Theory and Practice in Making Seamless Stereo Mosaics from Airborne Video[*]

Zhigang Zhu, Allen R. Hanson, Edward M. Riseman

UM-CS-2001-001

January 2001

Computer Vision Laboratory
Department of Computer Science
University of Massachusetts at Amherst
Amherst, MA 10003
{zhu, hanson, riseman}@cs.umass.edu

[*] This work is partially 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).

Keywords - Mosaicing, stereo vision, visual representation, epipolar geometry, image registration, view interpolation, airborne video analysis

Abstract

In this paper we present a novel method for automatically and efficiently generating stereoscopic mosaics by registration of optical data collected by a video camera mounted on an airborne platform that mainly undergoes translating motion. The resultant mosaics are seamless and will exhibit correct three-dimensional (3D) views. The basic idea is to fast construct stereo mosaics before 3D recovery for applications such as image-based rendering and environmental monitoring. Image mosaicing from a translating camera will result in a mosaic image with multiple viewpoints on a long motion path. It raises a set of different problems from that of circular projections of a rotating camera, including motion models, motion estimation, generation of geometric seamless image mosaic with large inter-frame motion parallax, and epipolar geometry of multiperspective projection.

First we will analyze two possible geometric representations for stereo mosaics - multiple-perspective mosaics and parallel-perspective mosaics. We will show that the parallel-perspective mosaic representation is superior to a multiple-perspective representation for the geometric seamlessness of its visual appearance and the simplicity of its epipolar geometry, which will benefit both interactive human viewing and automated stereo matching. Then 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. Next, a novel method based on interframe local matches and parallel-perspective view synthesis is proposed to generate dense and seamless parallel-perspective stereo mosaics in the presence of motion parallax. The advantages of stereo mosaics in representation, computation, 3D recovery resolution and 3D visualization are discussed throughout the paper. Experimental results on long video sequences will be given to validate our analysis.

1. Introduction

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 [1]-[5]. Creating stereo panoramas from two rotating cameras was proposed by Huang & Hung [6], and from one off-center rotating camera by Ishiguro, Yamamoto & Tsuji [7] , Peleg & Ben-Ezra [8], and Shum & Szeliski [9]. In these kinds of stereo mosaics, the viewpoint is limited to locations within a very small area, usually along a circle of a meter in diameter. Under more general motion other than rotation, a system for creating a global view for visual navigation was proposed by Zheng and Tsuji [10] by composing columns taken by a smoothly translating camera comprising only a vertical slit. The moving slit paradigm was used as the basis for manifold mosaicing [11], image-based rendering [12] and reconstruction of a 3D layered representation [13] for large-scale scenes.

As the motivation of our work, the paper is directed at the development of tools for environmental science and other aerial video applications requiring the mapping of information into absolute world coordinates. As a matter of fact, 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, being conducted jointly by researchers from the Computer Science Department and the Department of Natural Resources Conservation at the University of Massachusetts at Amherst, 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 utilized only a fraction of the available data due to the labor involved in hand interpreting the large amount of video data. For example, recent projects in Bolivia involved over 600 sites and more than 20 hours of video, and in Brazil 120 hours (10 terra bytes), which is prohibitive if the video is interpreted manually. A more compact representation and more flexible interactive 3D visualization interface are clearly necessary. Stereo mosaics - images with a large field of view and 3D perception capability- have proven to be very helpful for interpreting tree canopies. 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.

In fact, for many applications dealing with large-scale natural or urban scenes, extending the field of view (FOV) of a 2D image, and then introducing the third dimension of depth would be of great utility. Video surveillance [14], environmental monitoring [15][16], cultural scene archives and retrieval, robot navigation [10][13], and wearable camera systems [17] are just a few examples of the applications that would benefit from an extended and enhanced image-based representation. Theoretically, 3D reconstruction from pairs of images, and then integrating multiple 3D maps (as well as the texture maps) into a large texture-mapped digital elevation map (DEM) would be an ideal solution. However, there are many technical challenges that need to be solved, such as camera calibration, dense image correspondence, high performance computation, and consistency of 3D recovery across multiple frames. Consequently, the visual representation of large-scale 3D scenes is still an open problem.

In this paper we propose a novel method that automatically and efficiently generates large FOV stereo mosaics from a moving camera, which allows viewpoint changes over a large scale. The mosaics are both seamless and preserve 3D information for stereo viewing or later 3D recovery. In Peleg & Ben-Ezara' work on circular projections of stereo panorama [8], the authors mentioned that their technology could also be used to create stereo panoramic images from a translating camera. However, as far as we know, there is little serious work on stereo mosaics from a translating camera, which is the typical situation prevalent in airplane motion during forest surveys [15] and vehicular motion during navigation on roads [10][13]. Our previous work on airborne video stereo mosaic has shown that image mosaicing from a translating camera will result in a mosaic image with multiple viewpoints on a long motion path [16]. It raises a set of different problems from that of circular projections of a rotating camera, including motion models, motion estimation, generation of seamless image mosaics with large inter-frame motion parallax, and epipolar geometry of multiperspective projection.

Generally speaking, our stereo mosaicing is a 3D mosaicing process in which image mosaics preserve 3D information, and are generated from an image sequence of an arbitrary scene when motion parallax is apparent. Obviously use of standard 2D mosaicing techniques based on 2D image transformations such as a manifold projection [11] cannot generate a geometric seamless mosaic in the presence of large motion parallax, particularly in the case of surfaces that are highly irregular or with large different heights. Most research on seamless mosaics deals with video from a rotating camera. As a typical example, in generating seamless 2D mosaics from a hand-held camera, Shum & Szeliski [18] used a local alignment (de-ghosting) technique to compensate for the small amount of motion parallax introduced by small translations of the camera. In more general cases for generating image mosaics with parallax, several techniques have been proposed to explicitly estimate the camera motion and residual parallax by recovering a projective depth value for each pixel [19][20][21]. The most related work is the "3D mosaics plus parallax" representation [19] based on the “plane + parallax” approach. In order to construct a 3D corrected mosaic, at least three views of the scene are needed, which should partially overlap with each other. 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 each pair and the previous mosaic need to be estimated. Since a final mosaic is represented in a reference perspective view, there could be serious occlusion problems due to large viewpoint differences between a single reference view and the rest of the views in the image sequence.

An important part of our approach is a new mosaic representation that can well support large seamless mosaicing and also can capture the inherent 3D information during the mosaic process. Our interest in a new mosaic representation of a 3D scene is partially inspired by the techniques used in classical Chinese paintings. It is a common remark that the classical Chinese paintings are without a correct vanishing point and possess no exact laws for foreshortening of figures [18]. But this characteristic is nothing more than a convention by which artists endeavor to represent 3D objects. In a typical Chinese painting, "the scenery changes as one walks". A concrete reflection of this multi-focal perspective (i.e., no fixed vantage point) and connotative space can be seen in the structure of horizontal landscape scrolls and their seemingly endless extension. In this paper a *parallel-perspective* 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 its geometry allows us to generate seamless stereo mosaics.

There are other aspects that can be addressed. In some applications, such as environmental monitoring (e.g. estimation of the biomass of forests) and aerial surveillance, geo-referenced mosaics tied to real world coordinates are required. For example, Kumar, et al [14] presented a geo-registration method that can register video mosaics to a high-altitude reference image using the 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) in their approach. Our stereo mosaics will be geo-referenced naturally if we assume that the motion of the camera can be measured by position instrumentation, or can be estimated by some sophisticated techniques such as multi-frame bundle adjustment [23]. In other 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. Technically speaking, camera orientation estimation for geo-reference mosaicing remains challenging in both theory and practice, and we will leave this for future research.

In summary, the aim of stereo mosaicing in this paper is to discuss representations, conditions and algorithms for generating a stereo mosaic pair with extremely wide FOV that allows later 3D recovery of the terrain. A second goal is to use just two mosaic images and to allow a viewer to perceive the 3D terrain while changing viewpoints over a large scale. Four critical issues will be discussed in this paper: 1) Under what kinds of motion can we construct a pair of 2D stereo mosaics before we recover any 3D information? 2) How to make stereo mosaics seamless in the presence of motion parallax and for rather arbitrary scenes? 3) What is the epipolar geometry of multi-perspective stereo mosaics generated under rather general motion? 4) What are the benefits of generating stereo mosaics in computation, storage, 3D resolution and 3D visualization?

This paper is organized as follows. First we will analyze two possible geometric representations for stereo mosaics - multiple-perspective mosaics and parallel-perspective mosaics. We will show that the parallel-perspective mosaic representation is superior to a multiple-perspective representation for the geometric seamlessness of its visual appearance and the simplicity of its epipolar geometry, which will both interactive benefit human viewing and automated stereo matching. Then we will give the condition when such stereo mosaics can be created. Next, a novel method based on piecewise interframe matches and parallel-perspective view synthesis is proposed to generate dense and seamless parallel-perspective stereo mosaics. The advantages of stereo mosaics in representation, computation, 3D resolution and 3D visualization are discussed throughout the paper. Experimental results on long video sequences will be given to validate our analysis.

2. The Geometric Model of Stereoscopic Mosaics

A suitable representation is the first important element in generating stereo mosaics from a moving camera that allows viewpoint changes over a large scale. In this section the geometric model of the stereo mosaics will be presented in a somewhat ideal case, i.e., one-dimensional (1D) translation, however, the main concepts and features of multi-perspective stereo mosaics are covered. Here is the basic idea of the stereo mosaics: if the motion of the camera is a 1D translation of constant speed, the optical axis is perpendicular to the motion, and the frames are dense enough, then we can generate two spatio-temporal images by extracting a column of pixels near each of the rear and the front borders of the image that are perpendicular to the motion (Fig. 1a). These mosaic images are similar to *parallel-perspective* images captured by a linear pushbroom camera [24] , 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 imaging, these mosaics are obtained from two different constant 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 can be generated. In the following we will present this concept of stereo mosaics of a translating camera, and then compare two geometric representations of dense stereo mosaics: multiple-perspective projection and parallel -perspective projection.

2.1. Stereo mosaic geometry

Before constructing the stereo mosaics, let us first consider the stereo geometry of such two-slit windows in a perspective image (Fig. 1). Similar treatment has been proposed in Zheng & Tsuji [10] for landmark selection for a mobile robot. Here we extend it to 3D mosaicing of multi-perspective and parallel-perspective projections in the presence of motion parallax. Without loss of generality, we assume that two vertical slit windows have $d_y/2$ offsets to the left and right of the center of the image respectively. Now we will show why viewing through these two slits captures stereo information. Suppose that a curve C_l in the 3D scene can be seen through the front slit window of the image from viewpoint O_l of the "left eye", and p_l is the image of a point P on the curve. When the camera moves a certain distance B_y in the Y direction to viewpoint O_r of the "right eye", the point P can be seen from the rear slit window of the image from that viewpoint as image p_r, and on a 3D curve C_r. In this ideal motion model, the depth of the point P can be calculated as (Fig. 1b)

$$Z = D = F\frac{B_y}{d_y} \tag{1}$$

where D is the distance (in mm) of the point P from the moving camera, F (in mm or pixel) is the focal length of the camera, d_y (in mm or pixel) is the "disparity", which is a fixed distance between two slit window, and B_y (in mm) is the varying "baseline" (which is the y displacement between two locations of the camera where the same point P can be seen from two slit windows respectively).

Now we will connect the two-slit-window geometry to the stereo mosaic images. The "left eye" view (left mosaic) is generated from the front slit window, while the "right eye" view (right mosaic) is generated from the rear slit window. By definition, both of the digital stereo mosaics are represented as xy images whose two dimensions are in pixels. It is convenient that both F and d_y are also measured in pixels, and the origins of both the left and right mosaics are the same as the origin of a common reference frame (e.g., the first frame, see Fig. 1b). Hence the *parallel-perspective projection model* of the stereo mosaics is represented by the following equations

$$\begin{aligned} x_l &= x_r = F\,X/Z \\ y_r &= FY/H + (Z/H-1)\,d_y/2 \\ y_l &= FY/H - (Z/H-1)\,d_y/2 \end{aligned} \tag{2}$$

where the last two equations are derived from the slit-window perspective geometry as

$$y_r = FT_{yr}/H - d_y/2, \quad -d_y/2 = F(Y-T_{yr})/Z \tag{2a}$$

$$y_l = FT_{yl}/H + d_y/2, \quad d_y/2 = F(Y-T_{yl})/Z \tag{2b}$$

In the above equations, (X,Y,Z) is the 3D coordinates of the space point P in the reference camera coordinate system, H is the average height of the terrain, and T_{yl} and T_{yr} are the Y translational components of the left and right viewpoints of the moving camera when point P can be seen through the rear and front slits (so $B_y = T_{yr} - T_{yl}$). Note that Eq.(2) gives the relation between a pair of 2D points in mosaics, $p_l(x_l,y_l)$ and $p_r(x_r,y_r)$, and the corresponding 3D point $P(X,Y,Z)$, without using the assumption of the ideal translation with constant speed. It serves a function similar to the classical pin-hole perspective camera model. However, these equations do not imply that the construction of stereo mosaics from a video sequence require any 3D information of the scene (e.g., X,Y,Z). On the contrary, in our ideal case the two slit images in each viewpoint of the camera are directly copied from the original frames to the two mosaics. Note that in the *parallel-perspective projection model* (Eq. (2)), *the x coordinate obeys perspective projection geometry, but the y coordinate obeys parallel projection geometry* (Fig. 1c). Using this representation that has varying viewpoints along the y direction, we can easily generate a pair of stereo mosaics from a perspective video sequence when the camera translates along the y direction.

2.2. Dense multi-perspective stereo mosaics

In practice, image frames are captured only in discrete viewpoints (denoted by translational component T_y). In the ideal case, T_y=V k, where V (meters/frame) is the speed of the camera, k is the frame number (k=0,1,…). Eqs. (2a) and (2b) tell us how to generate a dense left mosaic or right mosaic in the ideal case: *We should take a slice whose width is FV/H pixels per frame for each of the mosaics, and put them in the corresponding mosaics centered at y_l and y_r, which are functions of the camera displacements as well as the fixed offset of the two slits* (Eqs. (2a),(2b)). The purpose of the scaling (*F/H*) in these equations is to balance the aspect ratio of each mosaic in x and y dimensions, so the value of H is not so critical. In fact, the mosaic can be constructed based on any plane with some distance H from the camera. Note that a mosaic generated in such a manner is a *multi-perspective* image, which means that *the sub-image (with multiple columns) obtained from a slit is of full perspective, but successive sub-images have different viewpoints* (Fig. 2a, Fig. 2c).

In a multi-perspective mosaic pair, for any point (x_l,y_l) in the left view mosaic, we can find its corresponding point (x_r,y_r) in the right mosaic (unless it is occluded). In general, this pair may not come from the centers of the slit windows in their original frames (Fig. 2c). In order to estimate the depth of their corresponding 3D point (X,Y,Z), we need to know which frames these two points come from. Let us assume that point (x_l,y_l) comes from frame (i.e., time) t_l of viewpoint T_{yl} , and that point (x_r,y_r) comes from frame t_r of viewpoint T_{yr}. Recall that the coordinate of the center of the front (or rear) slit windows in the right (or left) mosaic is given by Eq. (2b) (or Eq. (2a)). Then we can compute the y coordinates of these two mosaicing image points in their own frame coordinate systems as

$$y_{tr} = y_r - (FT_{yr}/H - d_y/2) - d_y/2 \qquad (3a)$$

$$y_{tl} = y_l - (FT_{yl}/H + d_y/2) + d_y/2 \qquad (3b)$$

Note that y_{tr} is always negative and y_{tl} is always positive by definitions of left and right mosaics (see Fig. 2c). The baseline between these two frames is $B_y = T_{yr} - T_{yl}$, and the disparity of the correspondence points is $y_d = y_{tl} - y_{tr} = F(T_{yr} - T_{yl})/H - (y_r - y_l)$. Let us define the stereo "*mosaic displacement*" as

$$\Delta y = y_r - y_l \qquad (4)$$

and the "*scaled baseline*" as

$$b_y = F(T_{yr} - T_{yl})/H \qquad (5)$$

Then the depth of the 3D point can be computed by

$$Z = F\frac{B_y}{y_d} = H\frac{b_y}{b_y - \Delta y} \qquad (6)$$

This is exactly the same disparity equation as in two-view perspective stereo, but there is a difference in the way it works. In the multi-perspective stereo mosaics, it is an adaptive-baseline stereo system: both the disparity (y_d) and the baseline (B_y) vary according to the depth of a point. The disparity is *almost* fixed by the distance of the rear-front slit windows except some variation depending on the width of the sub-images the points lie in; instead two suitable frames with an adaptive baseline is selected to calculate the depth of the point pair. By finding a correspondence pair in multi-perspective stereo mosaics, the baseline and the disparity (or equivalently the scaled baseline b_y and the mosaic displacement Δy) are determined by the correspondence pair using Eqs. (4) and (5). Then the depth can be computed using Eq. (6).

