consistent using our method as shown in the supplemental videos. The background model of Fig. 4 is used to fill in the occluded parts in the renderings. See the supplemental videos for a temporal reconstruction of the sequence.

Our method is computationally expensive and we have not explored even obvious ways of optimizing the implementation. The processing time for creating the observation for one frame of the breakdancing sequence is almost 4 minutes and processing three frames (at 1024×768) as described above takes 15 minutes for 150 iterations of BP.

6. Conclusions

We have presented an approach for temporal reconstruction of multiple video-streams based on an enhanced belief propagation framework. Our approach improves the temporal consistency of results whenever possible, but at the

(a) Depthmap by [23]

(b) Depthmap using our method

(c) Extrapolated rendering without background

(d) Extrapolated rendering using the background

Figure 9. A depthmap obtained by [23] and one by our method. The edges of the breakdancer are sharp in our depthmap in spite of not explicitly addressing boundary localization. (c) and (d) show renderings from viewpoints outside the camera arc. The background model is used in (d) to fill in occluded parts. Note that the mesh renderings do not connect points across depth differences above a threshold, so pixels on object boundaries are not rendered.

same time performs no worse than static multiple-view reconstruction if temporal correspondences are wrong. To achieve this, we construct graphs that include both spatial and temporal neighbors. We have augmented BP with the capability to disconnect neighbors whose beliefs are incompatible. We also made enhancements that address errors caused by occlusion. We believe that these enhancements are general enough to be applicable to other problems besides temporal multiple-view reconstruction. Our future work will focus on improving the efficiency of the implementation. We also intend to compute more than one depthmap and incorporate cross-validation in a similar way as [22, 7].

References

[1] 3DTV network of excellence, https://www.3dtv-research.org/publicSwLibrary.php. 2, 6

[2] Microsoft research: Video view interpolation, http://research.microsoft.com/~larryz/videoviewinterpolation.htm. 2, 6

[3] Middlebury stereo evaluation webpage, http://vision.middlebury.edu/stereo. 2, 6

[4] Mutli-view stereo evaluation webpage, http://vision.middlebury.edu/mview. 1, 2

[5] R. Carceroni and K. Kutulakos. Multi-view scene capture by surfel sampling: From video streams to non-rigid 3d motion, shape and reflectance. *IJCV*, 49(2-3):175–214, 2002. 1, 2, 3

[6] B. Goldluecke and M. Magnor. Space-time isosurface evolution for temporally coherent 3d reconstruction. In *CVPR*, pages 350–355, 2004. 1, 2, 3

[7] M. Gong. Enforcing temporal consistency in real-time stereo estimation. In *ECCV*, pages 564–577, 2006. 1, 2, 3, 8

[8] K. Kutulakos and S. Seitz. A theory of shape by space carving. *IJCV*, 38(3):199–218, 2000. 2

[9] J. Neumann and Y. Aloimonos. Spatio-temporal stereo using multi-resolution subdivision surfaces. *IJCV*, 47(1-3):181–193, 2002. 1, 2

[10] J. Pearl. *Probabilistic Reasoning in Intelligent Systems: Networks of Plausible Inference*. Kaufmann, San Fransisco, CA, 2nd edition, 1998. 3

[11] M. P. Perrone. *Improving regression estimation: Averaging methods for variance reduction with extensions to general convex measure optimization*. PhD thesis, Brown University, 1993. 3

[12] J. Pons, R. Keriven, and O. Faugeras. Modelling dynamic scenes by registering multi-view image sequences. In *CVPR*, pages 822–827, 2005. 2

[13] M. Proesmans, L. Van Gool, E. Pauwels, and A. Oosterlinck. Determination of optical flow and its discontinuities using non-linear diffusion. In *ECCV*, pages B:295–304, 1994. 5

[14] D. Scharstein and R. Szeliski. A taxonomy and evaluation of dense two-frame stereo correspondence algorithms. *IJCV*, 47(1-3):7–42, 2002. 2

[15] J. Sun, Y. Li, S. Kang, and H. Shum. Symmetric stereo matching for occlusion handling. In *CVPR*, pages 399–406, 2005. 3, 4

[16] H. Tao, H. Sawhney, and R. Kumar. Dynamic depth recovery from multiple synchronized video streams. In *CVPR*, pages 118–124, 2001. 2, 5

[17] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. In *ICCV*, pages 722–729, 1999. 1, 2

[18] S. Vedula, S. Baker, P. Rander, R. Collins, and T. Kanade. Three-dimensional scene flow. *PAMI*, 27(3):475–480, 2005. 1, 2, 3

[19] S. Vedula, S. Baker, S. Seitz, and T. Kanade. Shape and motion carving in 6d. In *CVPR*, pages 592–598, 2000. 2, 3

[20] J. S. Yedidia, W. T. Freeman, and Y. Weiss. Understanding belief propagation and its generalizations. In *Int. Joint Conf. on Artificial Intelligence*, Distinguished Papers Track, 2001. 3

[21] K. Yoon and I. Kweon. Locally adaptive support-weight approach for visual correspondence search. In *CVPR*, pages II: 924–931, 2005. 2, 5

[22] Y. Zhang and C. Kambhamettu. On 3-d scene flow and structure recovery from multiview image sequences. *IEEE Trans. on Systems, Man and Cyberentics*, 33(4):592–606, 2003. 1, 2, 3, 8

[23] C. Zitnick, S. Kang, M. Uyttendaele, S. Winder, and R. Szeliski. High-quality video view interpolation using a layered representation. *ACM Trans. Graph.*, 23(3):600–608, 2004. 1, 2, 6, 7, 8