Overcoming Parallax and Sampling Density Issues in Image Mosaicing of Non-Planar Scenes

Zhi Qi
Centre for Intelligent Machines
McGill University
qizhi@cim.mcgill.ca

Jeremy R. Cooperstock
Centre for Intelligent Machines
McGill University
jer@cim.mcgill.ca

Abstract

Image mosaicing constructs a wide field-of-view result from multiple source frames. In order to ensure a perceptually correct result, mosaicing typically requires either a planar or near-planar scene, parallax-free camera motion between source frames, or a dense sampling of the scene. When these conditions are not satisfied, various artifacts may result. A novel mosaicing approach is introduced that overcomes these limitations, building on the techniques of image-based rendering and manifold mosaicing, while permitting the synthesis of an effective mosaic of a non-planar scene from a sparse set of translated cameras. Our method first generates a series of intermediate virtual frames to reduce the disparities between neighboring images. Next, a series of vertical slices are chosen from the array of both real and virtual frames and connected according to a cost function that maximizes the similarity between adjacent slices. Experimental results indicate significant improvements over competing methods.

1 Introduction

In traditional imaging, with a fixed camera sensor, one must choose between capturing a wide field of view or high resolution details. *Image mosaicing* combines multiple input images, typically with some overlap, to produce an output with both wide field of view and high resolution. In general, this assumes that the camera motion is strictly rotational, as in QuicktimeVR [6], or that the scene is limited in depth variance, i.e. is planar or nearly planar. Otherwise, parallax effects result in the same point appearing at different relative positions on the multiple camera images. Such artifacts of duplication and missing objects can be avoided either by ensuring a parallax-free camera configuration or by increasing the sampling rate [13][15][16] of the scene.

Our motivation for investigating this problem stems from our three-screen videoconferencing configuration, illustrated in Figure 1a, in which a camera above each screen supplies video to a corresponding display at the remote location. At present, we have to restrict users to sit in the non-overlapping regions between cameras and require a uniform background to prevent the same objects appearing duplicated across multiple screens. However, it would be highly desirable to relax these constraints through the use of image mosaicing, which tolerates parallax effects thus is capable to generate reasonable

panorama view of the scene contents visible to the ensemble of cameras, without concern for the depth at which their respective fields of view begin to overlap.

We considered the simpler possibility of placing all three cameras in close proximity over the center screen,[1] rotated with respect to one another, as in Figure 1*b*. In this case, no restriction is necessary in the maximal depth at which users may sit from the cameras because the rotational camera configuration is insensitive to depth variance at all. However, unfortunately, such a configuration biases the correct view perspective to those users sitting near the center. Observers who are sitting at the sides would receive incorrect directional cues thus, precluding effective face-to-face interaction between these individuals sitting "across the table" from each other. An additional reason for preferring a translated camera configuration is that this provides depth cues that can be used later to generate a stereoscopic rendering for immersive applications.

Hence, the challenge we face is to generate a reasonable mosaic from a sparse sampling of the environment, in which depth variance of objects of interest may be significant, and where the camera configuration does not avoid parallax. To reduce the effect of parallax, our solution first synthesizes a number of virtual images along the path between input cameras and then combines these, along with the actual input images, using *manifold mosaicing*[9][13], which is robust to non-parallax-free camera configurations when provided a sufficiently dense sampling of the scene.

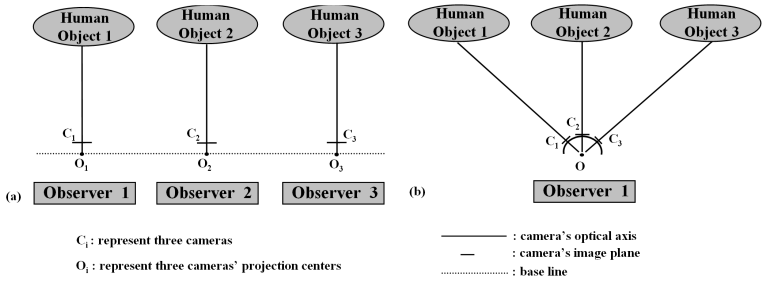

Figure 1: Two possible multi-screen videoconferencing camera configurations: (a) translational, which supports face-to-face communication across the table but suffers from overlapping fields of view; (b) rotational, which avoids problems of overlap but biases the perspective toward the center position.