2.3. Dense parallel-perspective stereo mosaics

If the original image sequence is dense enough to guarantee that only one column need to be extracted from each frame for each of the left and right mosaic, then a multi-perspective mosaic becomes its special case, i.e., a parallel-perspective mosaic. Otherwise, for an arbitrary scene, a multi-perspective mosaic has two problems due to motion parallax under translational motion of large interframe displacements. First, the geometry is somewhat complex since in the y direction, the projection model is a mixture of perspective and parallel projections. So for stereo matching using epipolar geometry and for computing 3D information, we need to remember which frame each sub-image in the left or right mosaic comes from and the corresponding viewpoint information. Second, the mosaic will have geometric seam since the border of two successive sub-images from two perspective views may not be able to be aligned completely due to motion parallax. Some points will have double images (ghosts) while some will not appear in the mosaics. This will have a negative effect on both human viewing and again on image match in stereo mosaics. Parallel-perspective stereo mosaics ideally solve these problems since in each of the mosaics, rays are parallel in the y direction (Fig. 2b). In Section 3, we will propose a 3D mosaicing method to generate seamless parallel-perspective mosaics. Here we will analyze the basic parallel-perspective stereo geometry, and show its advantages.

Let us denote a correspondence pair of the point P in a pair of parallel-perspective stereo mosaics as (x_l,y_l) and (x_r,y_r). From Eqs. (2a) and (2b), and using the same notation of mosaic displacements and scaled baseline in Eqs. (4) and (5), we can compute the scaled "baseline" directly as

$$b_y = d_y + \Delta y \tag{7}$$

where $\Delta y = y_r - y_l$ is the mosaic displacement in the stereo mosaics. In the ideal case, the epipolar line is the scanline in the y direction, i.e. $x_r = x_l$. The real "baseline" can be calculated as

$$B_y = b_y H / F \tag{8}$$

Substituting Eq. (8) into Eq. (1) we have

$$Z = H\frac{b_y}{d_y} = H(1+\frac{\Delta y}{d_y}) \tag{9}$$

This equation can also be derived directly from Eq. (2) (as $\Delta y = y_r - y_l = (Z/H-1)\, d_y$). In Eq. (9), both b_y and d_y are measured in pixels, and H and Z are in the same units, for example, in meters. Displacement Δy represents the depth variation around the fixation plane H. It is interesting to note that, since the selection of the two mosaic coordinate systems brings a constant shift d_y to the scaled "baseline", it produces the fixation of the stereo mosaics to a horizontal "*fixation plane*" of an average height H. This is highly desirable for both stereo matching and stereoscopic viewing.

We can establish a relation between a parallel-perspective model (Eq. (9)) and a multi-perspective model (Eq. (6)). If the correspondence points in a multi-perspective stereo mosaic pair locate right in the center of rear and front slit images, i.e. $y_r = (FT_{yr} /H - d_y/2)$ and $y_l = (FT_{yl} /H + d_y/2)$ (see Eqs. (3a) and (3b)), then Eq. (6) will become Eq. (9). In this sense, parallel-perspective stereo is a special case of multi-perspective stereo. It is a desirable special case since (1) parallel-perspective stereo mosaic are geometric seamless, and (2) the "disparities" (d_y) of all points are *completely* fixed as the results of a geometry of optimal/adaptive baselines (b_y) for all the points. In other words, for any point in the left mosaic, searching for the match point in the right mosaic means (but does not need) to find an original frame in which this match pair has a pre-defined disparity (by the distance of the two slit windows) and hence has an adaptive baseline depending on the depth of the point.

After we have established the point correspondence, the (X,Y) coordinates of any point P can be determined by using Eq. (2), as

$$X = \frac{H}{F}(1 + \frac{\Delta y}{d_y})x_l, \quad Y = \frac{H}{F}\frac{y_r + y_l}{2} \qquad (10)$$

Note that translational component T_y is not included in the 3D calculation. It means that once parallel-perspective stereo mosaics are generated, we no longer need to consider the real camera motion parameters. Only the ratio of the reference height H versus focal length is required. This analysis also implies that it is straightforward to represent the 3D information in a digital elevation map (DEM) as long as a stereo match is performed in the stereo mosaics for each elevation.

In comparison, for high quality mosaics and accurate 3D reconstruction, parallel-perspective mosaics are preferable; for computational efficiency, the general multi-perspective mosaics that can be generated via a simple 2D mosaicing technique still have the similar basic properties as the parallel-perspective mosaics.

2.4. Stereo-mosaic properties

Multi-perspective and parallel-perspective stereo mosaics are very useful for both 3D reconstruction and image-based rendering of large-scale terrain or urban scenes. Fig. 3 - Fig. 5 show an example of environmental monitoring of forest where stereo mosaics are quite valuable.

Fig. 3 shows three frames from a 165-frame video sequence, collected as part of a project with the Nature Conservancy (TNC) for determining biomass in preservation of tropical forest in Bolivia. The image resolution is 640(V)*480(H). Fig. 3a and Fig. 3b are two successive images (frames 20 and 21) with about 18-pixel interframe displacement. The two slit windows are 224-pixel apart. The sight difference in these two frames can be observed by focusing on the rear (or front) slit windows. Fig. 3c is the image (frame 33) where scene in the front-slit window of Fig. 3b can be seen in its rear-slit window. These three images also show the dilemma of baseline versus common FOV in a classic stereo/motion approach for each pair of images in a video sequence: A larger common FOV means a shorter baseline (Fig. 3a, Fig. 3b), but a longer baseline means a smaller common FOV(Fig. 3b, Fig. 3c). Fixated optical axes of binocular cameras can improve the situation, but fixation of optical axes brings in additional problems in data collection since such fixation needs 3D depths of scene points and probably more than one cameras.

Stereo mosaic methods solve this problem by extending the FOV in the baseline's direction. 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, and optimal/adaptive baselines for all the points. Fig. 4 shows the right mosaic and the left mosaic generated from the 165-frame video sequence. It should be noted that the motion of this sequence is not a 1D translation. These stereo mosaics are generated using the methods that will be discussed in the following sections. The reason for us to move the results here is to give readers a real example of stereo mosaics before we go into more technical details. Fig. 5a shows a window of the red-blue overlaid stereo pair of the stereo mosaics, where the red channel is the right mosaic and the blue channel is the left mosaic. By viewing through a right-red/left-blue glasses, dramatic 3D perception can be achieved. Experiences of both forestry experts and laymen show that the stereoscopically viewed mosaics of trees are both compelling and vivid to the viewers. It is especially true when a polarized stereo viewing system are used. The 3D reconstruction obtained from the stereo mosaics is very encouraging. Fig. 5b shows a 3D rendering result using the texture map (Fig. 4a) and the depth map (Fig. 4c) from a sub-pixel image match [25]. More high resolution stereo mosaics of both forest scenes and urban scenes can be found at our web site [27].

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

1) *Compact visual representation:* Stereo mosaics are a compact representation for long video sequences. Thousands of video frames can be represented in two mosaics that preserve 3D information of a static scene. The size of a mosaic is only a function of terrain coverage and image resolution. So it is worthwhile to generate stereo mosaics even under more complicated motions if the generation of stereo mosaics is cost effective (as discussed in Sections 3).

2) *Adaptive baseline and optimal FOV:* The disparity (d_y) is fixed for any point in a scene in these stereo mosaics. On the other hand, the baseline (B_y) in real world coordinates, or the scaled "baseline" (b_y) in the stereo mosaics, is proportional to the depth of a scene point. The "baseline adaptation" of the multi-perspective stereo can make full use of an image sequence by setting the "disparity" d_y as a fixed number (and as large as possible), and the depth resolution will be independent of the depth itself (A numerical analysis will be given in Section 4). In addition, multiple-perspective stereo mosaics approach is an ideal solution to the conflict in tradeoff between the magnitude of the baseline versus common field of view in traditional stereo vision.

3) *Efficient Computation*. Multi-perspective stereo mosaics greatly reduce the computational burden on 3D reconstruction by avoiding the processing of large amounts of data over multiple frames. As will shown in the following sections, 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.

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, an epipolar curve geometry can be nicely derived so that only a 1D search is needed for stereo correspondence matching (Section 4).

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 (Section 5).

3. Image Capture and Mosaic Generation

In real applications, such as environmental monitoring from an aerial camera, or urban scene modeling from a camera mounted on a vehicle, it is unlikely to have ideal 1D translational motion with image frames being evenly and densely captured. Now the question is: can we generate stereo mosaics with a more general and practical motion model? In this section we will show under what kinds of motion stereo mosaics can be generated. Then, assuming that we know the motion of the camera, we will focus on how to generate seamless mosaics under motion parallax. The challenging problem of estimating 3D motion is briefly discussed in Appendix 1 and we will leave it as a future research topic.

3.1. Camera model

First, the camera motion is modeled as general motion of six degree-of-freedom (DOF) rotation and translation relative to the reference camera system. Here the 3D rotation is represented by a rotation matrix **R**, and the 3D translation is denoted by a vector $\boldsymbol{T} = (T_x,T_y,T_z)^t$. A 3D point $\mathbf{X_k} = (X_k,Y_k,Z_k)^T$ with image coordinates $\mathbf{u} = (u_k, v_k, 1)^t$ at current frame *k* can be related to its reference coordinates $\mathbf{X} = (X, Y, Z)^T$ by the following equation

$$\mathbf{X} = \mathbf{R}\mathbf{X}_k + \mathbf{T} \qquad (11)$$

We find that the condition under which we can generate geometric-seamless stereo mosaics is when the motion track has a dominant motion direction (Fig. 6). Without loss of generality, assuming that vehicle's motion is primarily along the *y* axis, then we will have $T_x << T_y$, $T_z << T_y$. If these conditions cannot be satisfied, we can satisfy the requirement by applying an image rotation transformation as long as the motion has a dominant direction. We also assume that the motion in the optical axis direction is small compared to the average distance (height) of the camera to the scene , i.e. $T_z << H$. Hence Eq. (11) can be approximated quite well by

$$\begin{pmatrix} X \\ Y \\ Z \end{pmatrix} - \begin{pmatrix} T_x \\ T_y \\ 0 \end{pmatrix} = \mathbf{Q}\begin{pmatrix} X_k \\ Y_k \\ Z_k \end{pmatrix}, \ \mathbf{Q} = \mathbf{R} + \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & 0 \\ 0 & 0 & T_z/H \end{pmatrix} \qquad (12)$$

where the average height *H* can be estimated from the range profile information accompanying the image sequence. Note that the only approximation in Eq. (12) is the replacement of T_zZ/H with T_z. *Strictly speaking, the general case enabling stereo mosaicing is a motion constrained to 3D rotation and 2D translation (i.e.,* $T_z = 0$*). However, in practice, small* T_z *can be accounted for as a* *scaling factor* (as in Eq. (12)). Under this rather general camera model, we propose a two-stage algorithm to construct seamless, geo-corrected and 3D preserved stereo mosaics: image rectification followed by image mosaicing. A real-time 3D mosaicing method based on view interpolation and image morphing will be proposed in order to achieve both *seamless stitching* and *parallel-perspective geometry* in each mosaicing image.

3.2. Image rectification and mosacing basics

In the image rectification stage, a projective transformation **A** is applied to each frame of the video using the motion parameters from geo-data and global image registration[1]. The resulting video sequence will be a rectified image sequence as if it is captured by a "virtual" camera undergoing 2D translations (T_x,T_y) with the dominant T_y component (Fig. 6). It implies that the mosaic will be produced along the global flight direction. The projective transformation is expressed as

$$\mathbf{u}_k^{\mathbf{p}} \cong \mathbf{A}\mathbf{u}_k \quad \boldsymbol{A} = \boldsymbol{F}\boldsymbol{Q}\boldsymbol{F}^{-1}, \quad \boldsymbol{F} = \begin{pmatrix} F & 0 & 0 \\ 0 & F & 0 \\ 0 & 0 & 1 \end{pmatrix} \tag{13}$$

where $\mathbf{u}_k^{\mathbf{p}}$ is the reprojection image point of frame k, and equation "$\cong$" holds in the sense of projective transformation (for an expanded form, refer to Eqs (a-1) and (a-2)). From now on, for convenience, we will use $\mathbf{u}_k$ instead of $\mathbf{u}_k^{\mathbf{p}}$ as the rectified image coordinates unless indicated.

After image rectification, there is no rotation between frames, the nodal point of the camera will be in a single plane, and the Y direction of each frame will be the same (Fig. 6b). So in the initial step of image mosaicing stage we can apply a *direct cut-and-paste step*, which is a generalization of the multi-perspective mosaicing in Section 2.2: we take two x-slices (a set of columns in Fig. 7) from each frame, centered at the two slit windows. These x-slices will be called "mosaicing slices". Due to varying displacements between each pair of successive frames in the image sequence, the width of each of the left and right mosaicing slices in the current frame is computed from the displacements between the previous and the next frames. As an example, we give the equations of the left-view mosaic (Fig. 7). The center of the left-view slice in frame *k* is always at (*x*,*y*) = (0, d_y/2), and the widths of the slice from its center to the left and right borders are

[1] The solution to the challenging problem of camera motion estimation is enabled by a sophisticated aerial instrumentation package that augments the video data with 3D motion of the camera and range profile of the scene. If the data is not accurate, a refinement of the motion parameters is needed. A brief summary of the motion refinement algorithm is given in Appendix 1.

$$w_1^{(k)} = F\frac{|T_y^{(k)} - T_y^{(k-1)}|}{2H}, \quad w_2^{(k)} = F\frac{|T_y^{(k+1)} - T_y^{(k)}|}{2H} \tag{14}$$

where $T_y^{(k-1)}$, $T_y^{(k)}$ and $T_y^{(k+1)}$ are the Y translational components (relative to the reference frame) of frame *k-1, k* and *k+1* respectively. The slice will be put in the left-view mosaic centered at point

$$(t_x^{(k)},\ t_y^{(k)} + \frac{d_y}{2}) = (F\frac{T_x^{(k)}}{H},\ F\frac{T_y^{(k)}}{H} + \frac{d_y}{2}) \tag{15}$$

The central column of the ($w^{(k)}_1+w^{(k)}_2$)-wide mosaicing slice is called the *fixed line,* which is fixed by the camera's motion, and the two straight borders of the slice are called *stitching lines,* where the current slice will be stitched to the previous and the next slices. This approach is similar to the "manifold projection" method [11], where each image contributes a slice to the mosaic. For a translating camera, manifold mosaic can be modeled as a *multi-perspective image*: Each sub-image (with more than one column) taken from the original image is of full perspective, but sub-images from different frames will have different viewpoints (Fig. 2a). This often cause geometric seams in the mosaic due to motion parallax under translation. An error analysis of a direct cut-and-paste method is given in Appendix 2.

3.3. 3D mosaicing by view interpolation

Before we describe our 3D mosaicing algorithm, we first give definitions of seamless 2D and 3D mosaics. A *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 DOF motion. In these two cases, a homogenous image transformation can be used between two frames. In this paper, we define *3D mosaicing* as the generation of a stereoscopic mosaic pair, both mosaics are parallel-perspective composite images of an image sequence of an arbitrary scene under 3D rotation and 2D translation (where motion parallax between frames is apparent). This definition includes two aspects: 1) A 3D mosaicing 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. Obviously, using a 2D mosaicing technique such as manifold projection [11] to form a multi-perspective mosaic may produce a seam at the location of a stitching line due to large motion parallax, so 3D mosaicing techniques are needed. One existing approach is the "3D mosaic + parallax" approach [19]. In this approach, in order to construct a 3D corrected mosaic, at least three views of the scene are needed, which should partially overlap with each other. A synthetic image from a reference *perspective view* can be generated from any other pair 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 needs to be estimated. Here we will show that more efficient algorithms can be developed to generate seamless mosaics by selecting a more suitable geometric representation other than a perspective mosaic, namely parallel-perspective mosaic.

