Wide-Baseline Image Mosaicing for Indoor Environments

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

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

Abstract

We introduce an automated image mosaicing system using a limited number of cameras arranged along a wide baseline, which generates a perceptually acceptable panorama of an indoor environment, including human subjects at varying depths. The target application for this research is the production of a seamless, extra-wide display for videoconferencing. In such a case, we deliberately employ a configuration of translated cameras in order to preserve sensible semantics of eye contact for a viewer located toward the sides of the display. This is preferable to co-locating the cameras at a common center position *in an attempt to approximate the case of rotation-only camera movement. Our system applies feature-based and direct image alignment techniques to register the input images, and then creates the final image mosaic through multi-perspective projections. The competitive results of our work with current state-of-the-art techniques are discussed.*

1 Introduction

Image mosaicing is a generic term referring to the process of combining a group of images to generate a result that has a wider field-of-view (FOV) and at least an equivalent resolution to the individual inputs. Early applications stitched together a set of images taken by handheld cameras to construct wide-angle, seamless panoramas [5][8]. For input images taken by cameras rotating around a fixed projection center, Shum and Szeliski [12] applied both global bundle adjustment and local patch-based deghosting to improve the quality of the resulting mosaics. Brown and Lowe [4] introduced a feature-based approach to recognize and align images for the generation of a high-quality panorama of distant (i.e. essentially co-planar) scenes. Peleg *et al.* [10][11] and Zhu *et al.* [13] eliminated the restriction of a fixed projection center and generated image mosaics based on a translationally dominant camera motion, provided a relatively dense image sampling.

Although there are many successful applications of this technique, image mosaics suffer from ghosting errors due to parallax. Satisfactory results are obtained only when the following constraints hold:

- camera rotation with a fixed projective center [12],
- dense sampling using a large number of overlapping camera views, as is obtained in the video sequence from a camera moving around the environment [11][13],
- limited depth variance in the scene, *i.e.* approximate co-planarity along the direction normal to the camera's optical axis [4].

Indoor environments, which are typically characterized by large variances in depth, are thus problematic cases for mosaicing algorithms when the projection center of the camera is not fixed. Peleg *et al.* [10] addressed this problem by using a dense sampling approach of a static environment, based on a pushbroom camera model. Gorges *et al.* [6] recently introduced the idea of reducing parallax-induced mosaicing errors by decomposing the scene into planar sub-scenes and creating mosaics for each plane individually. However, the example provided by Gorges for an office scenario assumed that the primary objects could be represented as polyhedrons, which is unlikely to be the case for humans.

Our goal is to develop an affordable image mosaicing system that generates seamless high resolution panoramas including both human participants and their associated backgrounds, automatically, using only a limited number of cameras. The obvious challenge is how to do so despite the unavoidable problems of parallax.

2 Camera Configuration

We use a limited number of cameras to acquire image content simultaneously from a wide field of view. There are two possible arrangements to consider: a rotational configuration in which the cameras are closely located and rotated

around a common center of projection, as in Figure 1(a), or a translational configuration, with the cameras mounted along a wide baseline and facing the same direction, as in Figure 1(b). The resulting mosaic is rendered so that the dominant foreground objects (people) appear life-size over the display.

In order to generate a result in which people (or objects) near the sides of the display appear in a perceptually consistent manner to the remote viewers directly facing them, it is important to keep the social cues such as eye contact and sense of pointing direction, as would be the case in the translational configuration. In the example of Figure 1(a), with all the subjects looking at the same target (lower right), the co-located cameras capture gaze direction of all the subjects as being at the same angle, θ to the optical axis. This results in a mosaic that is rendered as if all the human subjects are viewing parallelly towards certain targets, which is conflict to the true situation as shown in Figure 1(b). In contrast, the translational configuration captures different gaze directions of the subjects in front of each camera, and thus, the mosaic appears perceptually correct in the area in front of viewers, regardless of their positions.