Our work overcomes the conventional constraints of parallax-free camera motion or limited depth variance. Building on the techniques of image-based rendering and manifold image mosaicing, we further dispense with the need for dense sampling in the case of non-parallax-free camera motion. The result is an effective approach to image mosaicing where traditional mosaicing algorithms fail. The computational requirements of the algorithm we present here preclude real-time operation at video frame rates, but this should be overcome by taking advantage of the power of the programmable GPU. Even in its current implementation, we believe this represents an important step in the right direction.

The remainder of this paper is organized as follows. Previous work in image mosaicing is summarized in Section 2, followed by a detailed description of our approach in Section 3. The results of our algorithm, applied to various image databases, are presented

[1]This is the configuration favored by Cisco in their high-definition *Telepresenc their fiele* system.

in Section 4. Experimental results demonstrate that our system makes significant progress in mosaicing quality over *Autostitch* [4], in particular with respect to a reduction of ghost errors caused by parallax. Finally, possible improvements to our system are discussed in Section 5.

2 Previous Work

Traditional image mosaicing techniques include three basic steps. First, image registration is applied to find the geometric relationship between input images. Second, images are warped so that their regions of overlap match each other. Third, these warped images are stitched together into a common mosaicing plane.

The Concentric Mosaic [18] is an image-based rendering technique that constrains camera movement to planar concentric circles and creates panoramas by composting input images taken at different positions onto these circles. Shum and Szeliski's global and local alignment [19] greatly improved the accuracy of image registration without the prior known camera motion models so that tremendously enhanced the quality of mosaicing result from images taken by hand held cameras. Brown and Lowe [4] presented a fast image mosaicing system using robust feature-based image alignment, considered by many to be among the best algorithms currently available.[2] Levin *et al.*[11] investigated seamless image stitching in the gradient domain, which overcomes both the problems of photometric inconsistency and geometric misalignment between input images. In order to obtain reasonable mosaicing results, all of these systems require either the camera motion model to be parallax-free or the scene to exhibit little depth variance.

Another approach known as manifold or strip mosaicing [9][13][14], generates a multiperspective panorama by projecting thin strips from input images onto the mosaicing plane. The shape of the strips depends on the motion of input cameras and the width is proportional to the amount of camera motion.

Rav-Acha [15] addressed the problem of more complex camera motion in $3D$ space by time warping, which resamples the time axis to generate straight feature line in EPI space and further provides the panorama mosaic. Zomet *et al.* [23] introduced a different way of producing mosaics called *cross-slit projection*, which offers the benefit that the generated mosaics are closer to perspective images than traditional pushbroom mosaics. Recent efforts led to additional improvements in the smoothness of strip connections. Wexler and Simakov [20] minimized appearance disagreement error between slices by searching for the best path in the space-time domain. Agarwala *et al.* [1] computed a panorama for a long street scene using Markov Random Field Optimization.

Although the various manifold mosaicing techniques are far less constrained by camera motion model and depth variance than the more traditional image mosaicing algorithms, they do require a dense sampling, such as a video sequence of a static scene taken by one moving camera. In practice, however, obtaining this sampling density is often impractical or prohibitively expensive for typical applications, for example the videoconferencing system introduced in Section1.

[2] As the reference implementation is freely available, Autostitch also serves as a useful basis for comparison with other algorithms.

3 System Details

Our approach is intended to produce smooth mosaicing results even in the case of a highly limited number of input sources and a scene with large depth variance. In summary, starting from calibrated camera inputs, we synthesize a set of virtual images taken from intermediate positions to compensate for the limited sampling rate. It is not necessary to provide rectified inputs to the system, because both of the view synthesis and manifold mosaicing techniques that we employ are robust to the motion model of cameras. We then search in the space-time domain for the best strip stitching path, starting from the first column on the leftmost image to the last column on the rightmost image. The remainder of this section describes each of these steps in further detail.