3.3.1. View interpolation by ray re-projection

How can we generate seamless mosaic in a computational effective way? The key to our approach lies in the parallel-perspective representation and an interframe view interpolation approach. For each of the left and right mosaics, we only need to take a front (or rear) slice of a certain width (determined by interframe motion) from each frame, and perform local registration between the overlapping slices of successive frames, then generate parallel *interpolated views* between two known discrete perspective views for the left (or right) mosaic. Our approach is similar to view synthesis by image interpolation, which has been well-studied in image-based rendering [26]. Fortunately in our case, the distance between two successive views are small, thus the synthetic views of parallel-perspective projection between the two known views reflect the 3D structure of the scene, and yet neither need explicit 3D estimation nor have serious occlusion problem. In addition, since we only need to stitch two narrow slices from two successive images, the point matching 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 given a 3D translation after image rectification.

Let us examine this idea more rigorously in the case of 2D translation after image rectification. We have defined the central column of the front (or rear) mosaicing slice in each frame as a *fixed line*, which has been determined by the camera's location of each frame and the pre-selection of the front (or rear) slice window (Fig. 7). In Fig. 8a, an interpretation plane (IP) of the fixed line is a plane passing through the nodal point and the fixed line. By the definition of "parallel"-perspective stereo mosaics, the IPs of fixed lines for the left (or right) mosaic are parallel to each other. We take the left mosaic as an example. Suppose that (S_x, S_y) is the translational vector of the camera between the previous (1st) frame of viewpoint (T_x, T_y) and the current (2nd) frame of view point (T_x+S_x, T_y+S_y). We need to interpolate views between the *fixed lines* of the 1st and the 2nd frames. For each point (x_l, y_1)(to the right of the fixed line y_0=d_y/2) in frame(T_x, T_y), which will contribute to the left mosaic, we can find a corresponding point (x_2, y_2) (to the left of the fixed line) in frame (T_x+S_x, T_y+S_y). We assume that (x_1, y_1) and (x_2, y_2) are represented in their own frame coordinate systems, and intersect at a 3D point (X,Y,Z). Then the parallel reprojected viewpoint (T_x, T_{yi}) of the correspondence pair can be computed as

$$T_{yi} = T_y + \frac{(y_1 - d_y/2)}{y_1 - y_2} S_y, \; T_{xi} = T_x + \frac{S_x}{S_y}(T_{yi} - t_y) \qquad (16)$$

where T_{yi} is calculated in a synthetic IP that passes through the point (X,Y,Z) and is parallel to the IPs of the fixed lines of the first and second frames , and T_{xi} is calculated in a way that all the viewpoints between (T_x,T_y) and (T_x+S_x, T_y+S_y) lie in a straight line. In the case of left mosaic and forwarding motion (Fig. 8a), we always have $y_1 \geq d_y/2$ and $y_2 <= d_y/2$. For points in the first frame starting from the point with $y_1 = d_y/2$ to the point that matches the point with $y_2 = d_y/2$ in the second frame, the "parallel" viewpoints interpolated must lie between (T_x, T_y) and (T_x+S_x, T_y+S_y). The viewpoint of the point pair is determined by this pair of points and hence by its 3D information[2]. The mosaicing coordinates of this pair are

$$x_i = t_{xi} + x_1 - \frac{S_x}{S_y}(y_1 - \frac{d_y}{2}), \; y_i = t_{yi} + \frac{d_y}{2} \qquad (17)$$

where

$$t_{xi}=F\, T_{xi}/H\,, \;\; t_{yi}=F\, T_{yi}/H. \qquad (18)$$

are the "scaled" translational components of the interpolated view.

Note that the interpolated views are also parallel-perspective - perspective in the x direction and parallel in the y direction. The reason for us not to use a full-parallel projection model is: while the changes between an interpolated view and the given views in the y direction is small, the changes between an interpolated parallel view and the given views in the x direction is large if we also want to generate a parallel projection in the x direction (Fig. 8b).

[2] Note that we may not be able to find correspondences for some points in the two views due to occlusion, thus some intermediate views cannot be generated in this way. The 3D mosaicing algorithm proposed in the next sub-section will deal with this problem.

3.3.2. A 3D mosaicing algorithm

We have designed a *3D mosaicing* algorithm based on the proposed local registration and view interpolation method. It should be noted that in the view interpolation, the transformations of the fixed lines for both left and right mosaics are determined only by the camera motion parameters. Therefore they are identical for the rear and front slit windows in the same frame, which enable the correctness of the parallel-perspective geometry that will be discussed in Section 4. More specifically, the image rectification $\mathbf{A_k}$ expressed in Eq. (13) is carried out for the entire frame (or the two slices for the sake of computational efficiency); and the left and right fixed lines are fixed into the suitable locations of the corresponding mosaics by the camera translational parameters ($t_x^{(k)}$,$t_y^{(k)}$). Then, in order to make seamless mosaics for both the left and right views, different local transformations are used in accordance with the different motion parallax from two oblique viewing angles in the same frame.

The proposed 3D mosaicing algorithm only requires matches between a set of point pairs in two successive images around their *stitching line,* which is defined as a virtual line in the middle of the two fixed lines (Fig. 9a). Note that this stitching line is where a *2D mosaic* method is supposed to smoothly interface the two successive slices (Fig. 7). The 3D mosaicing algorithm consists of the following four steps:

Step 1. *Slice determination* - Determine the fixed lines in the current frame *k* and the previous frame *k-1* by their 2D scaled translational parameters $(t_x^{(k)}, t_y^{(k)})$ and $(t_x^{(k-1)}, t_y^{(k-1)})$ (as in Eq. (18)). Then an "ideal" straight stitching line lies in the middle of the two fixed lines. Thus we have two overlapping slices, each of which starts from the fixed line and ends at a small distance away from the stitching line to ensure overlapping (Fig. 9a).

Step 2. *Match and view interpolation* - Match a set of corresponding points as control point pairs in the two successive overlapping slices, $\{(P_{1i}, P_{2i}), i = 1,2,\ldots N\}$, in a given small region along epipolar lines, around the straight stitching line. We use a correlation-based method to find a pair of *matching curves* passing through the control points in the two frames. The control point pairs are determined by measuring both the gradient magnitudes and the correlation values of a small window centered at the control point. Then compute the destination location Q_i ($i=1,\ldots,N$) of the interpolated view in the mosaic of each corresponding pair (P_{1i}, P_{2i}) using Eq. (17). A curve that passes through the point list $\{Q_i$ ($i=1,\ldots,N$)$\}$ is defined as the *stitching curve* where the two slices will be stitched after image warping (Fig. 9a). Both the matching pairs and the destination points form *curves* instead of straight lines due to the depth variation of the control points.

Step 3. *Triangulation* - Select two sets of control points R_{mi} $(m=1,2; i=1,\ldots N-1)$ on the fixed lines in the two frames, whose x coordinates are determined by the fixed lines and whose y coordinates are the averages of P_{mi} and $P_{m,i+1}$ (m=1,2) for good triangulation. Map R_{1i} and R_{2i} into the mosaic coordinates as S_{1i} and S_{2i} $(i=1,\ldots N)$, by solely using the interframe translations $(t_x^{(k)}, t_y^{(k)})$ and $(t_x^{(k-1)}, t_y^{(k-1)})$. For *kth* frame, we generate two sets of corresponding triangles (Fig. 9a): the source triangles by point sets $\{P_{2i}\}$ and $\{R_{2i}\}$, and the destination triangles by point sets $\{Q_i\}$ and $\{S_{2i}\}$. Do the same triangulation for the *(k-1)st* frame.

Step 4. *Warping* - For each of the two frames, warp each source triangle into the corresponding destination triangle, under the assumption that the region within each triangle is a planar surface given small interframe displacements. Since the two sets of destination triangles in the mosaic have the same control points on the stitching curve, the two slices will be naturally stitched in the mosaic.

Fig. 9b shows how this algorithm can be improved to work in the presence of occlusion. First, a pair of correspondence points that can be seen in both images is selected. Then those points that can only be seen from the first image are warped from the first image to the mosaic.

3.3.3. Experimental analysis and discussions

Fig. 10 to Fig. 12 show a real example of 3D mosaicing by local match and view interpolation for a Umass campus scene. Fig. 10 shows the local match of two successive frames where the interframe motion is $(s_x, s_y) = (3, 36)$ pixels, and points on the top of a narrow building have 1-2 pixel additional motion parallax. Fig. 11 shows the local match and view interpolation of another successive frame pair where the interframe motion is $(s_x, s_y) = (27, 48)$ pixels, and points on the top of a tall building have about 4 pixel additional motion parallax. As we will see next, the 1-2 pixel geometric misalignments, especially of linear structures, will be clearly visible to human eyes, and 4 pixel misalignments will definitely creates obvious geometric seam. Moreover, perspective distortion causing the seams will introduce errors in 3D reconstruction using the parallel-perspective geometry of stereo mosaics. In this example of stereo mosaics, the distance between the front and the rear slice windows is $d_y = 192$ pixels, and the average height of the aerial camera from the ground is $H = 300$ m. As an example, we will show the numerical analysis for images in Fig.

10. The relative y displacement of the building roof (to the ground) in the stereo mosaics is about Δy = -12 pixels. Using Eq. (9) we can compute that the "absolute" depth of the roof from the camera is Z = 281.25 m, and the "relative" height of the roof to the ground is ΔZ = 18.75 m. A 1-pixel misalignment will introduce a depth (height) error of δZ = 1.56 m, though stereo mosaics have extremely large "disparity" (d_y =192). While the relative error of the "absolute" depth of the roof ($\delta Z/Z$) is only about 0.55%, the relative error of its "relative" height ($\delta Z/\Delta Z$) is as high as 8.3%. The numerical results for Fig. 10 and Fig. 11 are summarized in Table 1. It can be seen that geometric-seamless mosaicing is very important for accurate 3D estimation.

While we are still working on 3D camera orientation estimation using bundle adjustments when writing this paper, Fig. 12f shows mosaic results where camera orientations were estimated by registering the planar ground surface of the scene via dominant motion analysis. However the effect of seamless mosaicing is clearly shown in this example. Please look along many building boundaries associating with depth changes in the entire 4160x1536 mosaics at our web site [27] for a comparison of the 3D mosaicing via 2D mosaicing. Fig. 12a shows a tile of the multi-perspective mosaic generated using 2D mosaic method from a temporally sub-sampled image sequence (every 10 frames, i.e. the interframe motion is about 40 pixels). Geometric misalignments (seams) at interfaces of successive slices are obvious, especially along building boundaries with large depth changes. Fig. 12b shows a tile of the parallel-perspective mosaic of the same temporal sub-sampled image sequence as in Fig. 12a but this time the proposed 3D mosaicing method is used. Most of the geometric seams visible in Fig. 12a are eliminated in Fig. 12b. Fig. 12c shows a tile of a multi-perspective mosaic when all the frames are used (i.e. the interframe motion is less than 4 pixels). In this case the multi-perspective mosaic is very close to a parallel-perspective mosaic; however, there are still "seams" in some places, e.g. areas indicated by a rectangle. It can be seen that the sparse-sampled parallel-perspective mosaic is better than the dense-sampled multi-perspective mosaic since local matches along stitching lines eliminate misalignments between two successive slices. These misalignments may be introduced by 3D structure of the scene, errors in motion modeling and errors in camera motion estimation. The 3D mosaicing algorithm can also result in a great saving of space and time since the algorithm will work on a highly sub-sampled sequence. Fig. 12d and Fig. 12e show a comparison of the 2D and 3D mosaics for the part of the scene with a tall building. In Fig. 12d 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 are missing due to misalignments. In Fig. 12e all these problems are fixed by using the parallel-perspective mosaicing.

In principle, we need to match all the points between the two fixed lines of the successive frames to generate a complete parallel-perspective mosaic. In an effort to reduce the computational complexity, the above algorithm only matches points on a "stitching curve", and the rest of the points are generated by image warping, assuming that each triangle is small enough to be treated as a planar region. Using sparse control points and image warping, the proposed 3D mosaicing algorithm only approximates the parallel-perspective geometry in stereo mosaics. However, the proposed 3D mosaicing algorithm can be easily extended to use more feature points (thus smaller triangles) in the overlapping slices so that each triangle really covers a planar patch or a patch that is visually indistinguishable from a planar patch. Further experiments are underway.

4. Depth from Stereoscopic Mosaics

Parallel-perspective stereo mosaics greatly reduce the computational burden of 3D reconstruction. Correspondence and 3D estimation are only performed on two mosaics, which are constructed using only sparse yet robust image matches. After the image rectification and mosaicing of a long image sequence, simple epipolar geometry can be derived in parallel-perspective mosaics. The disparity is fixed for every point in a scene in stereo mosaics; on the other hand, the "baseline" is proportional to the range of the point, which means that the depth resolution is independent of the depth itself. The parallel-perspective geometry does not impose additional problems for the 3D reconstruction stage since it is an exact geometric model although it is different from the traditional perspective stereo model.

4.1. Epipolar curves

In the case of the general motion model in Section 3, the stereo mosaics are generated from a rectified image sequence of only 2D translation (T_x,T_y), with the component T_y dominant. The epipolar lines are no longer horizontal scanlines as in the ideal case if T_x is not zero for every frame. Let us denote a pair of correspondence points in the left mosaic as (x_l,y_l) and in the right mosaic as (x_r,y_r); they are originated from frame t_l and frame t_r respectively. The displacements in x and y are (Fig. 13a)

$$\begin{aligned}\Delta x &= x_r - x_l \\ \Delta y &= y_r - y_l\end{aligned} \tag{19}$$

Note the difference between these two displacements: Δx is related to d_x , the disparity in the x direction; while Δy is related to b_y, the "scaled" baseline in the y direction, by the following equations

$$\begin{aligned}d_x &= b_x - \Delta x \\ b_y &= d_y + \Delta y\end{aligned} \tag{20}$$

where

$$b_x = F\frac{T_{xr} - T_{xl}}{H} = t_{xr} - t_{xl} \tag{21}$$

and T_{xr} and T_{xl} are the global x translational components of frame t_l and t_r respectively. The introduction of "scaled" baseline b_x and image translation t_x is just for consistency in the x and y directions. The depth of the point can de derived as

$$Z = H\frac{b_y}{d_y} = H\frac{b_x}{d_x} \tag{22}$$

This is the *distance-baseline-disparity equation* for multi-perspective stereo mosaics. Note that if we already know the correspondence, the depth can be calculated from the first part of Eq. (22). However, the second part of this equation will help us build up the correspondence relation. Recalling the generation of stereo mosaics, we can easily find that in Eq. (21) we have

$$t_{xr}(y) = t_{xl}(y + d_y)$$

Here t_{xr} and t_{xl} are expressed as functions of coordinate y in the stereo mosaics. They are also interpolated during the view interpolation of 3D mosaicing. The corresponding point (x_r,y_r) in the right-view mosaic of any point (x_l,y_l) in the left-view mosaic will be constrained to the following *epipolar curve* (From Eqs. (20)-(22))

$$\Delta x = b_x(y_l, \Delta y)\frac{\Delta y}{\Delta y + d_y} \tag{23}$$

where

$$b_x(y_l, \Delta y) = [t_{xl}(y_l + d_y + \Delta y) - t_{xl}(y_l)] \tag{24}$$

is the baseline function of y_l and Δy. Note that Δx is a function of position y_l as well as displacement Δy , which is quite different from the epipolar-line geometry of a two-view perspective stereo. In other words, if a point (x,y) is on the epipolar curve C_e of a given point (x_l,y_l), the epipolar curve of that point (x,y) may not be C_e. The reason is that image columns of different y_l in parallel-perspective mosaics are projected from different viewpoints that are represented by (t_x , t_y). To understand the geometry of the epipolar curve better, we can find some special points in an epipolar curve (Fig. 13a):

(i) $\Delta x=0$ *if* $\Delta y=0$;

