GENERATING SEAMLESS STEREO MOSAICS FROM AERIAL VIDEO

Zhigang Zhu, Allen R. Hanson, Harpal S. Bassali
Howard J. Schultz, Edward M. Riseman
Computer Vision Laboratory
Department of Computer Science
University of Massachusetts at Amherst
Amherst, MA 10003
(zhu|hanson|harpal|hschultz|riseman)@cs.umass.edu

ABSTRACT

This paper is directed at the development of tools for environmental science and other aerial video applications requiring the mapping of video into absolute world coordinates. As part of this effort, we propose a new method that automatically and efficiently generates seamless stereo mosaics from an aerial video camera moving over a large distance. An important part of our approach is a parallel-perspective mosaic representation that can well support large seamless mosaics and also can capture the inherent 3D information. We will show that the condition under which such stereo mosaics can be created is a rather general motion (3D rotation and translation) with a dominant direction, and the parallel-perspective stereo mosaics can be converted into a digital elevation map (DEM) after 3D recovery. A novel method PRISM (parallel-perspective ray interpolation for stereo mosaicing) is proposed to generate dense and seamless parallel-perspective stereo mosaics from sparse image sequences with motion parallax. The advantages of stereo mosaics are shown in representation, computation, 3D recovery and 3D visualization. Issues on the implementations of geo-references mosaicing as well as free mosaicing are discussed. Experimental results on several long video sequences are given.

INTRODUCTION

A critical issue among nations in the coming decades will be how to manage the use of land and natural resources. Unfortunately, the use of satellite data has not enabled general and automatic ecosystem modeling because many of the dynamic changes of interest in ecosystems take place at a finer level of resolution than is available. Thus applying high resolution low-altitude video sequences is highly required for interpreting the lower resolution data. Our interdisciplinary NSF environmental monitoring project aims at developing a methodology for estimating the standing biomass of forests. The instrumentation package mounted on an airplane consists of two bore-sighted video cameras (telephoto and wide-angle), a Global Position System(GPS), an Inertial Navigation System (INS), and a profiling pulse laser. The previous manual approach used by our forestry experts (e.g. Slaymaker et al, 1996) utilized only a fraction of the available data due to the labor involved in hand interpreting the large amount of video data. A more compact representation and more flexible interactive 3D visualization interface are clearly necessary. Automatically generated mosaics of forests that are both geo-referenced and support stereo viewing are of crucial importance when huge amounts of video data of the forest must be processed.

We have already demonstrated the significant potential of the emerging digital video technology soft-copy photogrammetry (including geo-referenced stereo mosaics) through a set of initial collaborative projects with environmental science partners, in regions of New England, Bolivia, Brazil, and Madagascar. One noteworthy recent success was the collection of 130 hours of digital video consisting of 10 Terabytes of uncompressed imagery over the Amazon rain forest. This was carried out in collaboration with the Smithsonian Institute, the University of California at Santa Barbara (UCSB) and I.N.P.E. (an arm of the Brazilian Space Agency). It was the largest ever aerial video data collection over the Amazon Basin, and we produced the first geo-registered aerial maps of the Smithsonian site for study of the dynamics of forest fragmentation. Another example project in Bolivia involved The Nature Conservancy (TNC), Winrock International, and American Electric Power (AEP), to evaluate our technology for the determination of biomass via aerial video analysis.

In fact, for many applications dealing with large-scale natural or urban scenes, extending the field of view (FOV) of 2D images, and then introducing the third dimension of depth would be of great utility. Video surveillance (Kumar,

et al. 1998), environmental monitoring (Zhu et al, 1999a; 1999b), home scene archives and retrieval, robot navigation (Zheng & Tsuji 1992; Zhu et al, 1999c), and wearable camera systems (Mann 1998) are just a few examples of the applications that would benefit from an extended and enhanced image-based representation. Theoretically, reconstructing 3D information from pairs of images in a video sequence, and then integrating multiple 3D maps (as well as the texture maps) into a large texture-mapped digital elevation map (DEM) is a general way to fulfill this goal. However, there are many technical issues that need to be solved, such as camera calibration, dense image correspondence, prohibitive computation, and consistency of 3D recovery across multiple frames. Consequently, the visual modeling and representation of large-scale 3D scenes is still an open problem.

Mosaic representations are suitable for modeling large-scale 3D scenes. Many are familiar with two dimensional (2D) mosaics - composite images with extremely large field of view. *Two dimensional (2D) mosaicing* is the generation of a 2D composite image from an image sequence of (1) an arbitrary scene under (almost) pure camera rotation (e.g. panning); or (2) a planar scene under 6 degree of freedom (DOF) motion. In these two cases, a 2D projective transformation (homography) can be used to model the motion between two frames. Here, we will address the problem of creating seamless and geo-corrected 3D mosaics from a moving camera whose viewpoints change over a large scale. In general, *three dimensional (3D) mosaicing* is the generation of composite images that preserve 3D information, from an image sequence of an arbitrary scene under more general motion where motion parallax between frames is apparent. Obviously, using 2D mosaicing techniques based on 2D image transformation cannot generate seamless 3D mosaics in the presence of large motion parallax.

Existing Approaches

Recently, the construction of panoramic images and high quality mosaic images from video sequences has attracted significant attention. Many of the current successful image mosaic algorithms, however, only generate 2D mosaics (either a 360-degree panorama or a full sphere omni-directional image) from a camera rotating around its nodal point (Chen 1995; Sawhney, et al. 1998; Shum & Szeliski, 1997; Xiong & Turkowski 1998; Zhu et al., 1999d). Creating stereo panoramas from two rotating cameras was proposed by Huang & Hung (1998), and from one off-center rotating camera by Ishiguro, Yamamoto & Tsuji (1990), Peleg & Ben-Ezra (1999), and Shum & Szeliski (1999). In these kinds of stereo mosaics, the viewpoint is limited within a very small area, usually along a circle of a meter in diameter. However mosaicing from a translating camera is the typical situation prevalent in airplane motion during forest surveys and car motion during navigation on roads. Image mosaicing from a translating camera will result in a mosaic image with multiple viewpoints along the motion path. It raises a set of different problems from that of circular projections of a rotating camera, including motion modeling and motion estimation (e.g. from airplane turbulence), generation of seamless image mosaics under motion parallax, and epipolar geometry of multiperspective projection (Zhu et al 1999a, 1999b; 2000a, 2000b).

In applications such as environmental monitoring and aerial surveillance, geo-referenced mosaics tied to real world coordinates are required. For example, Kumar, et al (1998) presented a geo-registration method that can register video mosaics to a high-altitude reference image using additional engineering data of the camera's poses (geo-data). A fine geo-registration requires knowledge of a reference image (geo-referenced aerial image with broader coverage) as well as an accompanying co-registered DEM (digital elevation map). Our proposed stereo mosaics will be geo-referenced naturally if the motion of the camera at every frame can be measured by position instrumentation, or can be estimated by some sophisticated techniques such as multi-frame bundle adjustment that is used in photogrammetry. Our goal is to generate seamless, continuous stereoscopic image mosaics that exhibit correct 3D views before we have a 3D DEM.