3.1 Preprocessing

An initial calibration step is required to estimate the intrinsic and extrinsic parameters of the cameras. For this purpose, we use Zhang's [22] method as implemented by Bouguet's calibration toolbox [2]. When applying our algorithm to external data sources, we recover the camera calibration parameters though the structure-from-motion algorithm [10]. Since our system is intended to operate in conjunction with identical high-definition cameras in an environment with controlled illumination, we ignore at present the correction of lens distortion, removal of vignetting, and compensation for exposure or color differences between input images. For lower quality imaging hardware, these factors could be addressed by a number of methods known in the literature.

3.2 Synthesis of Virtual Images

We compensate for a low sampling rate by synthesizing frames that would be generated by a number of virtual cameras positioned between the physical cameras. This decreases the effective baseline distance between neighboring cameras, whether real or virtual, and in turn, minimizes disparity effects between adjacent images, resulting in substantially improved mosaicing results.

The synthesis of virtual frames is performed using the plane sweep algorithm [8], which combines depth from stereo and 3D warping. Plane sweep projects input images onto the image plane of a virtual camera and then sweeps the image plane along the depth axis, examining the color consistency of projected input pixels. Moreover, the algorithm can benefit from implementation on a programmable GPU to provide very fast processing [21] thus likely permitting real-time operation of our image mosaicing system on high-definition video.

The number of virtual camera images is determined as a function of the difference between minimum and maximum disparities observed in the input images and the size of the search window used in the space-time domain. As the difference in disparities increases or the search window shrinks, a higher sampling rate, and thus, a greater number of virtual images, is required to avoid aliasing in the mosaic result. Although beyond the scope of this paper, a more detailed explanation of this topic is provided elsewhere [5][12].

Critical to the stitching operation, the movement (both translation and rotation) between neighboring cameras must be smooth and continuous. Therefore, the virtual sources

are positioned at equal intervals between the physical cameras and their orientations are interpolated smoothly between the rotation matrices of the nearest cameras.

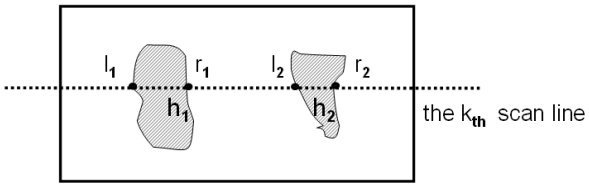

Figure 2: There are two shaded hole regions in this figure. On the k_{th} scan line, points l_1, r_1 and l_2, r_2 are the pairs of non hole region depth value on both sides of the segments across hole region h_1 and h_2 respectively. We fill segment on kth scan line in the hole region h_1 with the bigger depth value of l_1 and r_1 and apply the same way to fill the segment in hole h_2.

When two input cameras are located along a wide baseline, as is the case in our test data, there are likely to be some portions of the 3D scene that are visible by one camera but occluded in the other. This results in holes appearing in the virtual images, which we fill using depth information from neighboring pixels. We scan the hole region row by row. As illustrated in Figure 2, for any row, each hole h_i can be delineated on both sides by the nearest non-hole points, l_i and r_i, with corresponding depths determined by the plane sweep operation. Assuming that the hole exists because of an occluder in one camera view, we estimate its depth by choosing between the two points l_i and r_i the one furthest from the camera. Once all the depth values for the hole regions are estimated in this manner, we can use a foreword forward mapping to project input images onto the virtual image plane at the specified depth, thereby filling in the missing texture of the virtual images. Although more advanced hole-filling algorithms, such as inpainting, are well known, our experiments suggest that the simple strategy described above works sufficiently well for the small holes typically observed in the virtual images.

3.3 Optimal Slice Stitching

Given that no ground truth is available for the image mosaicing result, determining the quality of a particular image mosaic is largely a subjective matter, based on criteria of appearance smoothness and continuity. Wexler *et al.* [20] proposed a metric for these characteristics, and suggested an approach to optimal stitching together of slices to generate a perceptually satisfying mosaic, in which every pair of local neighboring slices should resemble some regions in input frames.