(ii) $\Delta x= -b_x(y_l,\Delta y)$ *if* $\Delta y= -d_y/2$ *(i.e., Z=H/2);*

(iii) $\Delta x= b_x(y_l,\Delta y)/2$ *if* $\Delta y= d_y$ *(i.e., Z=2H).*

We have the following important conclusions. From special case (i) we have

Conclusion 1: *The epipolar curve for a given point (x_l,y_l) in the left mosaic is a curve passing through this point. It means that the stereo mosaics are aligned for all the points whose depths are H.*

Special case (ii) and case (iii) show how big the x displacement in the stereo mosaics if the depth is half of or twice the average height H. To see how big Δx is in the general case, let us assume that the depth variation is $1/s_h$ (s_h >1) of the average height H, and the average x translation (i.e. baseline b_x) is $1/s_b$ (s_b >1) of the y displacement d_y. Substituting $\Delta y = d_y/s_h$, $b_x = d_y /s_b$ into Eq. (23), we have

$$\Delta x = \frac{d_y}{s_b(s_h+1)} \approx \frac{\Delta y}{s_b} \tag{25}$$

which means that the x displacement is only $1/s_b$ of the y displacement. In an actual mosaicing experiment of the forest scene in Fig. 4, d_y=224 (pixels), $s_h >= 10$, $s_b >=10$, so we have $\Delta x < d_y$ /100 = 2.24 pixels on average. Here we have another two conclusions:

Conclusion 2: *If we know the depth variation range* $\pm\Delta Z_m$, *the search region for the corresponding point can be determined by* $\Delta y \in [-\frac{d_y}{H}\Delta Z_m, +\frac{d_y}{H}\Delta Z_m]$, *and along a 1D epipolar curve (Eq. (23)).*

Conclusion 3: *If the x translation t_x is zero everywhere, there will be no x displacement in the mosaic pair and the epipolar curves will become horizontal epipolar lines.*

As an example, Fig. 4c shows the derived "depth" map (i.e., displacement map) from the pair of parallel-perspective stereo mosaics of a forest scene in Fig. 4. The depth map is obtained by using a hierarchical sub-pixel dense correlation method [25]. In the depth map, mosaic displacement (*Δy* in

Eq. (9)) is encoded as brightness (brightness is from 0 when Δy = 18.3 pixels, to 255 when Δy = -16.2 pixels). So higher elevation (i.e. closer to the camera) is brighter. The 3D rendering result is shown in Fig. 5b. It should be noted here that the parallel-perspective stereo mosaics were created by the proposed 3D mosaicing algorithm, with the camera orientation parameters estimated by the same dominant motion analysis (Appendix 1) as in Fig. 12. Here, the fixation plane is a "virtual" plane with an average distance ($H=390\ m$) from the scene to the camera. However, promising depth information has been obtained. Fig. 13 shows the statistics of the mosaic displacement and the epipolar geometry of the real stereo mosaics in Fig. 4. The histogram of the displacement map in Fig. 13b shows the numbers of pixels for different mosaic displacements from -16.2 to +18.3 pixels. It can be found that the Δy displacement distribution has almost a zero mean, which indicates that the stereo mosaics indeed fixate to a virtual plane. And most of the pixels have displacements within -10.0 pixels to +10.0 pixels. Using Eq. (9) we can estimate that the range of depth variations of the forest scene (from the fixation plane) is from -17.4 m (of tree canopy) to 17.4 m (of the ground). Fig. 13c shows the x translational component (t_{xl}) of the left mosaic in the common field of view of the stereo mosaics in Fig. 4. Fig. 13d shows the *x*-direction search range [Δx_{min}, Δx_{max}] of a corresponding point in the right mosaic for each point in the left mosaic of every *y* coordinate, given the y-direction search range as [-10, 10] pixels. It can be seen that for the most part of the mosaics, the *x*-direction search range is within ±3 pixels except the right tilted tail (up to ±8 pixels). The more accurate 3D recovery from parallel-perspective mosaics with accurate 3D camera orientation and sub-pixel geometric-seamless mosaicing is underway.

Recall that in the view interpolation of the proposed 3D mosaicing algorithm, we do not move the viewpoints of the parallel rays of the existing frames so that the viewpoints of the stereo mosaics are on the original camera path. Conclusion 3 implies that if we want to generate a pair of parallel-perspective stereo mosaics with horizontal epipolar lines, we need to generate new views for every pixel in the stereo mosaics other than directly using the original camera locations for the "fixed lines". It is possible in principle by using a similar view synthesis procedure as the view interpolation given that the motion in the *x* direction is relatively smaller than in the dominant motion direction y, so that the synthetic view points of the stereo mosaics are not far away from the original camera path. It is one of the interesting issues we will further investigate.

4.2. Depth resolution

In a pair of parallel-perspective stereo mosaics, depth is proportional to the image displacement Δy (Eq.(9)). On the other hand, for a two-view perspective stereo pair with parallel axes, we can derive a disparity equation in a similar way as Eq.(9)

$$Z = H \frac{b_y}{b_y - \Delta y} \tag{26}$$

where $b_y = FB_y/H$ is proportional to baseline B_y in the Y direction. Remember that Δy is measured in discrete images. Without loss of generality, let us assume that the image localization resolution is s pixels (s <=1), so that Δy =0, ± s, ± 2s, …. *The depth resolution in the parallel-perspective stereo of Eq. (9) is a constant value* (Fig. 14a)

$$dZ = \frac{H}{d_y} s = \text{constant} \tag{27}$$

whereas the depth resolution in the two-view perspective stereo of Eq. (26) is a function of Z and Δy

$$dZ = H \frac{b_y}{(b_y - \Delta y)^2} s = \frac{Z}{b_y - \Delta y} s \tag{28}$$

In contrast to multi-perspective stereo, in two-view perspective stereo, the larger the depth of a point, the coarser is the depth resolution (Fig. 14Fig. 13b).

5. Stereoscopic Viewing

Multi-perspective /parallel-perspective stereo mosaics can serve as an image-based representation for large-scale stereoscopic viewing. Without 3D recovery, true 3D perception can be achieved (by a human viewer) while allowing the viewer to change viewpoint across a large spatial scale.

Multi-perspective geometry is not full perspective. However the human observer has no problem perceiving accurate 3D surfaces from these images when viewed stereoscopically. This may be due in part to the fact that human fovea has about 1-degree acceptance angle (the area over which the image is sharp). When viewing a stereo mosaic, the data in sharp focus is almost a true perspective image (Fig. 15). Moreover, there are two additional advantages of the stereo mosaics over a traditional stereoscopic pair: *viewing fixation and moving viewpoints.* As we will explain in the following, these advantages of stereo mosaics enable the viewer to fly over the scene with fixated 3D perception.

5.1. Viewing distance and resolution

For best 3D perception, there are two constraints: the fixation angle constraint and the size-depth-ratio constraint.

Fixation angle constraint - If we can observe the stereo images from the same viewing angles as those when the images are captured (Fig. 15), the viewing scene will have the same occlusion relation as the real situation. Human eyes always focus on the screen when viewing a stereoscopic image pair. Given the fixed baseline B_e of a human observer, the viewing distance H_e determines the fixation angle of the two eyes. Comparing Fig. 1 and Fig. 15, the following relation holds

$$H_e = \frac{F}{d_y} B_e \tag{29}$$

As an example, if $B_e = 80\ mm$, $d_y = 224\ pixels$, $F = 3000\ pixels$, the best distance from eyes to the screen should be $H_e \approx 1m$.

Multi-perspective to single-view perspective approximation - Each of the stereo mosaics is generated by a multi-viewpoint perspective projection model in the y direction (Eq. (2)). Mapping the stereo geometry model in Eq. (9) to the scale of the screen display for human viewing, we have following relation

$$Z_e = H_e (1 + \frac{\Delta Y}{B_e}) \tag{30}$$

where ΔY is the disparity on the screen. On the other hand, human vision works under a perspective projection model

$$Z_e \approx H_e \frac{B_e}{B_e - \Delta Y} \tag{31}$$

which is derived from parallel axis stereo. Since the fixation angle is very small ($H_e/B_e > 10$), the approximation is close to the real situation. The depth error due to the difference in mosaicing and viewing is

$$\Delta Z = Z_m - Z_e = H_e \left(\frac{\Delta Y}{B_e} \right)^2 \tag{32}$$

which is only a second order residual of the depth from the multi-perspective stereo geometry in Eq. (30).

Size-depth ratio constraint - In order that the 3D perception of the scene has the right size-depth ratio (i.e. image dimension versus perceptual depth), a suitable display resolution of the stereo mosaics should be roughly satisfied. Comparing the real scene and the stereoscopic viewing of the same scene, we have

$$H_e \frac{B_e}{B_e - \Delta Y} \propto H \frac{d_y}{d_y - \Delta y} \qquad (33)$$

so the image resolution on the screen should be

$$\frac{\delta Y}{\delta y} = \frac{B_e}{d_y} \qquad (34)$$

where δY/δy (mm/pixel) is the display resolution for the mosaic images. For example, it is 0.36 mm/pixel given the above parameters (i.e., B_e = 80 mm, d_y = 224 pixels).

5.2. Viewpoint change and automatic fixation

The inherent nature of the stereo mosaics allows the viewer to change his viewpoints over a large scale by moving the head or scrolling the images on the screen. The left and right mosaics are pre-aligned based on the assumption that every location has the same average height H. It means the stereo mosaics will not be in fixation if heights changes dramatically across the full spatial extent. Fortunately, in geo-referenced mosaics, we have the range profile from the laser range finder along the flight path. This data give us the flexibility to shift the image in accordance with the changes of the average depths so that an automatic re-fixation can be easily implemented by shifting the mosaics when changing the viewpoint. In the free mosaic case, the stereo mosaics are generated on fixation planes of varying heights so we can always have the fixation of a scene. Fig. 16 shows red-blue copies of a wide FOV window and two zoom windows of the stereo mosaics in Fig. 12. Interested readers can perceive 3D effect using red-blue glasses (in the electronic version of these images at [27]).

Experiments shows that humans have surprisingly good 3D perception with our stereo mosaics, which is also true for free mosaics. Stereo mosaics (before 3D recovery) could be an effective way to provide 3D information for such applications involving realistic human-computer visual interaction. Also it differs in the following three aspects from the usual way of stereoscopic viewing in tele-presence where a pair of video images is displayed in a head-mounted display (HMD). First, only a single camera is needed in stereo mosaics instead of a binocular vision system. Second, stereo mosaics allow a human to change viewpoints to observe motion parallax effect instead of passively accepting the binocular video streams. Finally, stereo mosaics provide a compact way to represent large-scale 3D scenes before 3D reconstruction, which is affordable with a personal computer.

6. Conclusions and Discussions

In this paper, a parallel-perspective mosaic representation and a 3D mosaicing method have been presented for automatically and efficiently generating stereoscopic mosaics by seamless registration of optical data collected with a video camera mounted on a moving platform (e.g., an airborne vehicle or a car). The aim of this work is to generate seamless, continuous and stereoscopic image mosaics that can be used either for stereoscopic viewing during a virtual "fly-through" of the environment, or the 3D recovery of a digital elevation map of a large-scale scene. This paper provides a theoretical base for stereo mosaicing under large translational motion plus 6 DOF interframe rotation and translation. Important issues, such as motion models, large scale mosaic representations, seamless image mosaic techniques, and the epipolar geometry of parallel-perspective mosaics have been thoroughly discussed in the paper. The advantages in 3D reconstruction and visualization using stereo mosaics, including data compression, depth resolution and computational efficiency, have also been analyzed in comparison to traditional perspective stereo vision methods.

Acknowledgements

This work is partially 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). The authors would like to thank Prof. Howard Schultz for his valuable discussion and the helping in generating the depth map using his UMASS Terrest system. Thanks are also given to Kris Denio, Chris Holmes, Chris Hayward and Harpal Bassali for their efforts in data collection, programming and helping in evaluating results.

References

[1]. S. E. Chen, QuickTime VR - an image based approach to virtual environment navigation, *Proc. SIGGRAPH 95*:29-38.

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

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

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

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

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

[7]. H. Ishiguro, M. Yamamoto, and Tsuji, Omni-directional stereo for making global map, *Proc. IEEE ICCV'90*, 540-547.

[8]. S. Peleg, M. Ben-Ezra, Stereo panorama with a single camera, *Proc IEEE CVPR'99*: 395-401

[9]. H. Shum and R, Szeliski, Stereo reconstruction from multiperspective panoramas, *Proc. IEEE ICCV99*, 14-21, 1999.

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

[11]. S. Peleg, J. Herman, Panoramic Mosaics by Manifold Projection. *Proc IEEE CVPR'97*: 338-343.

[12]. P. Rademacher and G. Bishop, Multiple-center-of-projection images, *Proc. SIGGRAPH'98*, 199-206.

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

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

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

[16]. Z. Zhu, A. R. Hanson, H. Schultz, F. Stolle, E. M. Riseman, 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.

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

[18]. H. Shum and R, Szeliski, Construction and refinement of panoramic mosaics with global and local alignment, ICCV'98,: 953-958.

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

[20]. H.S. Sawhney, Simplifying motion and structure analysis using planar parallax and image warping. ICPR'94: 403- 408

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

[22]. S. Jenyns, *A Background to Chinese Paintings*, Schocken Books Inc. 1966

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

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

[25]. H. Schultz. Terrain Reconstruction from Widely Separated Images, In *SPIE*. Orlando, FL, 1995.

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

[27]. Z. Zhu, Parallel-perspective stereo mosaics, http://www.cs.umass.edu/~zhu/StereoMosaic.html.

[28]. A. Gelb, *Applied Optimal Estimation*, MIT Press, Cambridge, MA, 1974.

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

Appendix 1. Camera orientation estimation: free mosaic and geo-mosaic

The geographical data ("geo-data") from our aerial instrumentation package includes 3D locations from the GPS, 3D rotation angles from the Inertial Navigation System (INS), and the range data along the flight path from a laser range finder. 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 [23] for multiple (many!) frames. In the following sub-sections, we will show how we can obtain reasonably good results while greatly simplifying the problem. 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 if 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 [23][27][29].