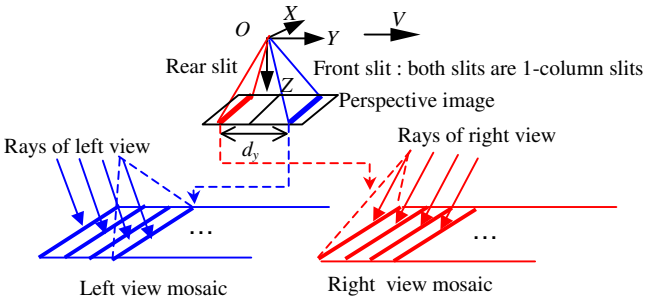
For generating seamless 2D mosaics from a hand-held camera, Shum & Szeliski (1998) use a local alignment (deghosting) technique which warps each image based on the results of pairwise local image registrations. The purpose of this deghosting process is to compensate for small amounts of motion parallax introduced by small translations of the camera, which are deviations from the pure parallax-free motion model. In more general cases of motion, for generating image mosaics with parallax and associated 3D recovery, several other techniques are proposed to explicitly estimate the camera motion and residual parallax, in which a projective depth value is recovered for each pixel of every frame (Kumar et al., 1995; Sawhney 1994; Szeliski and Kang 1995). Among them the most related work is the "3D mosaics plus parallax" approach by Kumar et al (1995) in which synthetic images from a reference view are generated from any other pairs of known images. However, first, it is computational intensive. In order to construct a 3D corrected mosaic, at least three views of the scene are needed, which should partially overlap with each other. It means that for a long image sequence, a dense parallax map needs to be calculated for every pair of frames, and the relation between a pair and the previous mosaic need to be estimated. Second, since the final mosaic is represented in a reference perspective view, there could be serious occlusion

problems due to large viewpoint differences between the reference view and the rest of the views in the image sequence.

In summary, there are several drawbacks in the existing approaches to generate stereo mosaics or 3D mosaic maps: (1) the dominant motion is constrained to rotation in stereo mosaicing; (2) the computation is extensive in the "planar + parallax" approach to generate 3D mosaics; (3) the geometric representation of 3D mosaics is inadequate. A "planar + parallax" model may not be an effective way to represent a forest scene with irregular depth changes. Both a perspective and an orthogonal representation need to estimate dense 3D information of the images or to provide a DEM in advance, and cannot avoid the serious occlusion problem when the viewpoints of the original image sequence change dramatically.

GEOMETRY, REPRESENTATION AND PROPERTIES

We have realized that the first important thing for 3D mosaic of a perspective image sequence is a new representation that can support seamless mosaicing and also can capture the inherent 3D information during the mosaic process. A ***parallel-perspective*** projection model is selected for representing mosaics in our approach since it is the closest form to the original perspective video sequence under large motion parallax, yet it's geometry allows us to generate seamless 3D mosaics.



(a) Stereo mosaic geometry

Parallel-Perspective Stereo Equation

$$Z = H\frac{b_y}{d_y} = H(1+\frac{\Delta y}{d_y})$$

Z: distance of a scene point to camera (depth)
d_y : distance betw. two slit windows (disparity)
b_y :"scaled" version of the baseline
Δy: stereo displacement of the point in mosaics
H : average distance of the terrain to camera

(2) Stereo mosaic equation

Fig. 1. Parallel-perspective stereo geometry: In each mosaic, each column is a full perspective projection, but successive columns have different viewpoints. Consequently, the depth accuracy is independent of absolute depths.

Here is the basic idea of the stereo mosaics (Fig. 1): Let us first assume the motion of a camera is a 1D translation, the optical axis is perpendicular to the motion, and the frames are dense enough. Then, we can generate two spatio-temporal images by extracting two columns of pixels near the two borders of the image that are perpendicular to the motion. These mosaic images thus generated are similar to *parallel-perspective* images captured by a linear pushbroom camera (Gupta & Hartley, 1997), which has *perspective projection in the direction perpendicular to the motion and parallel projection in the motion direction*. In contrast to the common pushbroom aerial image, these mosaics are obtained from two different oblique viewing angles of a single camera's field of view, one set of rays looking forward and the other set of rays looking backward, so that a stereo pair of left and right mosaics can be generated as the sensor moves forward. Since a fixed angle between the two viewing rays is selected for generating the stereo mosaics, the "disparities" of all points are fixed; instead geometry of optimal/adaptive baselines for all the points is created. In other words, for any point in the left mosaic, searching for the match point in the right mosaic means finding an original frame in which this match pair has a fixed disparity and hence has an adaptive baseline depending on the depth of the point.

We first give a real example to show the usefulness of parallel-perspective stereo mosaics for both 3D reconstruction and image-based rendering of large-scale terrain. Fig. 2a and 2b shows the results of stereo mosaics from a 166-frame video sequence, collected as part of a project with the Smithsonian Institute, UCSB and I.N.P.E. for studying the dynamics of forest fragmentation in the Amazon rain forest. The dense depth map reconstructed from the stereo mosaics is shown in Fig. 2c and a window of stereoscopic viewing is shown in Fig. 2d. In this example, the dynamic range of the displacements for most of the points is from -10.0 to +10.0 pixels with sub-pixel accuracy. The average height of the airplane is H = 385 m, and the distance between the two slit windows is d_y = 160. Using equation in

Fig. 1b we can estimate that the range of depth variations of the forest scene (from the fixation plane) is from -24.0 m (tree canopy) to 24.0 m (the ground).