The set of real and virtual camera images are stacked together into a cube, as shown in Figure 3*a*. Each image occupies a slice in $x-y$ space, and these are arranged along the third axis, t.[3] According to the work of adaptive manifold by Peleg *et al.* [14], in the case of translational dominant camera motion model, it is reasonable to choose vertical strips

[3] In video sequences, the meaning of the temporal dimension is obvious; in our case where the images are captured (approximately) simultaneously from a number of cameras, this can be thought of as equivalent to capturing a video sequence of a scene from a single camera moving over time across different input camera positions. Thus, for consistency of terminology, we retain t as a virtual time axis.

from input images to build a mosaicing result. The mosaic is then generated by combining vertical strips from these images along a minimum-cost path in $x-y-t$ space, starting from the first column strip of the first image (at t_0) and ending with the last column strip of the final image (at t_{end}). The cost is determined simply as the sum of squared differences between every pair of adjacent strips being combined. Starting from the initial node in the path, a dynamic programming search is conducted to find the best choice of successive node among all the possible choices within the three-dimensional search volume. This process continues until the final node is reached. An example is illustrated in Figure 3*b*.

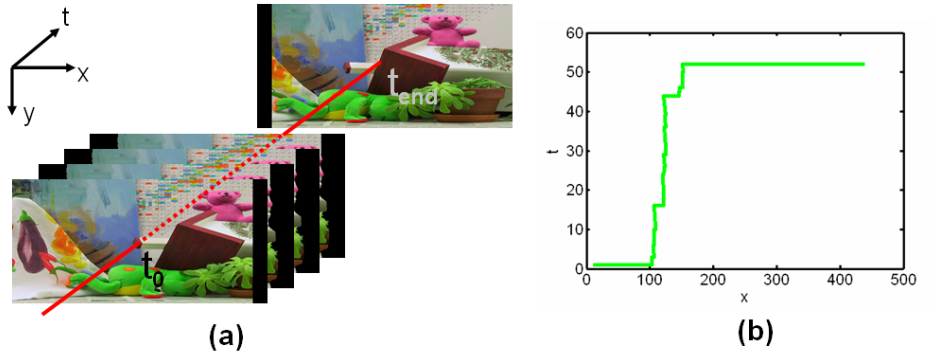

Figure 3: (a)the image cube in the space-time domain and (b)one example of the minimum-cost path found in this domain to stitch the strips from multiple images.

4 Experiments

To evaluate the quality of our algorithm, we compare its results to those of Autostitch [3], which generates excellent mosaics from source images constrained by the conditions outlined in Section 2. This comparison highlights the capacity of our approach to generate significantly improved mosaics from source images exhibiting non-trivial parallax, even with a very low sampling rate.

The first two data sets used for our comparisons are of indoor scenes. The Middlebury teddy data set [7] contains nine color images taken from distinct positions on the same baseline, from which we choose the leftmost and rightmost images as the inputs to our system. Given the range of object depths, the maximum difference in x directional disparities between foreground and background objects is 88 pixels. The second example, collected by Seitz [17] exhibits a maximal disparity difference of 126 pixels while the outdoor (house) scene exhibits a lower disparity difference of only 46 pixels.

Autostitch first deforms the two input images, compressing objects closer to the cameras and expanding distant objects to equalize their respective disparities. This normalizes the amount of overlap between the images, allowing for a smooth combination of the two deformed inputs.

The results of Autostitch, seen in Figures 4 and 5, contain a foreground region at the bottom of the mosaic, which has shrunk relative to the background at the top. We verified that this differential stretching effect is due to image content by flipping the source images

upside down. As expected, this produces a mosaic in which the foreground, now at the top of the image, shrinks relative to the background at the bottom. In contrast, the results of our algorithm are free of such deformation, as the parallax effects have been greatly reduced by the incorporation of synthesized virtual images. Furthermore, our mosaics do not exhibit the unpleasant ghosting effects seen in the Autostitch results, stemming from the difficulty of accurate alignment given the huge depth variances in the scene.