A.1.1. Free mosaic

To build a free mosaic, we will use the same camera model in Section 3. However, we will assume that two of the rotation angles, tip β (around x axis) and tilt γ (around y axis), are small random variables, whereas the heading angle α around the z axis can be large. In addition, we assume that the depth of the terrain varies only around a virtual horizontal plane, i.e., the fixation plane, which is often true for many video sequences of forest scenes. That is to say, the projective transformation in Eq. (13) can be established by matching image points between the current frame and the reference frame for mosaicing on a virtual fixation plane. This is done by fitting a projective motion model to the motion field between two successive frames. The correspondence matching pairs of points (i.e., tie points in photogrammetry) are selected and weighted using a weighted least mean square approach. In the coordinate system of the reference frame, the virtual plane can be represented by $Z = H$. Consider the projective relation between coordinates (u_k,v_k) in frame k with coordinates (u,v) in the reference frame. Denote **R** = $(r_{ij})_{3x3}$ in Eq. (12). From Eq. (12) we have

$$\begin{aligned} u &= F\,\frac{r_{11}u_k + r_{12}v_k + r_{13}F}{r_{31}u_k + r_{32}v_k + r_{33}F + t_z} + t_x \\ v &= F\,\frac{r_{21}u_k + r_{22}v_k + r_{23}F}{r_{31}u_k + r_{32}v_k + r_{33}F + t_z} + t_y \end{aligned} \qquad \text{(a-1)}$$

where $t_x = F\frac{T_x}{H}$, $t_y = F\frac{T_y}{H}$, $t_z = F\frac{T_z}{H}$, and are treated as a constant parameter for every point in the frame given the fixation plane assumption. From image matches using a pyramid-based correlation algorithm [5][15], we can fit a pseudo-projective transformation to the motion vector field as

$$\mathbf{u} \cong \mathbf{B}\mathbf{u}_k$$

where **B** = $(b_{ij})_{3x3}$, b_{33} =1; or in the expansion form

$$\begin{aligned} u &= \frac{b_{11}u_k + b_{12}v_k + b_{13}}{b_{31}u_k + b_{32}v_k + 1} \\ v &= \frac{b_{21}u_k + b_{22}v_k + b_{23}}{b_{31}u_k + b_{32}v_k + 1} \end{aligned} \qquad \text{(a-2)}$$

Comparing Eq. (a-1) with Eq. (a-2), we know that **B** is the combination of rotation **R** and translation **T**. The processing of decomposing **B** to obtain **R** and **T** are sensitive to noise, which is especially true if the interframe rotations are small. However, if we make use of the small angle assumption for tip β and tilt γ, the rotation matrix can be approximates as

$$\begin{bmatrix} r_{11} & r_{12} & r_{13} \\ r_{21} & r_{22} & r_{23} \\ r_{31} & r_{23} & r_{33} \end{bmatrix} \approx \begin{pmatrix} r_{11} & r_{12} & -\gamma \\ r_{21} & r_{22} & \beta \\ \gamma & -\beta & 1 \end{pmatrix}$$

Then Eq. (a-1) can be rewritten as

$$\begin{aligned} u &= s\,\frac{r_{11}u_k + r_{12}v_k - \gamma F}{\gamma u_k s/F - \beta v_k s/F + 1} + t_x \\ v &= s\,\frac{r_{21}u_k + r_{22}v_k + \beta F}{\gamma u_k s/F - \beta v_k s/F + 1} + t_y \end{aligned} \qquad \text{(a-3)}$$

where $s = \frac{F}{F + t_z}$. We can re-write Eq. (a-2) in the same form as Eq. (a-3)

$$u = \frac{b'_{11}u_k + b'_{12}v_k + b'_{13}}{b_{31}u_k + b_{32}v_k + 1} + t_x$$
$$v = \frac{b'_{21}u_k + b'_{22}v_k + b'_{23}}{b_{31}u_k + b_{32}v_k + 1} + t_y \qquad \text{(a-4)}$$

so that the motion transformation between frame *k* and the reference frame can be decomposed as

$$\mathbf{u} \cong \mathbf{A}_k \mathbf{u}_k + \mathbf{t}_k \qquad \text{(a-5)}$$

By integrating Eq. (a-2) , Eq. (a-3) and Eq. (a-4), we can find that the 2D translational vector in the plane *xoy* is

$$\mathbf{t}_k = (t_x, t_y) = (b_{13} + F^2 b_{31}, b_{23} + F^2 b_{32}) \qquad \text{(a-6)}$$

and the warping matrix is

$$\boldsymbol{A}_k = \begin{pmatrix} b_{11} - b_{31}t_x & b_{12} - b_{32}t_x & -F^2 b_{31} \\ b_{21} - b_{31}t_y & b_{22} - b_{32}t_y & -F^2 b_{32} \\ b_{31} & b_{32} & 1 \end{pmatrix} \qquad \text{(a-7)}$$

In the free-mosaic approach, the rectification stage finds the warping transformation $\mathbf{A_k}$ and the mosaicing parameters $\mathbf{t}_k$ in a *recursive* manner (i.e., the computation of $\mathbf{A_k}$ and $\mathbf{t}_k$ depends on the values of $\mathbf{A_{k-1}}$ and $\mathbf{t}_{k-1}$), frame by frame, since most frames do not overlap with the reference frame for a long video sequence. From Eq. (a-5) we can derive a relation between frame *k* and *k-1* as

$$\mathbf{A}_{k-1}\mathbf{u}_{k-1} \cong \mathbf{A}_k \mathbf{u}_k + \Delta\mathbf{t}_k, \quad \Delta\mathbf{t}_k = (\mathbf{t}_k - \mathbf{t}_{k-1})$$

or

$$\mathbf{u}_{k-1} \cong \mathbf{M}_k \mathbf{u}_k \qquad \text{(a-8)}$$

where

$$\mathbf{M}_{k=}\mathbf{A}_{k-1}^{-1}\mathbf{B}_k \qquad \text{(a-9)}$$

and $\mathbf{B}_k$ is a combination matrix (between frame *k* and the warped frame *k-1*) from which $\mathbf{A}_k$ and $\Delta\mathbf{t}_k$ can be derived the same way as Eqs. (a-6) and (a-7). The algorithm is outlined as follows.

Step 1. Assume that the first frame is the reference frame, so that $\boldsymbol{u} = \boldsymbol{u}_0 + \boldsymbol{0}$, i.e. , $\boldsymbol{A_0} = \boldsymbol{I}, \ \boldsymbol{t_0} = \boldsymbol{0}$.

Assign *k* = 1.

Step 2. Match frame *k* and frame *k-1*, find a projective transformation $\mathbf{M}_k$ (Eq. (a-8)).

Step 3. By decomposing $\mathbf{B}_{k=}\mathbf{A}_{k-1}\mathbf{M}_k$ into $\mathbf{A}_k$ and $\Delta\mathbf{t}_k$, we can warp the current frame using $\mathbf{A_k}$ and find the translational parameters

$$\mathbf{t}_k = \mathbf{t}_{k-1} + \Delta\mathbf{t}_k \tag{a-10}$$

which is the camera location at time *k* in the mosaic representation.

Step 4. Assign $k = k+1$, go to Step 2.

After image rectification, stereo mosaics can be generated using the algorithms in Section 3. Analysis and real experiments show that the relative depth perception supports effective 3D viewing, since the local depth perception error is only relevant to the error propagation for a camera translation approximately equivalent to distance of the two slit windows, d_y. As an example, for the image mosaicing in Fig. 4, the slit window distance is 224 pixels, the interframe displacement is about 18 pixels/frame. So the relative depth error is connected to the cumulative mosaicing error of about 12 frames (224/18 frames).

A.1.2. Motion refinement for geo-mosaic

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 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 3D reconstruction. On the other hand, 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 zoom camera (Fig. 3). So it is impossible to be directly used for image rectification and fixed-line mosaicing. 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, without assuming small rotation angles and a virtual horizontal plane. An Extended Kalman Filter (EKF) approach could be used.

Assume that the focal length of the camera has been determined by camera calibration. The average height H of the terrain can be calculated from the range profiler. Fitting our problem into the EKF framework, the state is defined as $\mathbf{x}_k = (\mathbf{A}_k, \mathbf{t}_k)$, which consists of the warping matrix $\mathbf{A}_k$ and the mosaicing translation vector $\mathbf{t}_k$. The state is governed by the following equations

$$\begin{cases} \mathbf{A}_{k+1} = f_A(\mathbf{B}_{k+1}), \mathbf{B}_{k+1=}\mathbf{A}_k\mathbf{M}_{k+1} \\ \mathbf{t}_{k+1} = \mathbf{t}_k + \Delta\mathbf{t}_k, \Delta\mathbf{t}_k = f_t(\mathbf{B}_{k+1}) \end{cases} \tag{a-11}$$

where functions $f_A(\cdot)$ and $f_t(\cdot)$ are derived from equations (a-7) an (a-6). Eq. (a-11) can be expressed in the following nonlinear equation

$$\mathbf{x}_{k+1} = f(\mathbf{x}_k, \mathbf{w}_k) \tag{a-12}$$

where $\mathbf{w}_k$ is the error of the modeling in Eq. (a-11). The measurement $\mathbf{z}_k = (\mathbf{R}_k, \mathbf{T}_k)$ can be transformed to the state $\mathbf{x}_k$ by the following equations

$$\begin{cases} \mathbf{A}_k = \mathbf{F}\mathbf{Q}_k\mathbf{F}^{-1} \\ \mathbf{t}_k = (F\dfrac{T_x^{(k)}}{H}, F\dfrac{T_y^{(k)}}{H}) \end{cases} \tag{a-13}$$

The measurement can be expressed in the following equation

$$\mathbf{z}_k = h(\mathbf{x}_k, \mathbf{v}_k) \tag{a-14}$$

where $\mathbf{v}_k$ is the measurement error. After we build Eqs. (a-12) and (a-14) with the corresponding error models, an EKF algorithm can be applied [27][29]. The basic idea is that, even if the parametric motion model from image matching is only an approximation of the real motion, it is a good prediction of the motion parameters of the current frame. Then the measurement of the instrumentation package is used to correct this prediction in order to prevent the warping and translational parameters from drifting further away from the geo-referenced path. One of the key issues in the use of the EKF algorithm here is the error modeling of the "predictor" and the "corrector". The prediction error can be derived from the image match [5][15], while the measurement error can be obtained from the inherent error properties of the instrumentation. Obviously the relative errors in the image match are much smaller than errors in the geo-data, so roughly speaking, in the EKF approach, the geo-data is mainly used to constrain the error accumulating in the iterative image match. Future work is needed to validate this approach.

Appendix 2. Registration error in a multi-perspective mosaic

A simple analysis of the error due to a direct paste as in 2D manifold mosaic is given below, assuming that $T_x = 0$. We take the left mosaic as an example (Fig. A-2). Suppose that S_y is the y translational component between the current frame (of viewpoint T_y) and the next frame. The corresponding image displacement is $s_y=FS_y/H$. Ideally, the projections of the *parallel-perspective* mosaic in the y direction are parallel projections as the viewpoints continuously change from T_y to $T_y+S_y/2$. However, in generating a *multi-perspective* mosaic, because we only have discrete sampling along the y direction, we have to extract a "wide" slice - width introduced perspective distortion within the slice - hence in the generated multi-perspective mosaic, slices may not align at the "stitching" line.

More specifically, we need to extract a $s_y/2$-wide slice starting from the fixed line $y_0=d_y/2$ and ending at the stitching line $y_s = (d_y+s_y)/2$ in the current perspective image, due to the lack of frames between viewpoint T_y and $T_y+S_y/2$. There will be perspective distortions at other places within each slice other than the fixed line, since we use the same perspective projection from viewpoint T_y instead of the parallel projection from the changing viewpoints in the interval (T_y, $T_y+S_y/2$].

Now let us estimate the error of registration at one of the stitching lines; for example the right border y_s of the slice. Suppose that the point y_s pasted in the mosaic that can be seen from the front slit window of a "virtual" camera at viewpoint $T_y+S_y/2$ is (X, Y, Z), which is expressed in the coordinate system of viewpoint T_y. Then we have $Y = Z\frac{d_y/2}{F}+\frac{S_y}{2}$. Its image coordinate y in the perspective image of viewpoint T_y should be $y=\frac{d_y}{2}+H\frac{s_y/2}{Z}$ instead of y_s. However, due to the direct paste, a point at $y_s = (d_y+s_y)/2$ in the current is directly put at a point on the stitching line, so the error in y is

$$\delta y = y - y_s = \frac{H-Z}{Z}\frac{s_y}{2} \approx \frac{\Delta Z}{H}\frac{s_y}{2} \qquad \text{(a-15)}$$

where ΔZ = H-Z is the range of depth variation relative to the average height H, which can be connected to range of displacement Δy in the mosaic by $\Delta Z = H\frac{\Delta y}{d_y}$. Hence the y error can be calculated as

$$\delta y \approx \frac{\Delta y}{2d_y}s_y \qquad \text{(a-16)}$$

In conclusion, a 2D mosaicing method as directly cut-and-paste approach may produce a seam at the location of a stitching line due to large motion parallax under translational motion. *This misalignment error is proportional to the interframe displacement and the depth variation of a scene* (Eq. (a-15)). In a typical case of forest mosaics, the misalignment of the images is less than one pixel when the following conditions are met: depth variation is less than 10%, the interframe image displacement (s_y) is less than 10% of the distance (d_y) between two slit windows, and the intrinsic and extrinsic parameters of the camera are accurate enough . For example, if we have H = 390 m, d_y = 224 pixels, s_y = 16 pixels, and $\Delta y = \pm 6$ pixels. then the misalignment is $\delta y \approx \pm 0.2$ pixels. While the depth range is $\Delta Z = \pm 10.5\,\text{m}$, the depth error due to this misalignment is $\delta Z = H \dfrac{\delta y}{dy} = \pm 0.35\,\text{m}$ (i.e. 3%). Note that the amount of 0.2 pixels is equivalent to the sub-pixel match resolution. However the direct placement of each slice to form a so-called multi-perspective mosaic will not result in a seamless mosaic when the 3D variation is large, or if the velocity of the camera is high (i.e. s_y is large), or if the error in motion measurements cannot be neglected. For this reason we have developed a 3D mosaicing technique where local registration and view interpolation is applied between two successive frames.

Figure Captions

Fig. 1. The model of stereoscopic mosaics. (a). Left (blue) and right (red) views. (b). Stereo disparity geometry (side view). (c). Parallel-perspective stereo geometry: In each mosaic, each column is a full perspective projection, but successive columns have different viewpoints.

Fig. 2. Dense stereo mosaics. (a). In each multi-perspective projection mosaic, each sub-image (with more than one column) 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. (b). In a parallel-perspective projection mosaic, each column is full perspective, but successive columns are with different viewpoints of parallel rays. (c). Stereo geometry of multi-perspective mosaics. The shadowed cones show the sub-images from the two corresponding frames for right and left mosaics due to the sparseness of the frames. Usually a pair of corresponding points in the stereo mosaics will not come from the centers of the two slit windows, which are shown as red and blue lines.

Fig. 3. Three frames from a 165-frame image sequence. The inter-frame y-translations of this sequence are from 16.0 pixels to 24.0 pixels. (a) frame 20; (b) frame 21; and (c) frame 33. The two slit windows (rear red and front blue) are 224-pixel apart. These three images also show the dilemma of length of baseline versus common FOV in a classic stereo/motion approach. The common FOV of frame 20 and 21 is denoted as FOVab, while the common FOV of frame 21 and frame 33 is denoted as FOVbc.

Fig. 4. Stereo mosaics (3600x1382). (a) Left view mosaic, (b) Right view mosaic, and (c) Depth map. The offset of two slit windows is 224 pixels. A stereo pair of extended mosaics are virtually endless, hence producing almost the same FOVs. In the depth map, mosaic displacement is encoded as brightness (brightness is from 0 when Δy = 18.3 pixels, to 255 when Δy = -16.2 pixels). So higher elevation (i.e. closer to the camera) is brighter.

Fig. 5. Applications of stereo mosaics. (a). Stereoscopic viewing: red-blue overlaid stereo pair (Better stereo effect in the electronic version). (b). 3D rendering result: 3D recovery using stereo mosaics.

Fig. 6. Image rectification. (a) Original image sequence; (b) Rectified image sequence. In the rectified sequence, the nodal point of the "virtual" camera will move in a plane perpendicular to

its optical axis, and the Y axis remains to be parallel. In the video rectification process, the rotational components are compensated by rotational transforms, and the small translational component in Z is accounted by scaling factors. The translational components in X and Y are kept.

Fig. 7. Image mosaic basic: cut and paste

Fig. 8. View interpolation by re-projection of rays: (a) Given two successive views, how to interpolate all the views between them? (b). Why not full parallel projection? In Fig. 2a, IP represents the Interpretation Plane .

Fig. 9. Image morphing and stitching. (a). Local match, view interpolation and triangulation. (b) How to deal with occlusion?

Fig. 10. An example of local match and triangulation for both left and right mosaics. (a) The previous frame; (b) the current frame. (c) Frame difference (zoom window of b) after shifting (3, 36) pixels. The green crosses show the initially selected points (which are evenly distributed along the ideal stitching line) in the previous frame and its initial matches in the current frame by using the global transformation. The blue and red crosses show the correct match pairs by feature selection and correlation (red matches red, blue matches blue). The fixed lines, stitching lines/curves and the triangulation results are shown as yellow. The local match results (see zoom regions) show that points on the top of the narrow building have larger motion parallax than ground points. This can also be seen in the difference image in (c). Mosaicing results are shown in Fig. 12.

Fig. 11. An example of local match and triangulation for both left and right mosaics. (a) The previous frame; (b) the current frame. (c) Frame difference (zoom window of b) after shifting 27, 48) pixels. The green crosses show the initially selected points (which are evenly distributed along the ideal stitching line) in the previous frame and its initial matches in the current frame by using the global transformation. The blue and red crosses show the correct match pairs by feature selection and correlation (red matches red, blue matches blue). The fixed lines, stitching lines/curves and the triangulation results are shown as yellow. The local match results (see zoom regions) show that points on the top of the tall building have larger motion parallax (+4 pixels) than ground points. This can also be seen in the difference image in (c). Mosaicing results are shown in Fig. 12.