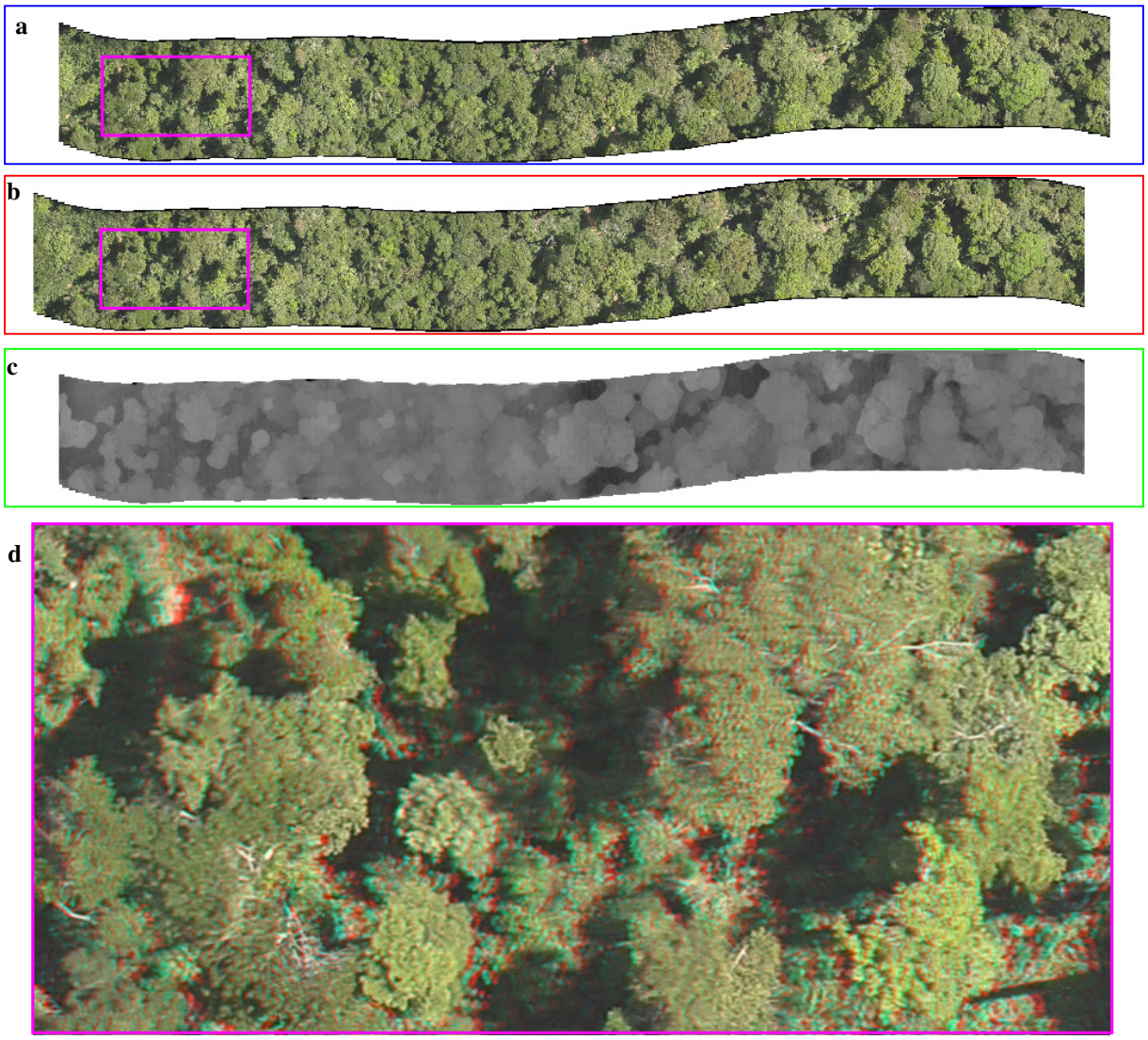


Fig. 2. Stereo mosaics and 3D reconstruction of a 166-frame telephoto video sequence. The size of each of the original mosaics is 7056*944 pixels. (a) left mosaics (b) right mosaics (c) depth map (stereo displacement Δy from 33 to -42 is encoded as brightness from 0 to 255) (d) stereoscopic viewing using left-blue/right-red glasses. Better effect may be achieved by choosing a suitable zoom factor (100% or 200%)

To summarize, there are several novel features and advantages of the parallel-perspective stereo mosaics:

1) *Compact visual representation:* Stereo mosaics are a compact representation for long video sequences. The size of a mosaic is only a function of terrain coverage and image resolution.

2) *Adaptive baseline:* The "baseline adaptation" of the multi-perspective stereo can make full use of an image sequence by setting the "disparity" as a fixed number (and as large as possible), and the depth accuracy is independent of absolute depth of a scene point.

3) *Efficient Computation*. The computation for generating stereo mosaics and recovering 3D information from a pair of stereo mosaics is only the function of the sizes of the stereo mosaics, which are independent to the number of video frames (see next section for a detail analysis).

4) *Parallel-perspective epipolar geometry*: In the case of a pure 1D translation, the epipolar lines are along the y direction in the stereo mosaics. Even in the more general case, epipolar curve geometry can be nicely derived so that only a 1D search is needed for stereo correspondence matching.

5) *Multi-viewpoint stereoscopic viewing:* Multi-perspective stereo mosaics can serve as an image-based representation for interactive 3D viewing. Without any computation for 3D recovery, realistic 3D perception by a human viewer can be achieved, while also allowing the viewer to change viewpoint over a large scale. This enables environmental researchers to estimate crown diameters of trees more effectively for a biomass study.

STEREO MOSAICING FROM REAL VIDEO

We have proposed a novel "3D mosaicing" technique called PRISM (parallel-perspective ray interpolation for stereo mosaicing) (Zhu et al, 2001b) to efficiently convert the sequence of *perspective* images with 6 DOF motion into the parallel-perspective stereo mosaics. In the PRISM approach for large-scale 3D scene modeling from real video, the computation of "match" is efficiently distributed in three steps: camera pose estimation, image mosaicing and 3D reconstruction. In estimating camera poses (for image rectification), only sparse tie points widely distributed in the two images are needed. In generating stereo mosaics, matches are only performed for parallel-perspective rays between small overlapping regions of successive frames. In using stereo mosaics for 3D recovery, matches are only carried out between the two final mosaics. This section gives a brief summary of the techniques in the three steps. Algorithms and discussions in detail can be found in our related papers (Zhu et al, 2001a, 2001b and 2001c).

Pose Estimation and Image Rectification

The stereo mosaicing mechanism can be generalized to the case of 3D translation if the 3D curved motion track has a dominant translational motion for generating a parallel projection in that direction. Under 3D translation, seamless stereo mosaics can be generated in the same way as in the case of 1D translation. The only difference is that viewpoints of the mosaics form a 3D curve instead of a 1D straight line. Further, the motion of the camera can be generalized to a 6 DOF motion with some reasonable constraints on the values and rates of changes of motion parameters of a camera(Fig. 3a), which are satisfied by a sensor mounted in a light aircraft with normal turbulence. There are two sub-steps necessary to generate a rectified image sequence that exhibits only 3D translation, from which we can generate seamless mosaics:

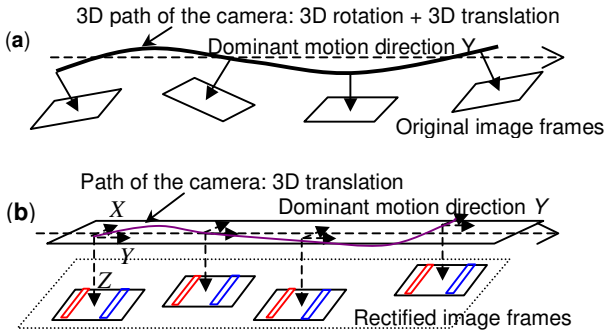


Fig. 3. Image rectification. (a) Original and (b) rectified image sequence.

1) *Camera orientation estimation*. Assuming an internally pre-calibrated camera, the extrinsic camera parameters (camera orientations) can be determined from our aerial instrumentation system (GPS, INS and a laser profiler) (Schultz et al 2000) and a bundle adjustment technique (Slama 1980). The implementation needs further efforts; the main point here is that we do not need to carry out dense match between two successive frames. Instead only sparse tie points widely distributed in the two images are needed to estimate the camera orientations.
2) *Image rectification*. A 2D projective transformation is applied to each frame in order to eliminate the rotational components(Fig. 3b). In fact we only need to do this kind of transformation on two narrow slices in each frame that will contribute incrementally to each of the stereo mosaics. The 3D motion track formed by the viewpoints of the moving camera will have a dominant motion direction (*Y*) that is perpendicular to the optical axis of the "rectified" images.