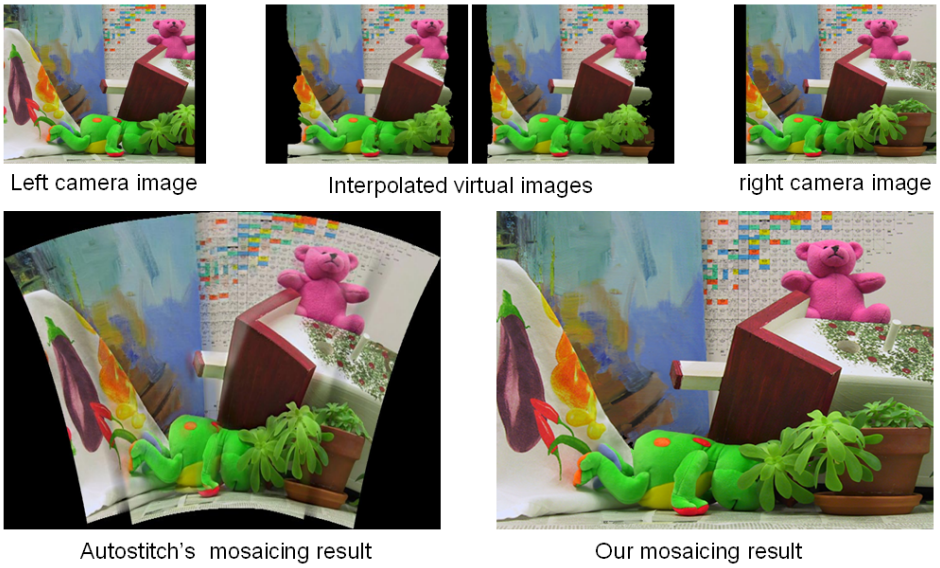

Figure 4: Experimental results from the Middlebury teddy data set.

The third data set used for our experiments is taken from an outdoor scene of a house with trees in the background and right foreground. This example has disparity difference within the range that Autostitch can handle. The resulting mosaic, pictured in Figure 6, still exhibits some radial distortion as well as ghost errors around the window, while our result, although free of radial distortion, contains ghost errors on the garage door. Potentially contributing factors to this error include the accuracy of camera calibration, the large textureless regions in the input images, and robustness of the plane sweep algorithm to color inconsistency.

In the experimental results, above, our algorithm used only the color channel of the virtual images, searching for the best stitching path in the spatiotemporal volume according to the color consistency principle. We also tested the algorithm using the depth channel of virtual images, connecting stripes only when their contents are similar in depth.

As shown in Figure 7, searching by depth consistency finds a similar path as with the color consistency rule. The results are slightly noisier than those of the earlier mosaicing example, based on color consistency, seen in Figure 4. This is likely because the estimation of depth by the plane sweep operator is less robust than that of color. However, the test confirms that depth cues are also important to determining alignments between input images. This may prove significant when generalizing our algorithm to dynamic scenes, with objects moving about the environment. Otherwise, we are required to recompute the entire stitching path for each frame, which generally results in observable jitter.

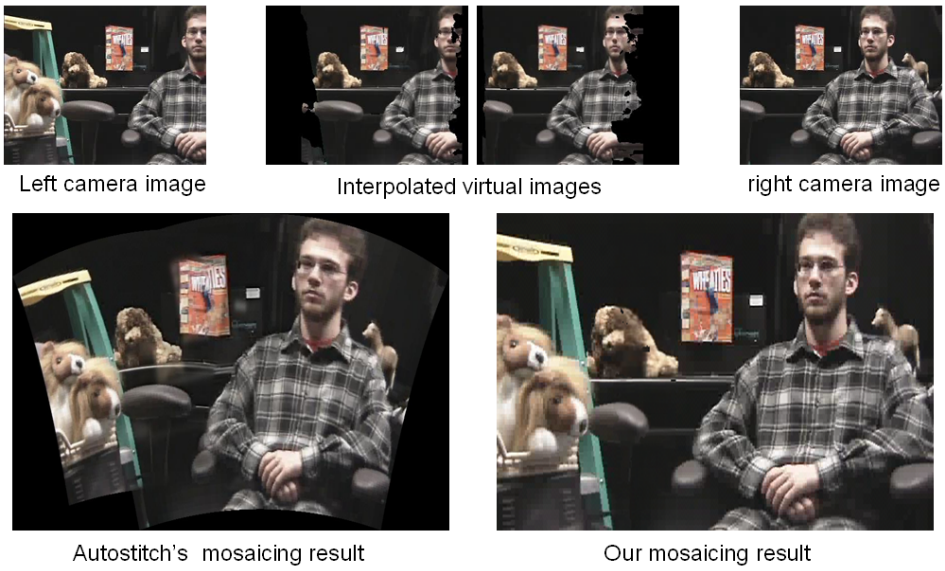

Figure 5: Experimental results from an indoor scene with greater depth variance.