Fig. 12. Parallel-perspective mosaics of a campus scene from an airborne camera. The (parallel-perspective) mosaic (f) 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. The top three zoom sub-images of mosaics compare (a) multi-perspective mosaic of sparse image sequence; (b) parallel-perspective mosaic of sparse image sequence and (c) multi-perspective mosaic of dense image sequence (using all the 1000 frames). The bottom two zoom sub-images show how 3D mosaicing deals with large motion parallax of a tall building (d) 2D mosaic result (e) 3D mosaic result. As shown in Table 1, the misalignments of 1-2 pixels in (a) and ~4 pixels in (d) introduce not only visual seams but also about 10% errors in height estimations of the buildings.

Fig. 13. Depth from stereo mosaics. (a). Illustration of epipolar curves in stereo mosaics. (b) Histogram of mosaic displacement H(Δy) for the depth map of Fig. 4(c).   (c). Curves of the x translation (t_{xl}) in the left mosaic. (d). The x displacements corresponding to the max and min y displacements in the stereo mosaics.

Fig. 14. (a) Depth resolution of stereo mosaics, and (b) Depth resolution of two-view stereo

Fig. 15. Stereoscopic viewing geometry

Fig. 16. Red-blue overlaid image of campus stereo mosaics. (a) A wide FOV . (b) and (c) zoom FOVs.

Fig. A-1. Motion Refinement by an Extended Kalman Filter

Fig. A-2. Registration error due to direct paste

Table 1. Error analysis in 2D image mosaics in Fig. 10 - Fig. 12. (d_y = 192 pixels, H = 300 m)

Measurement ($\Delta Z = H\ \Delta y/d_y$, $\delta Z = H\ \delta y/d_y$) ($Z = H + \Delta Z$)	Fine Arts Center (long narrow building in Fig. 10)	Campus Center (tall building in Fig. 11)
Interframe motion (s_x, s_y)	(3, 36) pixels	(27, 48) pixels
Interframe misalignment δy	1-2 pixels	4 pixels
Mosaic displacements Δy	-12 pixels	-29 pixels
"Absolute" depth from camera Z	281.25 m	254.68 m
"Relative" height to ground ΔZ	18.75 m	45.31 m
Depth (height) error δZ	**1.56 -3.13 m**	**6.25 m**
Relative depth error (δZ/Z)	**0.55% - 1.1%**	**2.45%**
Relative height error ($\delta Z/\Delta Z$)	**8.3%-16.6%**	**13.8%**

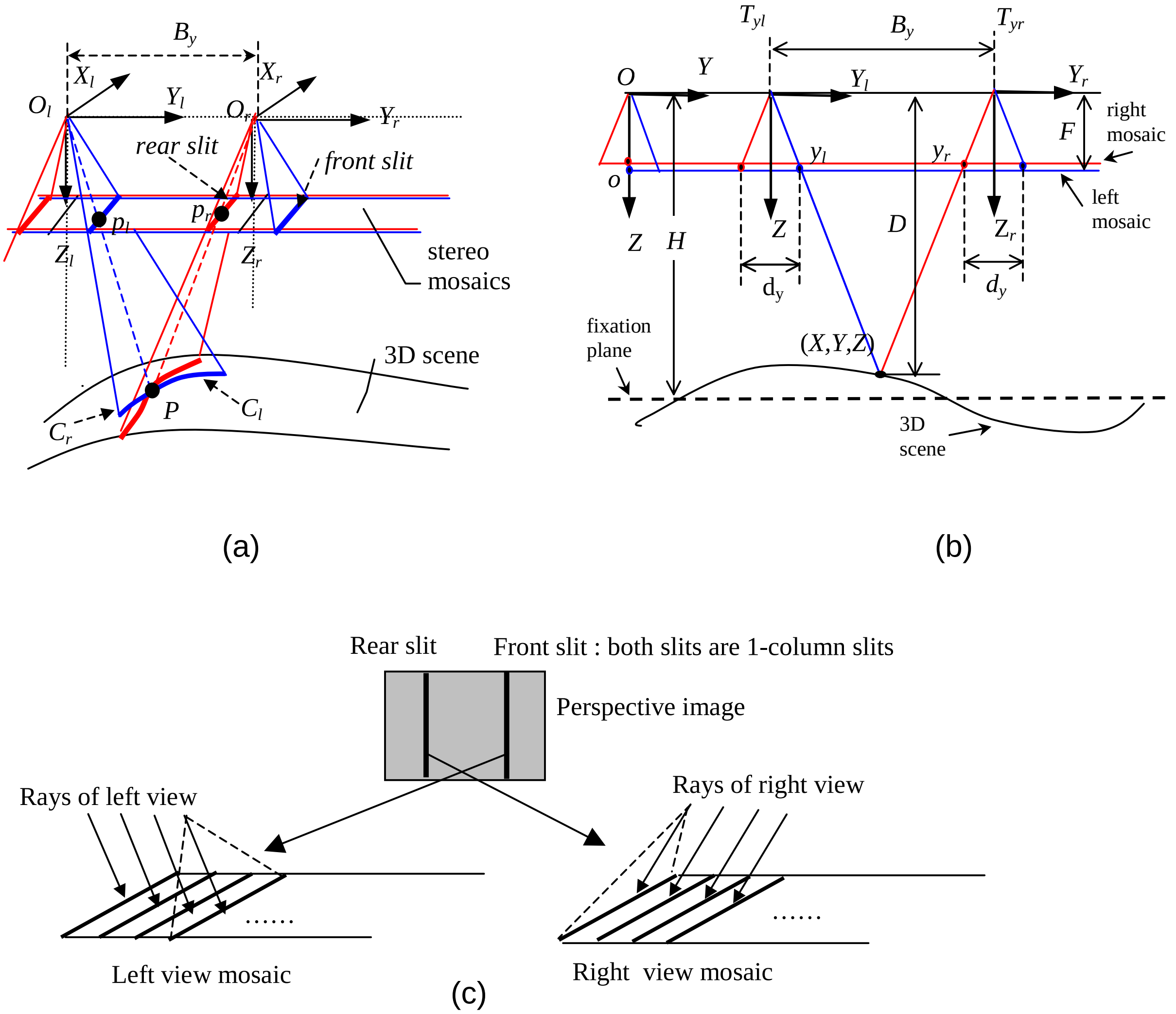


Fig. 1. The model of stereoscopic mosaics. (a). Left (blue) and right (red) views. (b). Stereo disparity geometry (side view). (c). Parallel-perspective stereo geometry: In each mosaic, each column is a full perspective projection, but successive columns have different viewpoints.

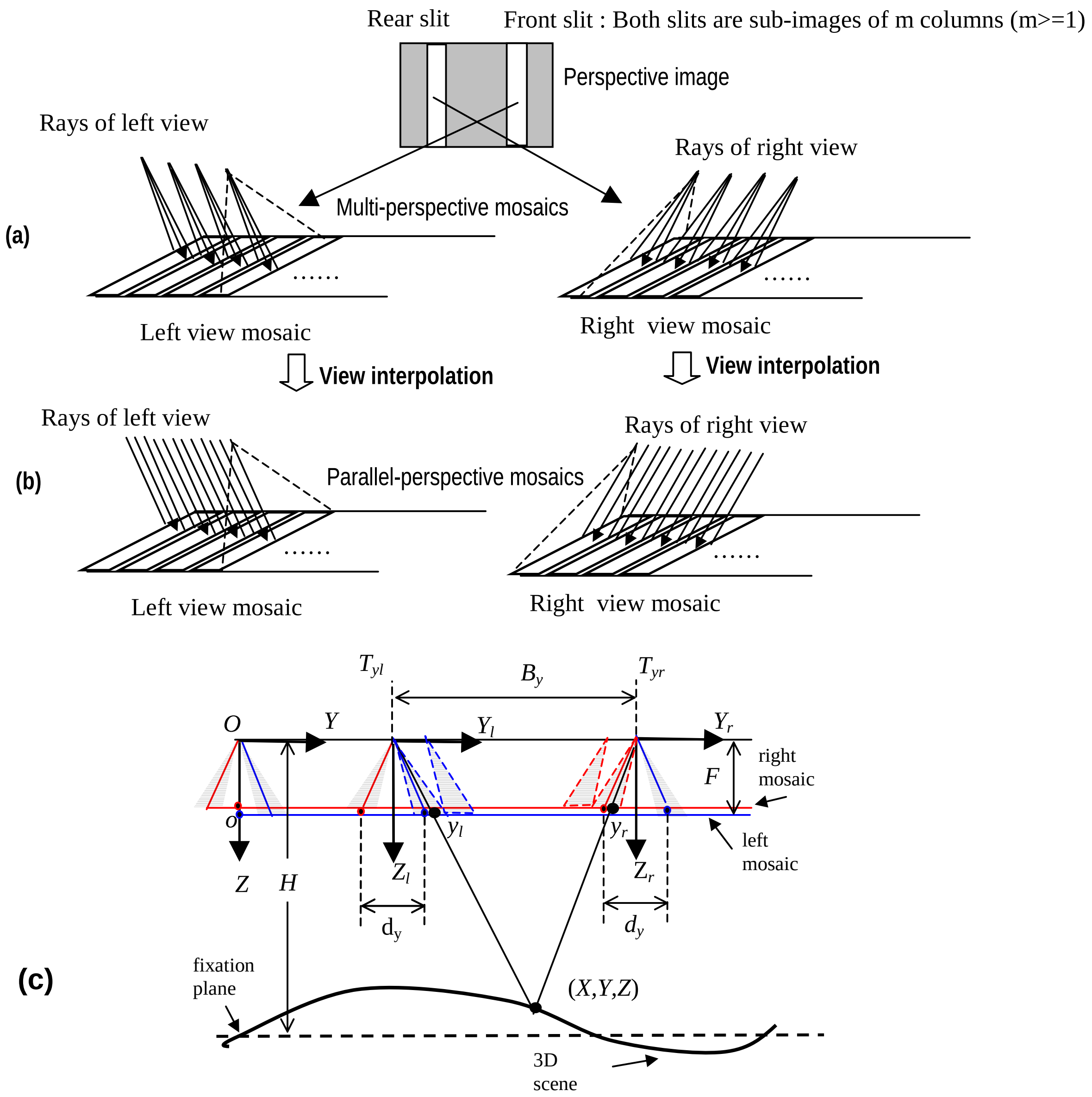


Fig. 2. Dense stereo mosaics. (a). In each multi-perspective projection mosaic, each sub-image (with more than one column) taken from the original image is of full perspective, but sub-images from different frames will have different viewpoints. This may cause seams in the mosaic due to motion parallax. (b). In a parallel-perspective projection mosaic, each column is full perspective, but successive columns are with different viewpoints of parallel rays. (c). Stereo geometry of multi-perspective mosaics. The shadowed cones show the sub-images from the two corresponding frames for right and left mosaics due to the sparseness of the frames. Usually a pair of corresponding points in the stereo mosaics will not come from the centers of the two slit windows, which are shown as red and blue lines.

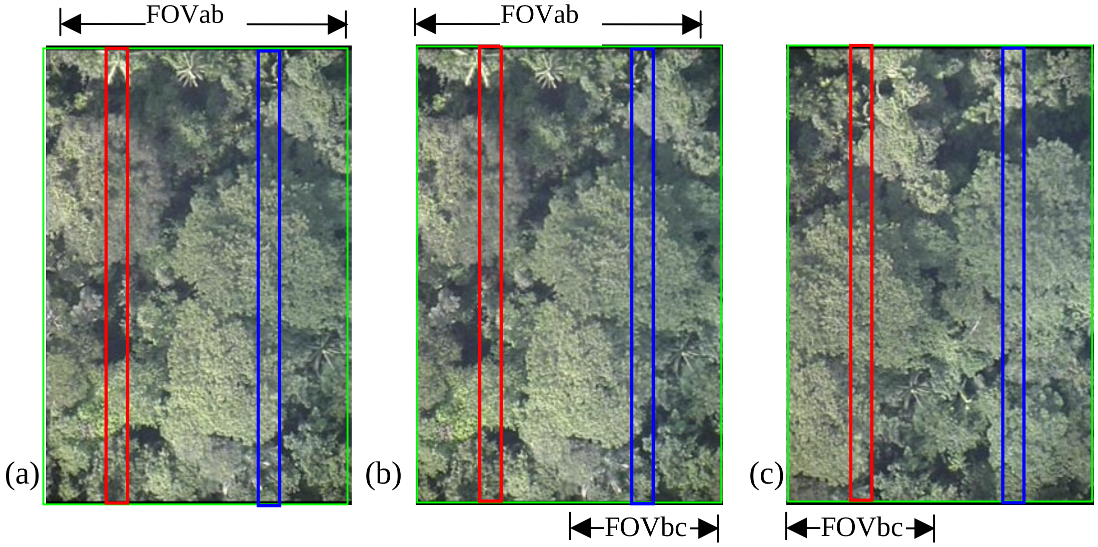


Fig. 3. Three frames from a 165-frame image sequence. The inter-frame y-translations of this sequence are from 16.0 pixels to 24.0 pixels. (a) frame 20; (b) frame 21; and (c) frame 33. The two slit windows (rear red and front blue) are 224-pixel apart. These three images also show the dilemma of length of baseline versus common FOV in a classic stereo/motion approach. The common FOV of frame 20 and 21 is denoted as FOVab, while the common FOV of frame 21 and frame 33 is denoted as FOVbc.

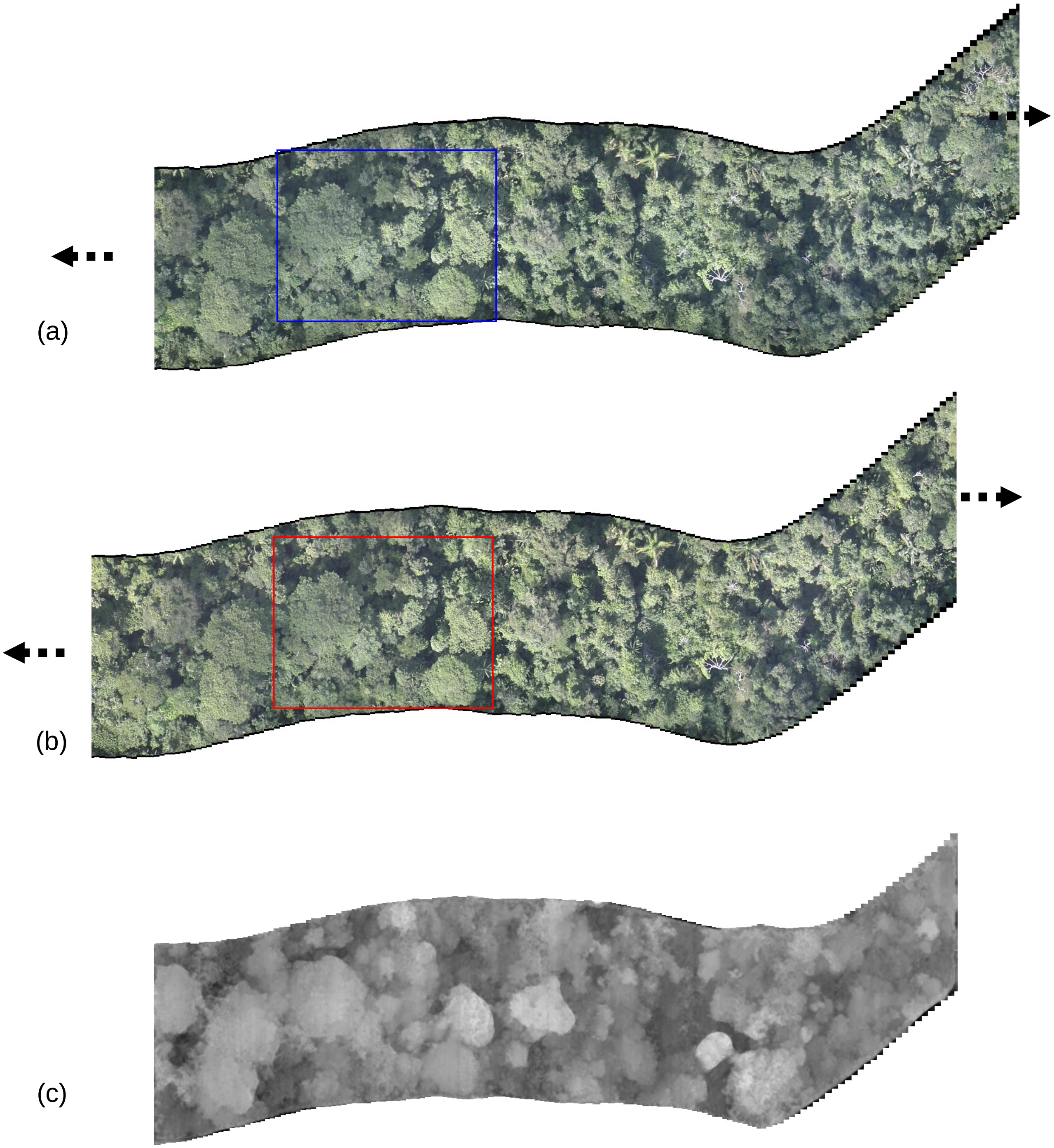


Fig. 4. Stereo mosaics (3600x1382). (a) Left view mosaic, (b) Right view mosaic, and (c) Depth map. The offset of two slit windows is 224 pixels. A stereo pair of extended mosaics are virtually endless, hence producing almost the same FOVs. In the depth map, mosaic displacement is encoded as brightness (brightness is from 0 when ∆y = 18.3 pixels, to 255 when ∆y = -16.2 pixels). So higher elevation (i.e. closer to the camera) is brighter.