Local Matching and Ray Interpolation

Due to large and varying displacements between each pair of successive frames in the image sequence, extracting one-column slices from each frame is not sufficient to form uniformly dense mosaics. There have been two existing approaches to solving this problem. In a "manifold mosaic" (Peleg & Herman, 1997), each image contributes a slice to the mosaic. The width of the slice is a function of the displacements between frames. For a translating camera, manifold mosaic can be modeled as a *multi-perspective image*: Each sub-image (with more than one columns) taken from the original image is of full perspective, but sub-images from different frames will have different viewpoints (Fig. 4a). This may cause seams in the mosaic due to motion parallax under translation. Another approach is the "3D mosaic + parallax" approach (Kumar et al, 1995). In order to construct a 3D corrected mosaic, synthetic images from a reference *perspective view* are generated from any other pairs of known images. It means that for a long

image sequence, a dense parallax map needs to be calculated for every pair of entire frames, and the relation between a pair and the previous mosaic need to be estimated.

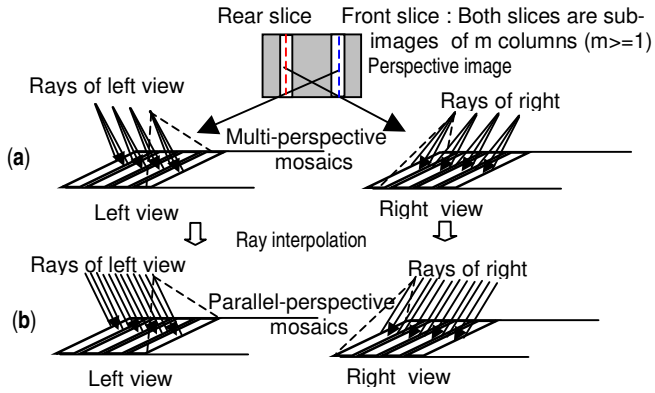


Fig. 4. Dense stereo mosaics. In each multi-perspective projection mosaic, each sub-image (with more than one columns) taken from the original image is of full perspective, but sub-images from different frames will have different viewpoints. This may cause seams in the mosaic due to motion parallax. However, in a parallel-perspective projection mosaic, each column is of full perspective with a moving viewpoint. It implies that a perspective sub-image in the multi-perspective mosaic is transformed into a parallel-perspective sub-image during local registration and ray interpolation.

How can we generate seamless mosaic in a computational effective way? The key of our approach lies in the parallel-perspective representation and the PRISM approach. For each of the left and right mosaics, we only need to take a slice of certain width from the corresponding left (or right) oblique viewing direction from each frame, and perform local registration between the overlapping slices of successive frames, then generate parallel *interpolated rays* between two known discrete perspective views (Zhu et al, 2001a; 2001b). Our approach is similar to view synthesis by image interpolation, which has been well-studied in image-based rendering (e.g. Seitz & Dyer, 1995). Fortunately in our case, the distance between two successive views are small (and not large enough for accurate 3D estimation), thus the synthetic rays of parallel-perspective projection between the two known views reflect the 3D structure of the scene, and more importantly do not have serious occlusion problem. In addition, since we only need to stitch two narrow slices from two successive images, the match of points is only carried out in small overlapping regions of the two images. As a result, each mosaic is a parallel-perspective composite image, the viewpoints of which are on a smoothly (interpolated) 3D curve. This is where a parallel-perspective representation is superior to a representation based on a full perspective projection or a full parallel projection.

Feature matching and slice triangulation - In the fast PRISM algorithm (Zhu et al, 2001b), we only match pairs of selected feature points ("tie points") in the overlapping slices of two successive frames, and calculate the destination coordinates of the interpolated rays in the mosaics. Then image slices are divided into small triangles on these tie points and each triangle is warped to the mosaics by an affine transformation. So the first and one of the most important steps in constructing a parallel-perspective mosaic is to choose the right set of feature points for local warping. It is necessary to make sure that the feature points are on the boundaries of depth change. This helps ensure that after triangulation every triangle covers a planar surface. This condition is essential in order to validate the affine transformation of each triangle. In our current implementation, feature points are chosen to be points along a "stitching line" between the two successive frames, where the brightness value change between consecutive pixels is more than a predefined threshold. The sudden changes in brightness value can be considered to be edges where we can have depth changes. Once the feature points are selected in one slice, the corresponding point is searched in the overlapping region of the consecutive slice. This is achieved by finding the expected position of a feature point in the overlapping region of the next slice and then doing a cross-correlation search in a region around the expected position. The expected position can be calculated using the translation information between the two frames.

A further refinement step to maintain this condition (of planar surface for triangles) is to break the image slice along the width into further smaller slices (hence multiple stitching lines between two frames) so that the triangles formed are smaller in size and there is a better chance that each individual triangle is a planar patch. These triangles however, require different affine transformation functions, therefore the computation may be more significant.

Epipolar Geometry and Stereo Match

We all know the dilemma of baseline versus common field of view (FOV) in a classic stereo/motion approach for each pair of images in a video sequence: A larger common FOV means a shorter baseline causing insufficient parallax and associated disparity. But a longer baseline means a smaller common FOV for 3D construction. A multiple-framers / multiple-baseline approach is one of the possibilities to handle this problem, but it is not an efficient way since dense matches for multiple frames hence intensive computations are needed. Stereo mosaic methods solve the baseline-FOV dilemma efficiently by extending the FOV in the direction of the dominant motion

that serves as the main baseline. From the same image sequence of a single video camera, a stereo pair of extended mosaics that are virtually endless can be generated, hence producing almost the same FOVs in the two mosaics. More important, since a fixed angle between the two viewing rays is selected for generating the stereo mosaics, the disparities of all points are pre-selected (by mosaicing) and fixed; instead geometry of optimal/adaptive baselines for all the points is created. Thus one of the distinguish features of stereo mosaics is its geometry of *adaptive baselines*, compared to the fixed baseline and multi-baseline approaches.

Stereo mosaics have multi-perspective geometry - the viewpoints of both stereo mosaics will be on the same smooth 3D motion track. In the general case, the corresponding point in the right mosaic of any point in the left mosaic will be on an epipolar curve determined by the coordinates of the left point and the 3D motion track. In the ideal case where the 3D curve becomes a 1D straight line, the epipolar curves will turn out to be horizontal lines. The depth maps of stereo mosaics in our experiments were obtained by using the Terrest system designed for perspective stereo match (Schultz 1995) without modification. The Terrest system was designed to account the illumination differences and perspective distortion of stereo images with largely separated views by using normalized correlation and multi-resolution un-warping, thus it is suitable for stereo mosaics with two largely separated viewing directions. Further work is needed to apply the epipolar curve constraints into the search of correspondence points in the Terrest to speedup the match process. Currently we perform matches with 2D search regions estimated from the motion track and the maximum depth variations of a scene. Fig. 2c shows the depth map where the stereo displacements are encoded as brightness. Afterward, the parallel-perspective mosaic (with both depth and texture) can be easily mapped into a DEM if so desired (Zhu et al 2001a).

Storage and Computational Efficiency

Assume that we have F frames of images, each frame of size M (rows) × N (columns) and each of the two mosaics is of size M (rows) × L (columns). Suppose the average overlap of successive frames is r (percent). Then the compression ratio is c = 1/2(1- r) (the factor 2 comes from 2 mosaics). For example, if r = 95%, then c = 10. The length of each mosaic L is approximately (1-r)NF. T*he length (number of the columns) of a mosaic is only a function of image resolution and the coverage of the scene, since neither mosaic has redundant information.* It is also true for the computation in generating and using the stereo mosaics except in the first step (camera pose estimation), since the computation of all the other steps only depends on the size of the stereo mosaics. The computation in each step of stereo mosaics is as follows.