5 Conclusions and Future work

Issues of parallax and camera motion rate prevent traditional image mosaicing algorithms from generating perceptually acceptable panoramas in many situations. The approach we introduce combines previous knowledge of image-based rendering and manifold mosaicing to overcome some of these limitations and produce reasonable mosaics from sparse input sources, situated along a wide baseline.

Our system may be improved along a number of dimensions. Taking advantage of the power of the programmable GPU could greatly accelerate the processing speed and allow for real-time operation on live video. The quality of synthesized virtual images could be improved by a global optimization algorithm and by tracking the movements of people (or other dynamic objects) in the scene.

References

[1] A. Agarwala, M. Agrawala, M. Cohen, D. Salesin, and R. Szeliski. Photographing long scenes with multi-viewpoint panoramas. *SIGGRAPH '06: ACM SIGGRAPH 2006 Papers*, pages 853–861, 2006.

[2] J.V. Bouguet. Camera calibration toolbox for matlab. `http://www.vision.caltech.edu/bouguetj/calib_doc/`. 2003.

[3] M. Brown and D.G. Lowe. Autostitch :: a new dimension in automatic image stitching. `http://www.cs.ubc.ca/~mbrown/autostitch/autostitch.html`.

[4] M. Brown and D.G. Lowe. Recognising panoramas. *ICCV '03: Proceedings of the Ninth IEEE International Conference on Computer Vision*, page 1218, 2003.

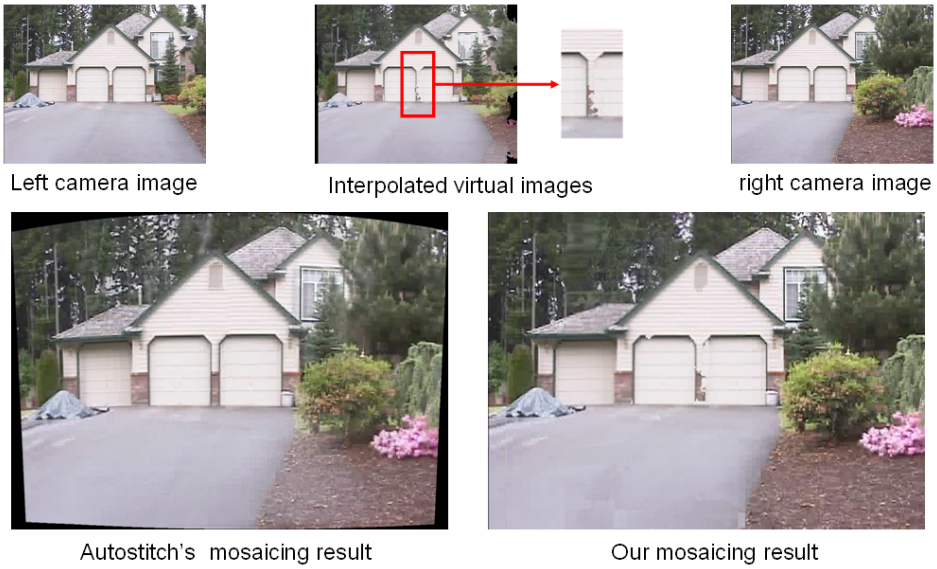

Figure 6: Experimental results from an outdoor scene with small depth variance. The boxed region of the synthesized virtual image is enlarged to the right in order to illustrate the source of the ghost error around the garage door.

[5] J.X. Chai, S.C. Chan, H.Y. Shum, and X. Tong. Plenoptic sampling. *SIGGRAPH '00: Proceedings of the 27th annual conference on Computer graphics and interactive techniques*, pages 307–318, 2000.

[6] S.C. Chen. Quicktime vr: an image-based approach to virtual environment navigation. *SIGGRAPH '95: Proceedings of the 22nd annual conference on Computer graphics and interactive techniques*, pages 29–38, 1995.