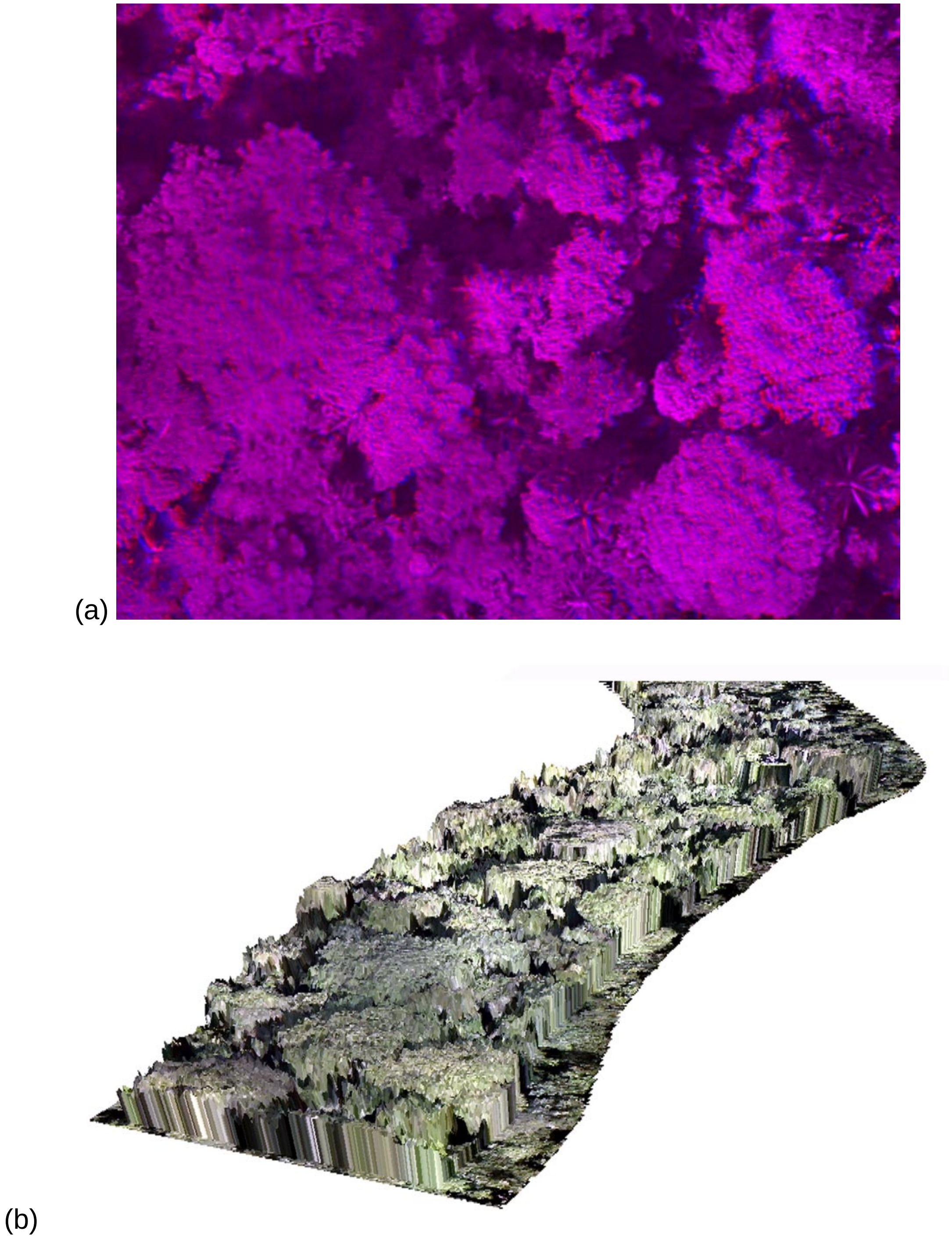


Fig. 5. Applications of stereo mosaics. (a). Stereoscopic viewing: red-blue overlaid stereo pair (Better stereo effect in the electronic version). (b). 3D rendering result: 3D recovery using stereo mosaics.

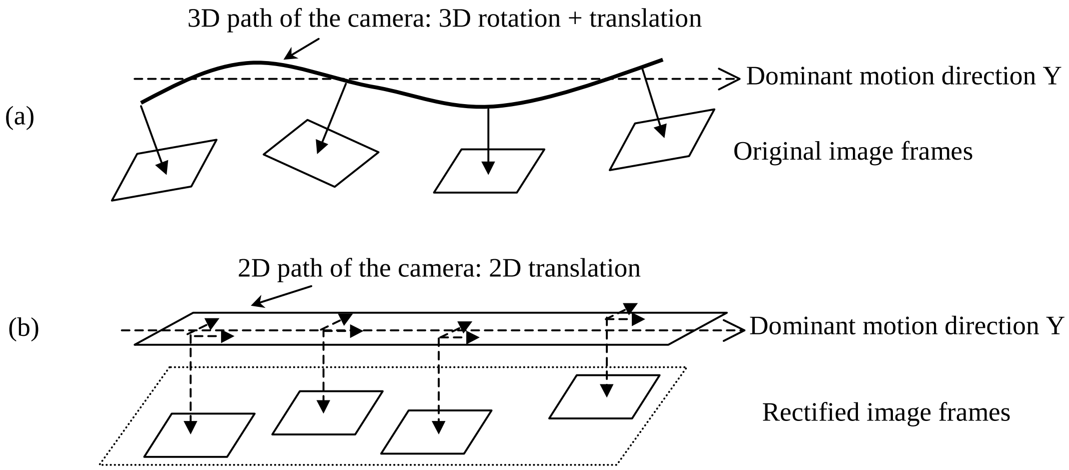

3D path of the camera: 3D rotation + translation
(a)
Dominant motion direction Y
Original image frames
2D path of the camera: 2D translation
(b)
Dominant motion direction Y
Rectified image frames

Fig. 6. Image rectification. (a) Original image sequence; (b) Rectified image sequence. In the rectified sequence, the nodal point of the "virtual" camera will move in a plane perpendicular to its optical axis, and the Y axis remains to be parallel. In the video rectification process, the rotational components are compensated by rotational transforms, and the small translational component in Z is accounted by scaling factors. The translational components in X and Y are kept.

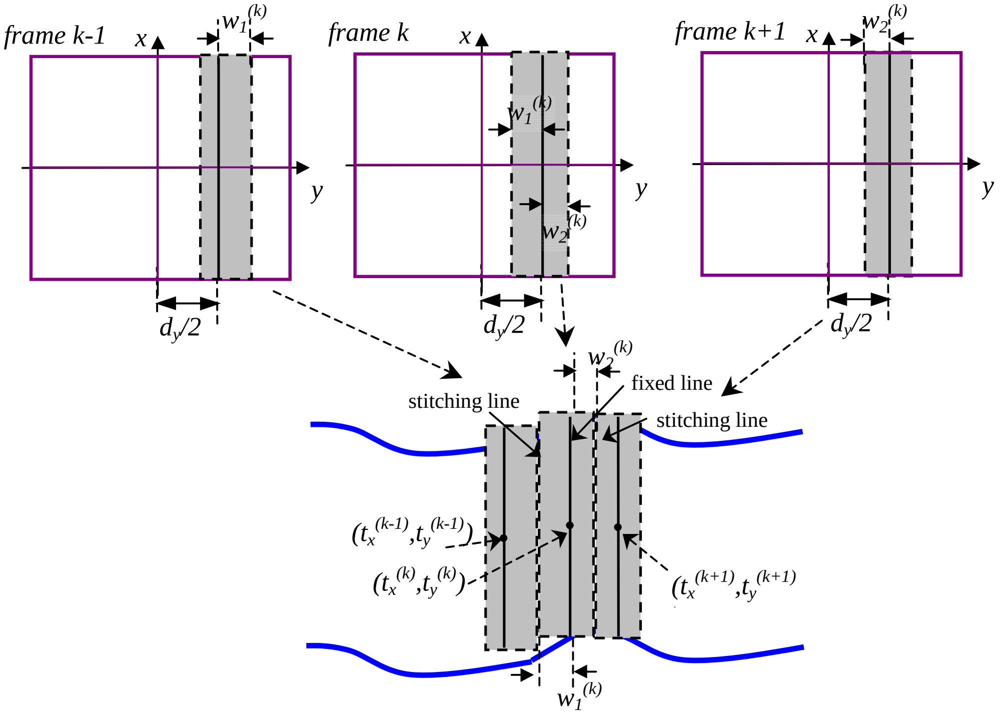


Fig. 7. Image mosaic basic: cut and paste

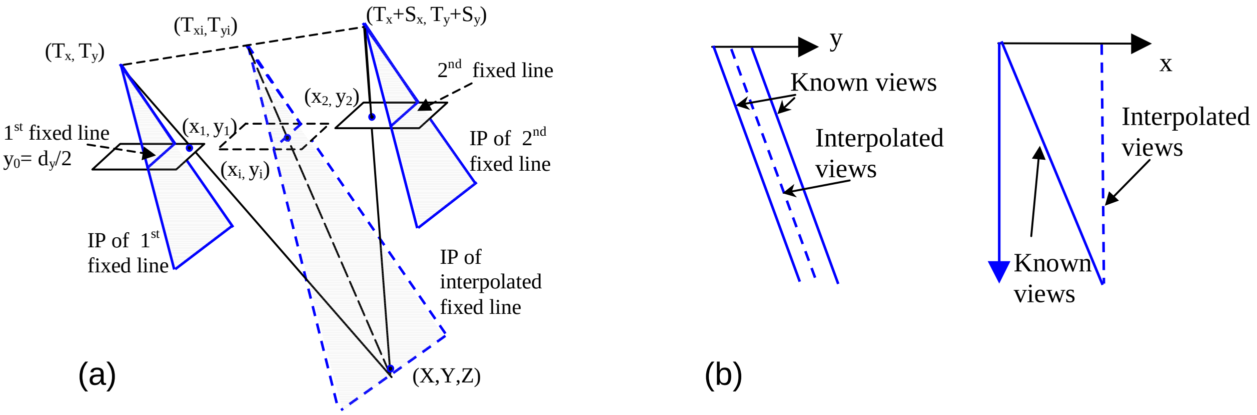


Fig. 8. View interpolation by re-projection of rays: (a) Given two successive views, how to interpolate all the views between them? (b). Why not full parallel projection? In Fig. 2a, IP represents the Interpretation Plane .

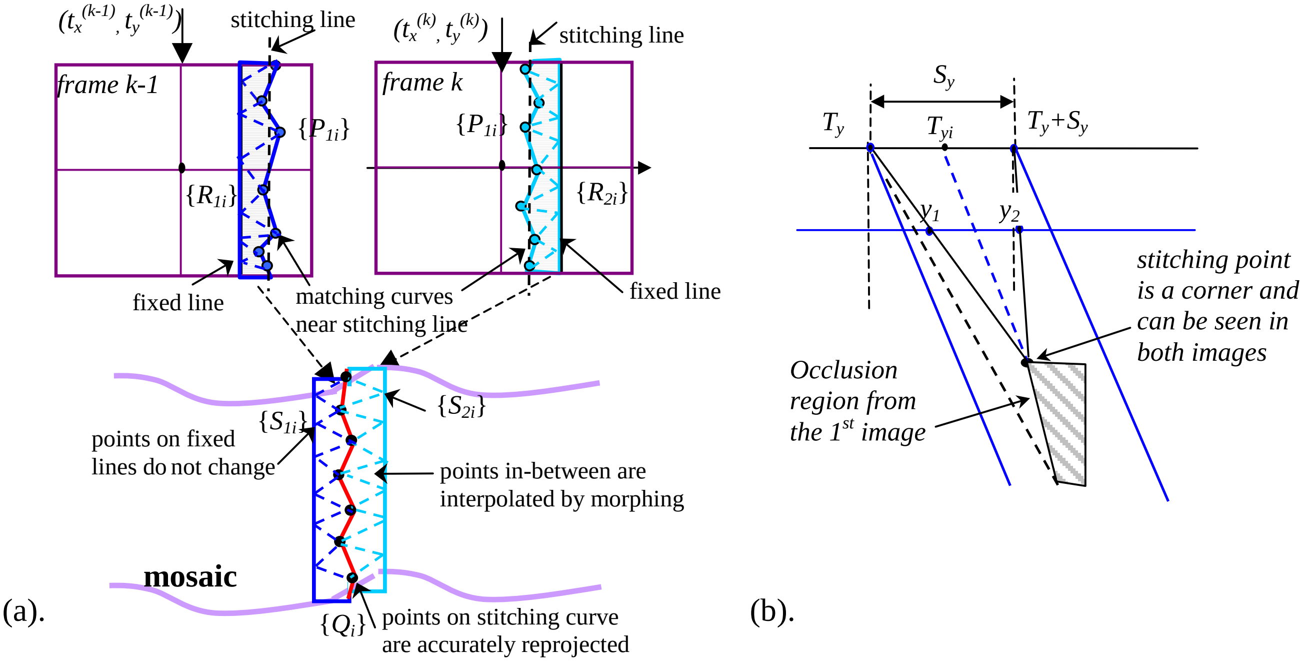

$(t_x^{(k-1)}, t_y^{(k-1)})$
stitching line
$(t_x^{(k)}, t_y^{(k)})$
stitching line
frame k-1
frame k
$\{P_{1i}\}$
$\{P_{1i}\}$
$\{R_{1i}\}$
$\{R_{2i}\}$
fixed line
matching curves near stitching line
fixed line
$\{S_{2i}\}$
$\{S_{1i}\}$
points on fixed lines do not change
points in-between are interpolated by morphing
mosaic
$\{Q_i\}$
points on stitching curve are accurately reprojected
(a).
S_y
T_y
T_{yi}
T_y+S_y
y_1
y_2
stitching point is a corner and can be seen in both images
Occlusion region from the 1st image
(b).

Fig. 9. Image morphing and stitching. (a). Local match, view interpolation and triangulation. (b) How to deal with occlusion?