1) Camera pose estimation: match of sparse tie points in each pair of successive frames
2) Image rectification: Data amount to be processed 2×2×M×L pixels (same as the next)
3) Image mosaicing: Data amount to be processed 2×2×M×L pixels (**2** mosaics of size **M×L**; for each mosaic, each pixel is a registration and interpolation from **2** points of successive image pair, which are transformed from original video frames in step 2)
4) Correspondence: Data amount to be processed 2×M×L pixels (only two images!)

Generally speaking, the computation complex of stereo mosaics is 2(1-r) of a multi-frame approach. For example, if r = 95%, it is 2(r-1) = 10%. Moreover, stereo mosaics provide us a unique way to neatly organize and represent video sequences with 3D information.

IMPLEMENTATIONS AND RESULTS

The geographical data ("geo-data") from our aerial instrumentation package includes 3D locations from the GPS, 3D rotation angles from the INS, and the range data along the flight path from a laser range profiler. In a real application, the motion parameters (geo-data) provided by our measurement instrumentation may not be accurate enough due to the sensor resolution and noise. In some circumstances, we may not have *a priori* motion parameters at all. In both cases, the complete solution of the problem becomes difficult because it turns out to require bundle adjustment of photogrammetry for multiple (many!) frames. In the following sub-sections, we will show how we can obtain reasonably good results while greatly simplifying the problem thus near real-time is possible. First, without any geo-data, we can model the scene as piecewise planar surface and a projective model is used for frame registration. In this so-called *free-mosaic method*, camera "calibration" is carried out implicitly during the image registration and mosaicing stage. If geo-data is available (even though it is inaccurate), then geo-corrected mosaics might be obtained. However, careful calibration of the intrinsic camera parameters (at least of the focal length) is needed in this case. Furthermore the motion parameters from GPS/INS/laser range profiler could be refined by a

motion refinement algorithm based on bundle adjustment and Extended Kalman Filter (EKF) approaches, which have had extensive work in both theory and practice (e.g. Gelb 1974; Broida & Chellappa 1991; Triggs et al, 1999).

Free Mosaics in the Absence of Geo-Data

In some applications such as image-based rendering, photometric and 3D geometric realism is the main goal rather than geo-referenced accuracy. In this case, under some assumptions of more constrained motion, we wish to generate seamless, continuous stereoscopic image mosaics that exhibit correct 3D views before we have recovered a 3D DEM. Fig. 5 shows mosaic results where camera orientations were estimated by registering the planar ground surface of the scene via dominant motion analysis. Global transformation parameters between all the frames and a reference frame (e.g. the first frame), defined as the "image track", are computed from the interframe motion parameters. Here we show the effect of seamless mosaicing with out PRISM approach. Please compare the results of 3D mosaicing (parallel-perspective mosaicing) vs. 2D mosaicing (multi-perspective mosaicing) by looking along many building boundaries associating with depth changes in the entire 4160x1536 mosaics at our web site (Zhu 2001d). Since it is hard to see subtle errors in the 2D mosaics of the size of Fig. 5a, Fig. 5b and Fig. 5c show close-up windows of the 2D and 3D mosaics for the portion of the scene with the tall Campus Center building. In Fig. 5b the multi-perspective mosaic via 2D mosaicing has obvious seams along the stitching boundaries between two frames. It can be observed by looking at the region indicated by circles where some fine structures (parts of a white blob and two rectangles) are missing due to misalignments. As expected, the parallel-perspective mosaic via 3D mosaicing method (Fig. 5c) does not exhibit these problems.

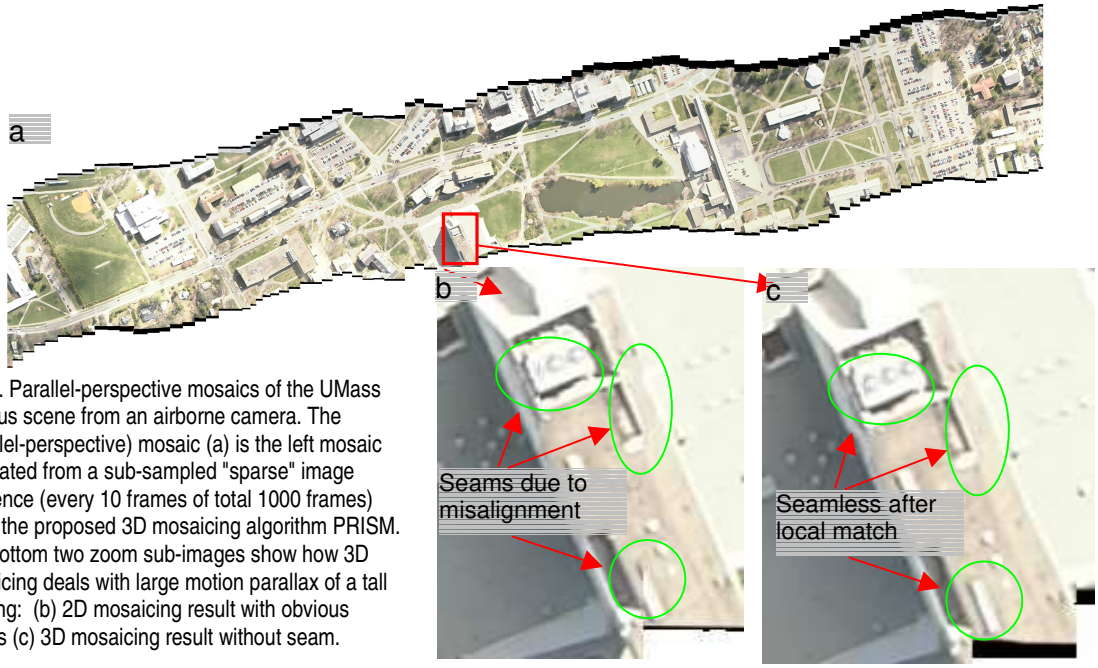


Fig. 5. Parallel-perspective mosaics of the UMass campus scene from an airborne camera. The (parallel-perspective) mosaic (a) is the left mosaic generated from a sub-sampled "sparse" image sequence (every 10 frames of total 1000 frames) using the proposed 3D mosaicing algorithm PRISM. The bottom two zoom sub-images show how 3D mosaicing deals with large motion parallax of a tall building: (b) 2D mosaicing result with obvious seams (c) 3D mosaicing result without seam.

Real-Time Geo-Referenced Mosaics with Inaccurate Pose Data

The free mosaic algorithm can generate a rather "realistic" pair of stereoscopic mosaics with low computational cost. However, the mosaics may not be faithful to the geo-referenced path due to the motion/scene model simplification and error accumulation. For example, even sub-pixel errors between two successive frames in the free-mosaic may lead to a drift from the correct path of up to hundred pixels at the end of a thousand-frame video sequence, even if the relative depth perception is good for 3D viewing as well as relative 3D reconstruction. On the other hand, for example, the GPS system used in our previous data collection only gives us an accuracy of about one meter in the ground, which means 5 to 10 pixels in the images of our telephoto camera. So it is impossible to be directly used for image rectification and seamless mosaicing since the relative orientation errors between two successive frames are obvious. Fortunately, errors in GPS locations and INS orientations do not accumulate in time, although the absolute values of the errors are large. Thus it is possible to combine the two different kinds of measurements from geo-data and image match with different error properties.