[7] Middlebury College. Stereo vision research page. `http://cat.middlebury.edu/stereo/newdata.html`.

[8] R. Collins. A space-sweep approach to true multi-image matching. *IEEE Computer Vision and Pattern Recognition*, pages 358–363, 1996.

[9] R. Gupta and R.I. Hartley. Linear pushbroom cameras. *IEEE Trans. Pattern Anal. Mach. Intell.*, 19(9):963–975, 1997.

[10] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, New York, NY, USA, 2000.

[11] A. Levin, A. Zomet, S. Peleg, and Y. Weiss. Seamless image stitching in the gradient domain. *Hebrew university technical report:2003-82*, 2003.

[12] Z.C. Lin and H.Y. Shum. A geometric analysis of light field rendering. *Int. J. Comput. Vision*, 58(2):121–138, 2004.

[13] S. Peleg and J. Herman. Panoramic mosaics by manifold projection. *CVPR '97: Proceedings of the 1997 Conference on Computer Vision and Pattern Recognition (CVPR '97)*, page 338, 1997.

[14] S. Peleg, B. Rousso, A. Rav-Acha, and A. Zomet. Mosaicing on adaptive manifolds. *IEEE Trans. Pattern Anal. Mach. Intell.*, 22(10):1144–1154, 2000.

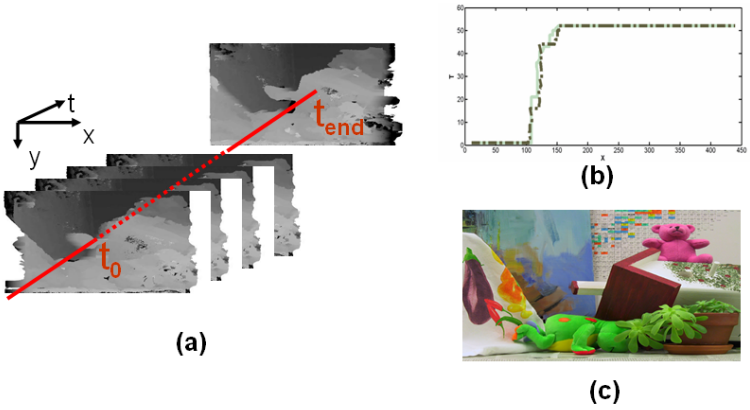

Figure 7: (a) The group of real and virtual images in the spatiotemporal domain, (b) the minimum-cost path found by searching for consistency in the depth channel (light solid line) and in the color channel (dark dashed line), (c) the corresponding color channel mosaicing result based on the path of depth channel (light solid line) in (b).

[15] A. Rav-Acha, Y. Shor, and S. Peleg. Mosaicing with parallax using time warping. *CVPRW '04: Proceedings of the 2004 Conference on Computer Vision and Pattern Recognition Workshop (CVPRW'04) Volume 11*, page 164, 2004.

[16] A. Román and H.P.A. Lensch. Automatic multiperspective images. *Rendering Techniques 2006: Eurographics Symposium on Rendering*, pages 161–171, 2006.

[17] S.M. Seitz and J. Kim. The space of all stereo images. *Int. J. Comput. Vision*, 48(1):21–38, 2002.

[18] H.Y. Shum and He L.W. Rendering with concentric mosaics. *Computer Graphics in SIGGRAPH '99*, pages 299–306, 1999.

[19] H.Y. Shum and R. Szeliski. Construction and refinement of panoramic mosaics with global and local alignment. *ICCV '98: Proceedings of the Sixth International Conference on Computer Vision*, page 953, 1998.

[20] Y. Wexler and D. Simakov. Space-time scene manifolds. *ICCV '05: Proceedings of the Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1*, pages 858–863, 2005.

[21] R. Yang, M. Pollefeys, H. Yang, and G. Welch. A unified approach to real-time multi-resolution. *International Journal of Image and Graphics, 2004*, 2004.

[22] Z.Y. Zhang. A flexible new technique for camera calibration. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 2000.

[23] A. Zomet, D. Feldman, S. Peleg, and D. Weinshall. Mosaicing new views: The crossed-slits projection. *IEEE PAMI*, pages 741–754, 2003.