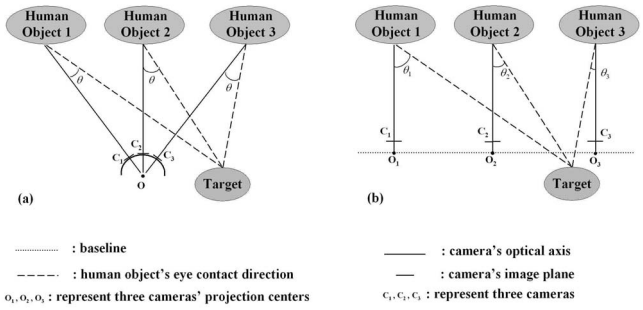

Figure 1. (a) rotational camera configuration: each camera captures gaze direction at the same angle θ to the optical axis (b) translational camera configuration: gaze direction with respect to the optical axis is preserved more closely for each camera

3 Background Theory

Image mosaic construction requires a step of image alignment, or registration, which finds the transformation between pixel coordinates of neighboring input images, followed by image stitching, which warps and pastes the relevant content from the input images to the reference mosaic plane. The remainder of this section describes these techniques in further detail.

3.1 Image Alignment

Image alignment algorithms estimate the parameters of camera motion models to determine the geometric relationship between images. These may be classified as either direct or feature-based. The former find the pixel-to-pixel matching by comparing cross-correlation values through a window search. These utilize all the pixel information in the overlapping regions between images and generates a more accurate estimation of the camera motion model. However, if the initial estimate is far from the optimal values, the algorithm may take a long time to converge or may not converge at all. In contrast, feature-based approaches estimate camera motion models through the correspondence of a limited number of features over different images. This approach is faster than direct image alignment but it is not well suited for images with an insufficient number of features, such as the case with large textureless regions.

Figure 2. Two neighboring input images are shown, with crosses indicating SIFT feature positions. In the first row, these are restricted to the human subject who appears in the overlapping region, whereas, in the second row, only the crosses from a nearest neighbor match are retained. The circles in the third row represent the RANSAC inliers, which are used to estimate initial camera motion model parameters.

Our system takes advantage of both approaches. We use

the feature-based approach to find an initial estimate of the camera motion model for a successive direct approach. The initial estimation step reduces the time needed for the direct approach to converge. The direct approach then uses all the information in the overlapping regions to generate a final, precise estimate of the camera motion model.

The wide baseline translational camera configuration, as in Figure 1(b), increases the level of occlusion, which complicates the search for correspondences between images. Based on the conclusions of Mikolajczyk and Schmid [9], we use Lowe's SIFT features [7] as a good choice for its invariance to scale, rotation, color, intensity, and geometric distortion. For every two neighboring input images, we first calculate SIFT features, then use a nearest neighbor algorithm to find the candidate matches of features between them. The Hough transform is used to cluster these matches into different groups, each of which satisfies a respective camera motion model. We then apply RANSAC to the group of candidate matches with the highest probability and separate this group into a set of RANSAC inliers, geometrically consistent with the estimated camera motion model, and inconsistent outliers. Based on the inlier feature correspondences, RANSAC also generates an initial estimate of parameters for an affine camera motion model, which represents the transformation between two neighboring images. Figure 2 illustrates how the initial estimate of camera motion model is obtained from the two input images.

Finally we use the Lucas-Kanade algorithm [2][3] to arrive at the final estimate of the camera motion model, updating the initial estimate iteratively until the overlapping region between two neighboring images is correctly aligned.

3.2 Image Stitching

For a group of translated cameras located along a wide baseline and covering a wide field-of-view, *manifold projection* [10] provides an efficient way to combine input images based on the estimated camera motion models between them. With the estimated camera motion model, we first warp the second (right) image into the plane of the first (left) image and find where the two overlap. Next, we cut the first image along the overlapping boundary to obtain a slice seen only by the first camera. We collect all the other slices by applying the same procedure to all remaining image pairs. Last, we warp and paste these slices onto the common plane to generate a final image mosaic. Since a single, translating (dollying) camera is used to collect all the input images, we may assume no color or intensity differences between these. The merging of slices is simply a weighted summation of pixel color values along the merge boundary.

One advantage of manifold projection is that it results in less distortion of objects on the borders of a wide field-of-view in the final image mosaic. If instead, we use one of the input images as a reference and project all the other images onto its coordinate system, pasted slices from those cameras furthest from the reference will exhibit greater distortion in the final mosaic, because perspective projection stretches pixels most as they are near the image boundary.