We have implemented a fast algorithm to obtain seamless and geo-corrected mosaics (Zhu et al 1999b). The key of our program lies in a frame-by-frame two-track geo-mosaic composition method that achieves the required mosaic

fidelity even if the geo-data is not accurate. First the geo-data is filtered using a simple Kalman filter. Then the location, orientation and range parameters of all camera positions are used to compute a "geo-track" on the ground in order to warp the original frames to rectified frames, and to localize each frame into the mosaic. A "stitching track", which is computed from the "geo-track" and the "image track" (from free-mosaicing), is used to stitch two slices from two successive frames seamlessly. The two-track algorithm is in fact a simplified version of the PRISM method, where global interframe motion estimation is used instead of point-wise feature matching. No complex global optimization is used, and the algorithm is robust and fast. With no effort yet expended on code optimization and speedup, the video alignment step of our experimental prototype system takes only about 1 second for a pair of 720*480 color images on a Pentium II 400 MHz PC, and the geo-correction and mosaic generation steps take far less time. The system has the potential to process data at video rate in the future.

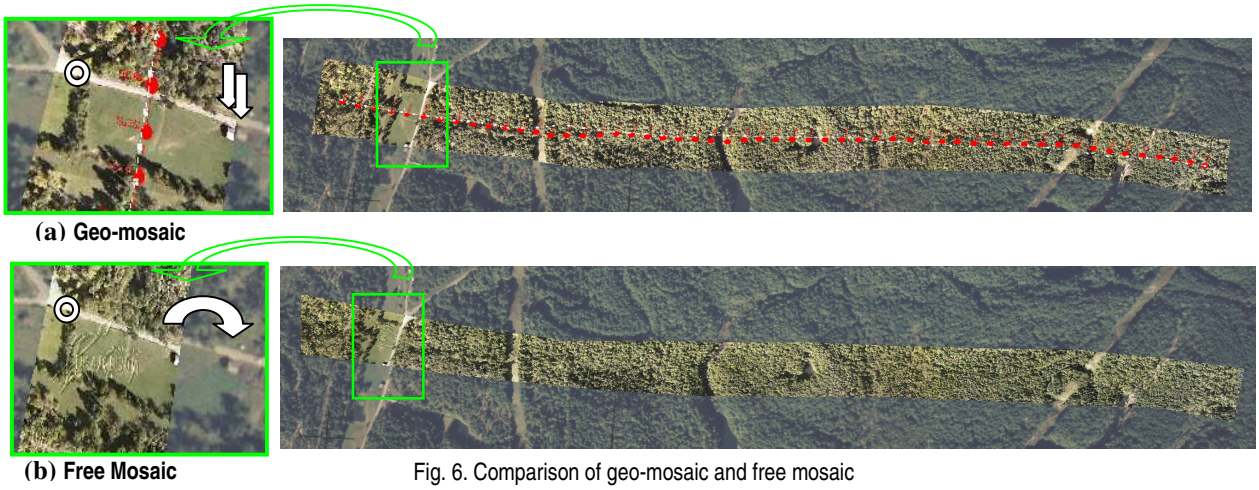


Fig. 6. Comparison of geo-mosaic and free mosaic

Fig. 6. shows the geo-referenced mosaic of a 53-frame sub-sampled video sequence. The two-track method corrects the accumulating errors in image matches frame by frame, while maintaining precise stitching of successive frames. For comparison, we also generated a free mosaic using the evaluation version of well-developed commercial software VideoBrush 2.0. From their paper (Sawhney et al, 1998) and the mosaic result, it appears that no scaling is applied to the mosaic. While the mosaic is seamless, and color/ luminance blending is handled, it is not geo-corrected; for example the scale is not changed according to the range of the ground. To evaluate our geo-mosaic result, we superimpose both our geo-mosaic and VideoBrush's mosaic in the high altitude image that is taken at 5k ft at about the same time as the video sequence. This high-altitude image has then been warped into a geo-reference frame, which is an approximation of a parallel projection. To register the mosaics with the high-altitude image, we manually select two points in the high-altitude image and find corresponding points in both the geo-mosaic and VideoBrush mosaic respectively, and use 2D rigid transformation of translation, rotation and scaling. Fig. 4 also shows a comparison of close-ups of the same portion of mosaics. Image points along the center track in our geo-mosaic register precisely with the high-altitude image, and at the border of the mosaic there are only small errors (due to distortions in both the mosaics and the high altitude image). Registration between the geo-mosaic and an available high-altitude image could improve the accuracy of the geo-referenced mosaics. As expected, the VideoBrush mosaic does not register with the high-altitude image well, since it does not change the scaling of the images according to the depth of the scene (Fig. 4b). It should be noted here that two feature points are selected in the head and tail of both of the mosaics; one of them is under the white circle (O) in each of the sub-images in Fig. 4. Notice the obvious different location errors in the geo-mosaic and the free mosaic of a white building (pointed by arrows in the close-up windows) below the road in the right of each close-up image. Remember that the video used in this example is only about 1 minute, but we are carrying out experiments on much longer image sequences in actual environmental applications. We have produced the first geo-registered aerial maps over the Amazon Basin with the geometric accuracy of 5 – 10 m in the telephoto mosaics using the proposed method.

From Free Mosaic to Geo-Mosaic

The geo-referenced mosaic generated by the two track algorithm on a frame-by-frame base has the problem of local distortion if the relative errors in the measurements of the GPS and INS systems are large. This makes the extraction of the digital elevation maps(DEM) from the mosaics using stereo matching difficult, as the disparity information gets hidden in the noise generated by the GPS. An alternative way to generate a "globally" geo-referenced mosaic is to warp a free mosaic that has been built to minimize the stitching error. The solution proposed is to find

corresponding control points in the free and the geo-mosaic using the correspondences between the geo-track and the image track, and "globally" warp the free mosaic to its geographical coordinates to generate the geo-mosaic. The resulting geo-mosaic will retain the smoothness of the free mosaic but at the same time be in the correct geographical location. The global warping means that we should not use all the corresponding pairs of coordinates of the original frame positions in the free and geo-mosaic but instead select coordinates for warping which are suitably far from each other so as to average out the error in individual GPS measurements. If instead we had used all the coordinates in warping then the resulting geo-mosaic will be similar to the geo-mosaic generated by the frame-by-frame two-track method. Also if we had selected the points to be too far off from each other, then error in individual pixels in the mosaic could be unacceptably high.

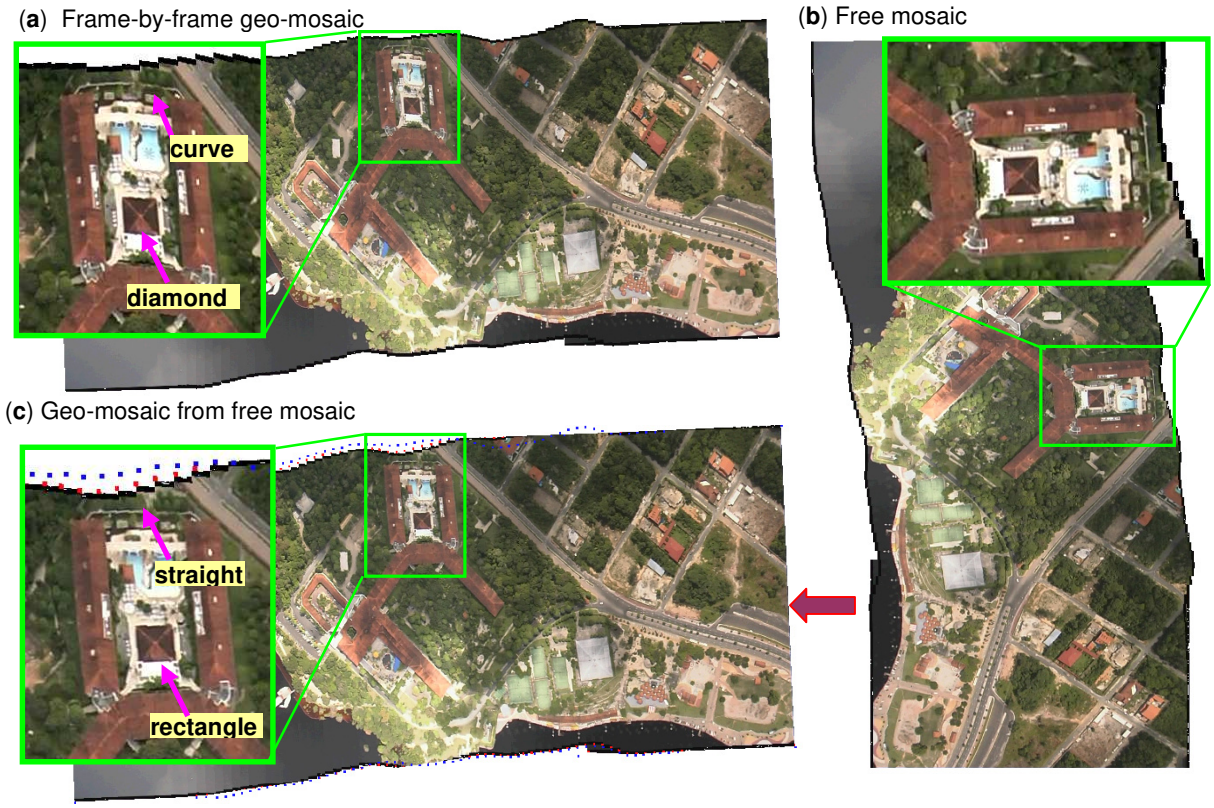


Figure 7: Construction of a geo-referenced mosaic from a free mosaic.

Thus we find that the choosing of the control points for global warping is very important. The coordinates of the control points are chosen to be as far apart as possible subject to the constraint that the coordinates in each warped region between two sets of control points have error below a "tolerance limit", where the errors are defined to be the differences between the expected geo-track from the geo-data and the "warped track" from the free mosaic. Now the new problem that arises is as to how much error is "tolerable". Different image sequences are found to have varying tolerance limits. It has been observed through experiments that when the slice contributed by individual frames is small, we need to have a higher value for tolerance to maintain smoothness of the mosaic since dense image sequence with slow movement is subject to more localizing errors in the GPS. This was observed in the video sequence in the Brazil data set that contains building and roads making it easy to detect any non-smooth regions (Fig.7). Fig 7a shows a geo-referenced mosaic generated by the two-track approach where the camera traveled mainly from west to east in the geo-referenced coordinate system. Errors of the geo-data introduce local distortions in the frame-by-frame geo-mosaic. For example, a square-shaped roof became diamond-shaped and a straight line turned to a curve. Fig. 7b shows the free mosaic of the same sequence without using any geo-data , although not geo-referenced (so the mosaic runs in the vertical direction), the shapes and smoothness of the structures are preserved. Fig. 7c shows the globally geo-referenced mosaic warped from the free mosaic, where control points of only five locations of geo-data are used. It can be seen that the straight line and the square keep their shapes. The red and blue dots in the upper and lower boundaries of Fig 7c represented the expected geo-track and the real warped image track respectively (In the frame-by-frame geo-mosaic, they should be identical).

CONCLUSIONS AND DISCUSSIONS

We define a 3D mosaicing as the generation of a stereoscopic mosaic pair, both mosaics are parallel-perspective, from an image sequence captured by a camera with 3D rotation and 3D translation. This definition of 3D mosaics includes two aspects: (1) A 3D mosaicing process consists of a global image rectification followed by a fine local transformation that accounts for the motion parallax due to *3D structure* of a scene. (2) The final mosaics are a stereo pair that embeds *3D information* of the scene. In the proposed stereo mosaicing approach, the computation of "match" is efficiently scattered in three steps: camera pose estimation, image mosaicing and 3D recovery. In estimating camera poses, only sparse tie points widely distributed in the two images are needed. In generating stereo mosaics, matches are carried out between small overlapping regions of successive frames. In using stereo mosaics for 3D recovery, matches are carried out between the two final mosaics, which is equivalent to finding a matching frame for every point in one of the mosaics with a fixed disparity. Thus stereo mosaics using parallel-perspective projection are a compact and efficient way to represent 3D information of a scene over a large scale under a general motion model of a camera. Here we discuss several important issues on stereo mosaicing that need further study.

Camera Orientation Estimation and Accurate Geo-Referenced Mosaics

In a real application, the motion parameters (geo-data) provided by the measurement instrumentation may not be accurate enough due to the sensor resolution and noise. In some circumstances, we may not have *a priori* motion parameters at all. In both cases, the complete solution of the problem becomes difficult because it turns out to require bundle adjustments of photogrammetry for multiple (many!) frames. While many efforts have been made in multiple frame bundle adjustments (Triggs, et al 2000), the extension to a long image sequence remains to be a challenging problem in both computation and robustness. In our PRISM approach, the bundle adjustments are only used for (relative) camera orientation estimation for stereo mosaicing, these could greatly reduce the burden in computation. The robust and accurate camera orientation estimation is essential for accurate geo-referenced mosaics. Until then we can really analyze errors in geo-mosaics and 3D reconstruction with ground truth.

View Interpolation for Both Geometric and Photometric Seamlessness

It has been observed that using local warping to produce parallel-perspective mosaics helps remove geometric seams due to motion parallax. The resulting mosaics are smooth even in the case of sparse image sequences. However, the performance of this algorithm depends highly on the quality of matches of the feature points. Any mismatch in the feature points will produce a distinctly visible error in the mosaic. Further, consecutive frames in an image sequence can differ in their brightness levels and/or colors. The resulting mosaic in such a case although being geometrically smooth can still be seen to have photometric seams. This visual artifact can be removed by the use of blending (e.g. Burt & Adelson 1983).