4 Experimental Results

An example result of our algorithm is shown in Figure 3 (Considering the limited space of this 4-page paper, we only list one of our experiment results). A Sony TRV-900 camera was dollyed across the room to acquire a video sequence of a semi-static environment.[1] Six frames, each of size 720×480 pixels, were selected from approximately equally separated positions across a wide baseline, with overlap between successive images less than 50%. The resulting mosaic has a resolution of 3344×525 before cropping.

We compare our result to that obtained from Brown and Lowe's AutoStitch [4]. AutoStitch is a powerful tool that automatically generates high quality panoramas when provided with images of a remote outdoor scene as input [1]. However, when applied to our data from an indoor environment, we find that both AutoStitch and our algorithm suffer from ghosting errors due to parallax, e.g. the duplicated outlet box, which appears behind the female subject in the leftmost two input images. Possibly due to the slight rotational motion between successive frames, the results of Autostitch contain arc-like connections between adjacent slices. Since our approach combines vertical shift and rectangular warping, the generated result exhibits significantly reduced vertical displacement, which is more acceptable for a videoconferencing application. The most glaring problem with our result is the stretched appearance of the individual slightly left of center, which was due to the significantly differing viewpoint of the camera between the second and third input images.

Our system also has difficulty when feature extraction fails to find correspondences in textureless regions, for example, when foreground subjects are dressed in a uniformly dark color. In our future work, we plan to exploit depth information to enhance the mosaic result regarding to these limitations.

5 Conclusions and Future Work

We introduced an automated image mosaicing system that generates mosaics of human objects in an indoor environment, suitable for wide field-of-view videoconferencing applications. Our approach uses both feature-based and direct approaches to improve the accuracy and efficiency of

[1] The human subjects were asked to remain still during the filming.

Figure 3. A sample mosaic result from six input images, compared with AutoStitch[4].

image alignment results, and the strategy of multiple perspective projection reduces object distortion near the borders. Together, these techniques improve robustness to the violation of parallax-free constraints, as is the case in most indoor environments.

In our future work, we hope to scale the system to real-time operation and address the problem of ghost errors due to 3D parallax. Ghost errors are of particular concern for environments containing many people, as they are perceptually much more distracting when they affect a human face. In this regard, manifold projection offers another important advantage, as it generates an image mosaic as being approximately taken by a *slit* or *pushbroom camera*. Provided that we can synthesize a sufficient number of (virtual) inter-camera frames along a wide baseline, these ghost errors caused by parallax can be avoided.

References

[1] *http://www.cs.ubc.ca/ mbrown/panogallery/panogallery.html*.

[2] S. Baker, R. Gross, I. Matthews, and T. Ishikawa. Lucas-kanade 20 years on: A unifying framework: Part 2. *Robotics Institute, Carnegie Mellon University*, (CMU-RI-TR-03-01), February 2003.

[3] S. Baker and I. Matthews. Lucas-kanade 20 years on: A unifying framework: Part 1. *Robotics Institute, Carnegie Mellon University*, (CMU-RI-TR-02-16), July 2002.

[4] M. Brown and D. Lowe. Recongnising panorama. *Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on*, 2:1218–1225, October 2003.

[5] S. E. Chen. Quicktime VR: An image-based approach to virtual environment navigation. *Computer Graphics*, 29:29–38, 1995.

[6] N. Gorges, M. Hanheide, W. J. Christmas, C. Bauckhage, G. Sagerer, and J. Kittler. Mosaics from arbitrary stereo video sequences. *DAGM-Symposium*, pages 342–349, 2004.

[7] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, January 2004.

[8] S. Mann and R. W. Picard. Virtual bellows: Constructing high quality stills from video. *ICIP (1)*, pages 363–367, 1994.

[9] K. Mikolajczyk and C. Schmid. A performance evaluation of local descriptors. *IEEE Trans. Pattern Anal. Mach. Intell.*, 27(10):1615–1630, 2005.

[10] 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)*, pages 338–343, 1997.

[11] 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.

[12] 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*, pages 953–958, 1998.

[13] Z. Zhu, A. R. Hanson, and E. M. Riseman. Generalized parallel-perspective stereo mosaics from airborne video. *IEEE Trans. Pattern Anal. Mach. Intell.*, 26(2):226–237, 2004.