Using Structural Information and Photographic Image in Geo-Mosaicing

Further approach to generate smooth and shape-preserved geo-mosaics could exploit the fact that human eye is more sensitive to the errors around edges. Thus smoothing of the edges could enhance the quality of a mosaic generated from an image sequence. We could use the original images from the sequence and run an edge detector on them to get the edges in the original images. Since we have known the transformations from the images to the mosaic, we can effectively build a virtual “edge mosaic” consisting of all the edges in the sequence. An edge detection algorithm could then be run on the real mosaic and corresponding edges be found between the virtual edge mosaic and the real mosaic. We could then transform curves in the real mosaic corresponding to the edges in the edge mosaic in order to obtain a smoothed and shape-preserved geo-mosaic.

On the other hand, if a high-altitude photographic image is available in the same area, we could run a registration between the mosaic and the photographic image. The match should be performed in feature spaces instead of the original images since the resolution and the photometric appearances could be significantly different between video images and photographic images. Then we can warp the mosaic to the ortho-rectified photometric image.

ACKNOWLEDGEMENTS

This work is supported by NSF EIA-9726401, and NSF CNPq EIA9970046. Additional funding is supported by TNC in conjunction with the Winrock Corporation and DURIP equipment grant (DAAD19-99-1-0016).

REFERENCES

Broida T. J., Chellappa R. (1991), Estimating the kinematics and structure of a rigid object from a sequence of monocular images, *IEEE Trans PAMI*, 13(6): 497-513.

Burt P., Adelson E. (1983), A multi-resolution spline with application to image mosaics, ACM Trans. Graphics, 2(4): 217-236.

Chen S. E. (1995), QuickTime VR - an image based approach to virtual environment navigation, *Proc. SIGGRAPH 95*:29-38.

Gelb A. (1974), *Applied Optimal Estimation*, MIT Press, Cambridge, MA.

Gupta R. , Hartley R. (1997), Linear pushbroom cameras, *IEEE Trans PAMI*, 19(9): 963-975

Huang H.-C. and Hung Y.-P. (1998), Panoramic stereo imaging system with automatic disparity warping and seaming, *Graphical Models and Image processing*, 60(3): 196-208.

Ishiguro H. , Yamamoto M. , and Tsuji S. (1990), Omni-directional stereo for making global map, *Proc. IEEE ICCV'90*, 540-547.

Kumar R. , Sawhney H. , Asmuth J., Pope J. and Hsu S. (1998), Registration of Video to Geo-referenced Imagery, *Proc. IAPR ICPR98*, vol. 2: 1393-1400

Kumar R., Anandan P., Irani M., Bergen J. and Hanna K. (1995), Representation of scenes from collections of images, In *IEEE Workshop on Presentation of Visual Scenes*: 10-17.

Mann S. (1998) , Humanistic Intelligence: WearComp as a new framework for Intelligent Signal Processing, Proceedings of the IEEE, Vol. 86, No. 11, November, 1998, pp 2123-2151

Peleg S., Herman J. (1997), Panoramic Mosaics by Manifold Projection. *Proc IEEE CVPR'97*: 338-343.

Peleg S. , Ben-Ezra M. (1999), Stereo panorama with a single camera, *Proc IEEE CVPR'99*: 395-401

Sawhney H.S. (1994), Simplifying motion and structure analysis using planar parallax and image warping. ICPR'94: 403- 408

Sawhney H.S., Kumar R., Gendel G., Bergen J., Dixon D., Paragano V. (1998), VideoBrushTM: Experiences with consumer video mosaicing, *Prof. IEEE Workshop on Applications of Computer Vision (WACV)*: 56-62

Seitz S. M. and Dyer C. R.(1995), Physically-valid view synthesis by image interpolation, In *IEEE Workshop on Presentation of Visual Scenes*.

Schultz H. (1995). Terrain Reconstruction from Widely Separated Images, In *SPIE*. Orlando, FL

Schultz H., Hanson A., Riseman E., Stolle F., Zhu. Z. (2000), A system for real-time generation of geo-referenced terrain models, *SPIE Symposium on Enabling Technologies for Law Enforcement*, Boston MA, Nov 5-8, 2000

Shum H.-Y. and Szeliski R. (1997), Panoramic Image Mosaics, *Microsoft Research, Technical Report, MSR-TR-97-23*

Shum H.-Y. and Szeliski R.(1998), Construction and refinement of panoramic mosaics with global and local alignment, ICCV'98: 953-958.

Shum H.-Y. and Szeliski R. (1999), Stereo reconstruction from multiperspective panoramas, *Proc. IEEE ICCV99*: 14-21.

Slama C. C. (Ed.) (1980), Manual of Photogrammetry, Fourth Edition, *American Society of Photogrammetry*

Slaymaker D. M., Jones K. M. L., Griffin C. R. and Finn J. T.(1996), Mapping deciduous forests in Southern New England using aerial videography and hyerclustered multi-temporal Landsat TM imagery. Gap Analysis, Am. Society for Photo. and Remote Sens.

Szeliski R. and Kang S. B. (1995), Direct methods for visual scene reconstruction, In *IEEE Workshop on Presentation of Visual Scenes*: 26-33

Triggs B., McLauchlan P., Hartley R., and Fitzgibbon A. (2000). Bundle adjustment - a modern synthesis. In *Vision Algorithms: Theory & Practice*. Springer-Verlag, pp. 298--375

Xiong Y., Turkowski K. (1998), Registration, calibration, and blending in creating high quality panoramas, *Prof. IEEE WACV'98*: 69-74.

Zheng J. Y. and Tsuji S. (1992), Panoramic representation for route recognition by a mobile robot. IJCV 9(1), 1992, 55-76.

Zhu Z., Hanson A. R., Schultz H., Stolle F., Riseman E. M. (1999a), Stereo Mosaics from a Moving Video Camera for Environmental Monitoring, *First International Workshop on Digital and Computational Video*, December 10, 1999, Tampa, Florida, USA, pp. 45-54.

Zhu Z., Riseman E. M., Hanson A. R., Schultz H.(1999b), Automatic Geo-Correction of Video Mosaics for Environmental Monitoring, *Technical Report TR #99-28*, Computer Science Department, University of Massachusetts at Amherst, April, 1999

Zhu Z., Xu G., Lin X. (1999c), Panoramic EPI Generation and Analysis of Video from a Moving Platform with Vibration, *Proc. IEEE CVPR'99*: 531-537

Zhu Z., Xu G., Riseman E. M. and Hanson A. R. (1999d), Fast generation of dynamic and mutliresolution panorama from video sequences, *Proc. IEEE ICMCS'99*: 400-406.

Zhu Z., Riseman E. M., Hanson A. R. (2001a), Theory and practice in making seamless stereo mosaics from airborne video, *Technical Report #01-01*, CS Dept., UMass-Amherst, Jan. 2001.

Zhu Z., Riseman E. M., Hanson A. R. (2001b), Parallel-perspective stereo mosaics, In *ICCV'01*, Vancouver, Canada, July 2001

Zhu Z., Hanson A. R., Schultz H., Riseman E. M. (2001c), Error characteristics of parallel-perspective stereo mosaics, *IEEE Workshop on Video Registration* (with ICCV'01), Vancouver, Canada, July 13, 2001

Zhu Z.(2001d), PRISM: Parallel ray interpolation for stereo mosaics, http://www.cs.umass.edu/~zhu/StereoMosaic.html