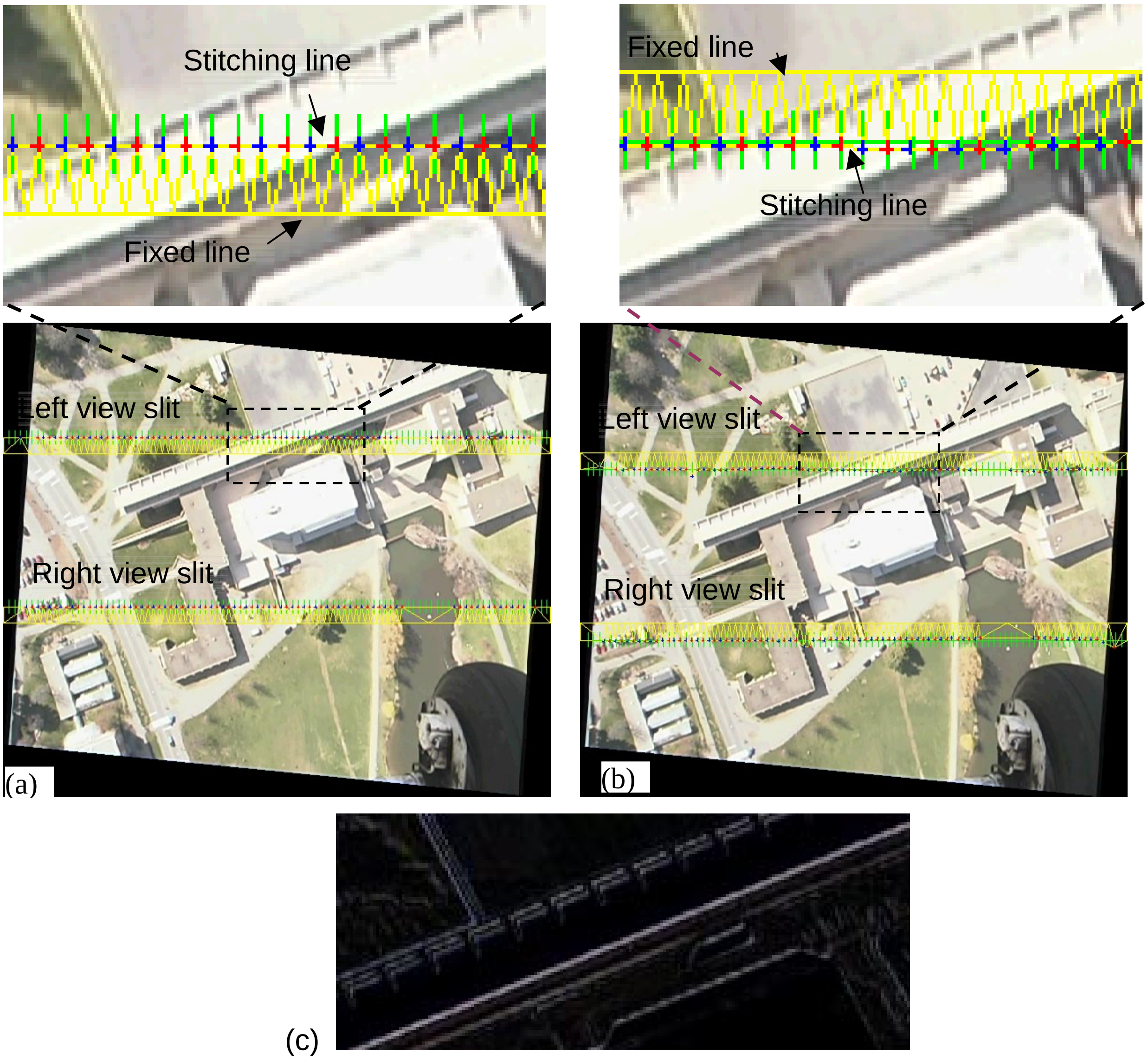


Fig. 10. An example of local match and triangulation for both left and right mosaics. (a) The previous frame; (b) the current frame. (c) Frame difference (zoom window of b) after shifting (3, 36) pixels. The green crosses show the initially selected points (which are evenly distributed along the ideal stitching line) in the previous frame and its initial matches in the current frame by using the global transformation. The blue and red crosses show the correct match pairs by feature selection and correlation (red matches red, blue matches blue). The fixed lines, stitching lines/curves and the triangulation results are shown as yellow. The local match results (see zoom regions) show that points on the top of the narrow building have larger motion parallax than ground points. This can also be seen in the difference image in (c). Mosaicing results are shown in Fig. 12.

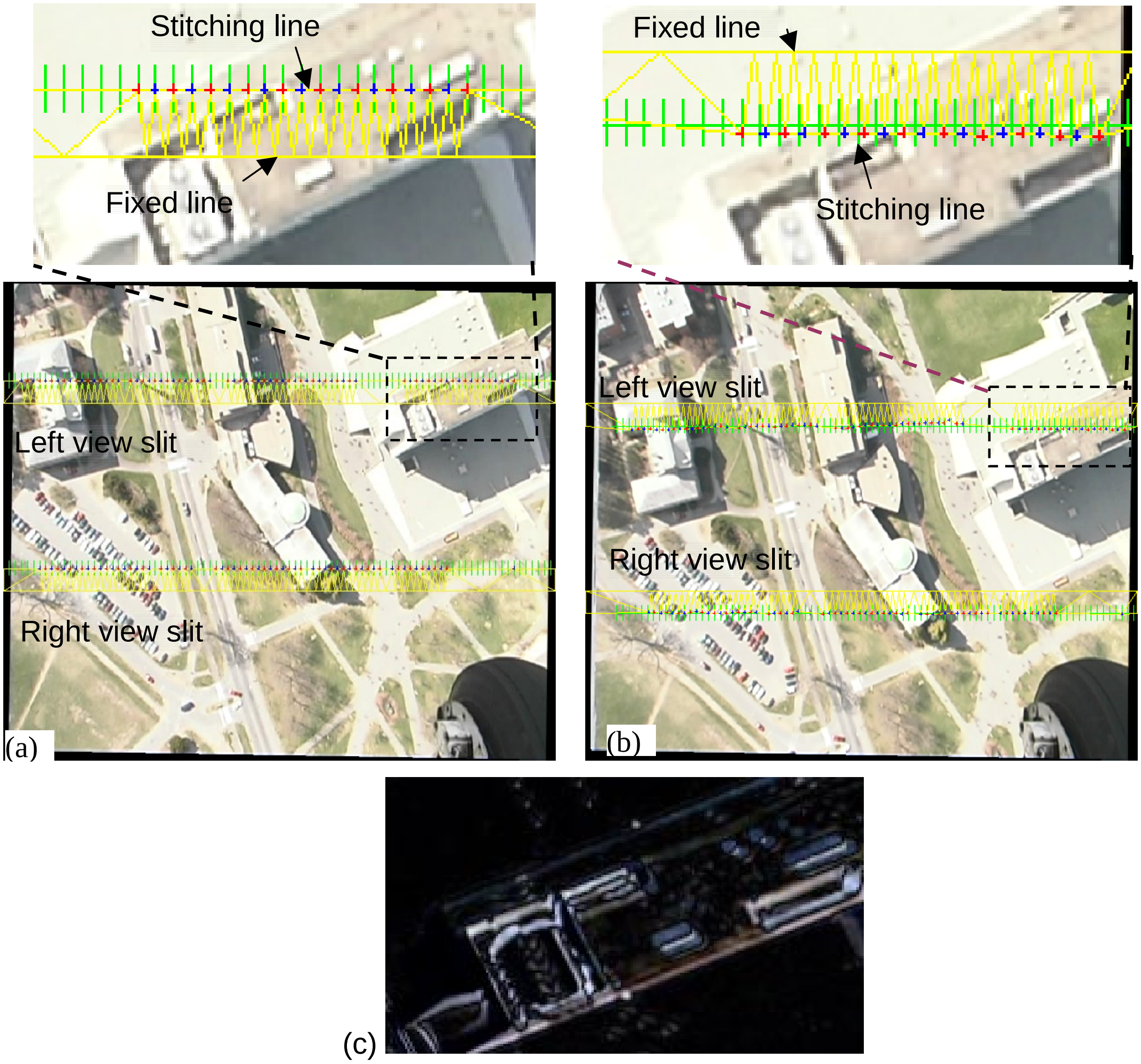
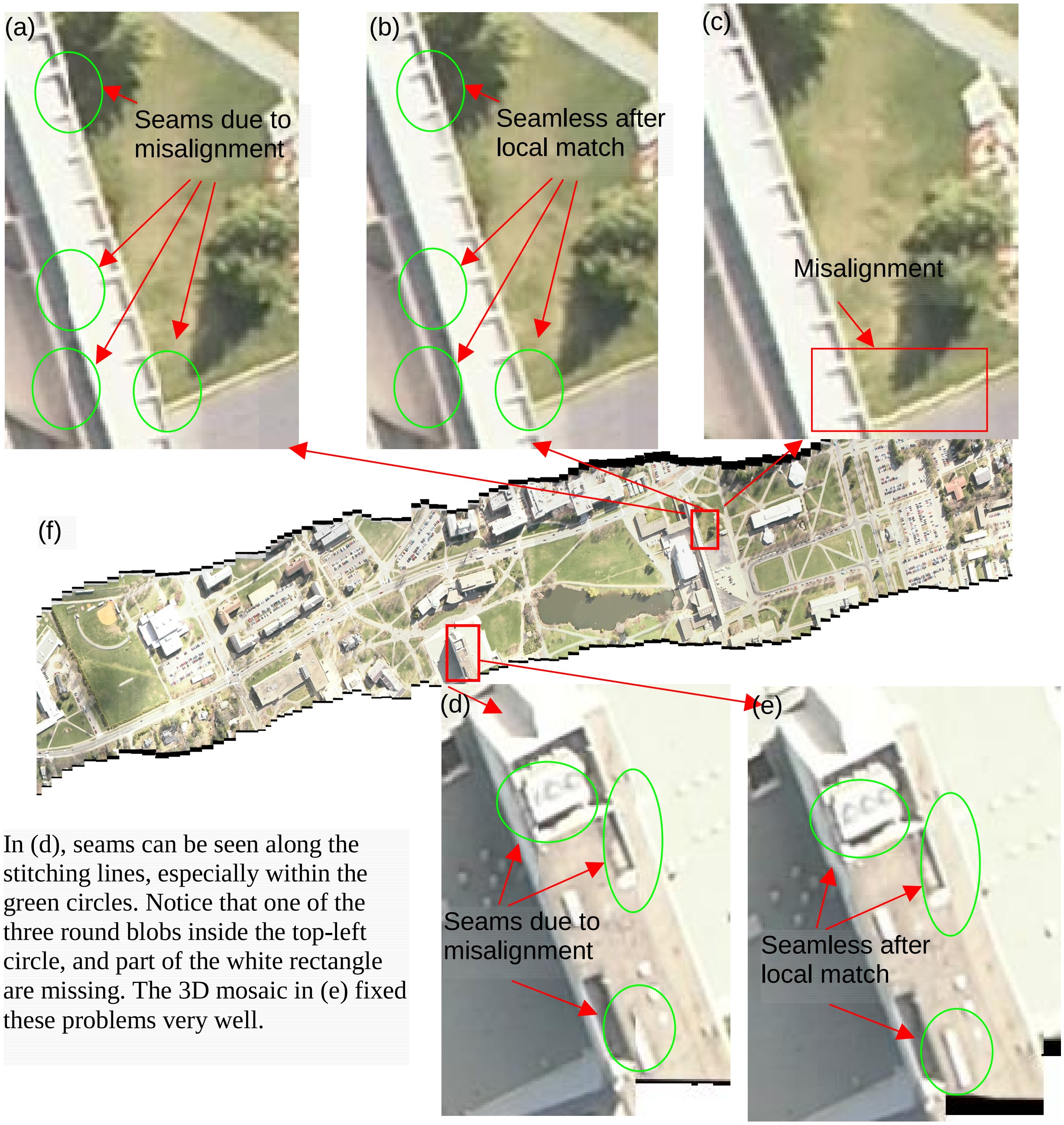
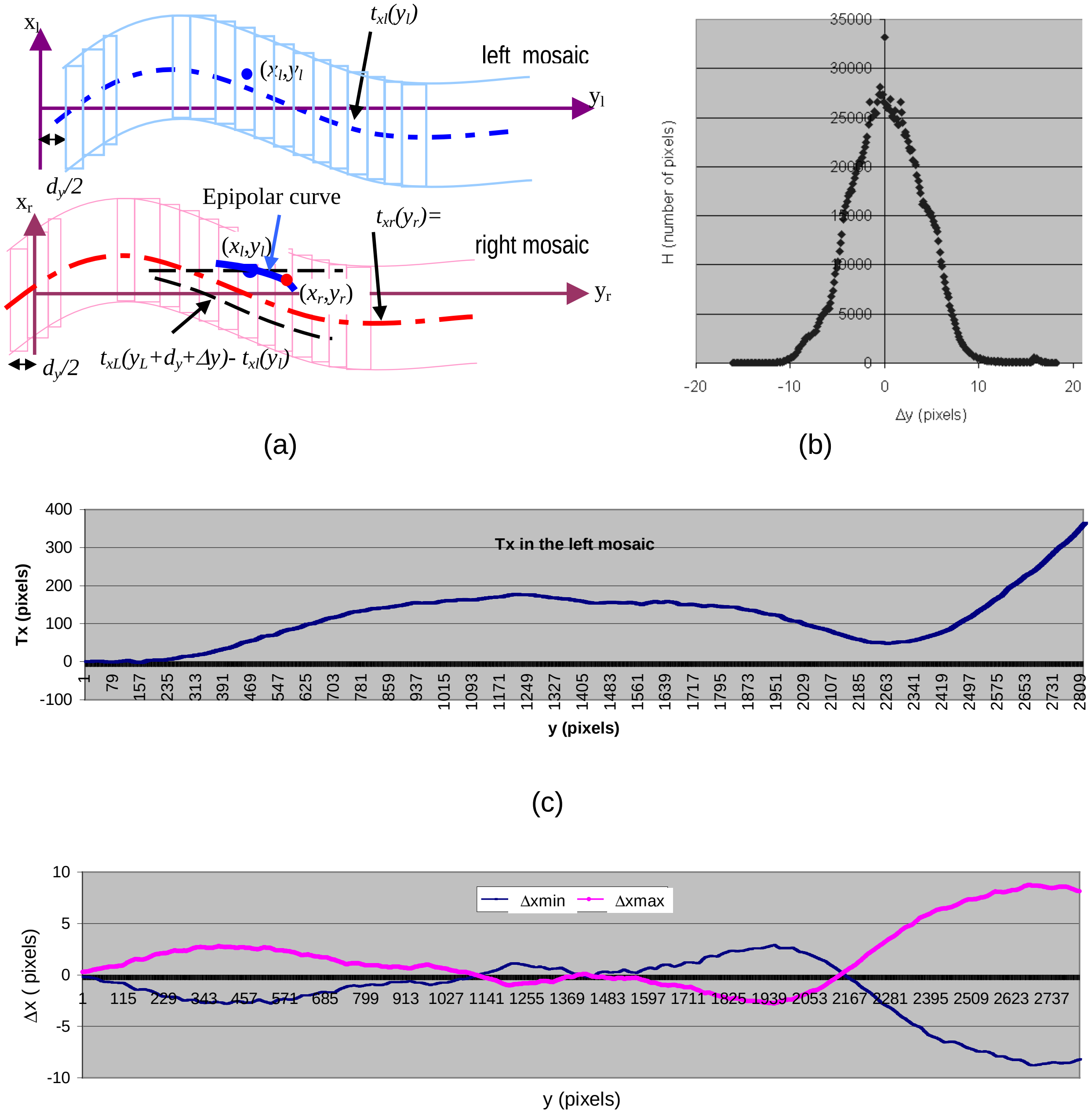


Fig. 11. An example of local match and triangulation for both left and right mosaics. (a) The previous frame; (b) the current frame. (c) Frame difference (zoom window of b) after shifting 27, 48) pixels. The green crosses show the initially selected points (which are evenly distributed along the ideal stitching line) in the previous frame and its initial matches in the current frame by using the global transformation. The blue and red crosses show the correct match pairs by feature selection and correlation (red matches red, blue matches blue). The fixed lines, stitching lines/curves and the triangulation results are shown as yellow. The local match results (see zoom regions) show that points on the top of the tall building have larger motion parallax (+4 pixels) than ground points. This can also be seen in the difference image in (c). Mosaicing results are shown in Fig. 12.


(a)
Seams due to misalignment
(b)
Seamless after local match
(c)
Misalignment
(f)
(d)
(e)
Seams due to misalignment
Seamless after local match

In (d), seams can be seen along the stitching lines, especially within the green circles. Notice that one of the three round blobs inside the top-left circle, and part of the white rectangle are missing. The 3D mosaic in (e) fixed these problems very well.

Fig. 12. Parallel-perspective mosaics of a campus scene from an airborne camera. The (parallel-perspective) mosaic (f) 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. The top three zoom sub-images of mosaics compare (a) multi-perspective mosaic of sparse image sequence; (b) parallel-perspective mosaic of sparse image sequence and (c) multi-perspective mosaic of dense image sequence (using all the 1000 frames). The bottom two zoom sub-images show how 3D mosaicing deals with large motion parallax of a tall building (d) 2D mosaic result (e) 3D mosaic result. As shown in Table 1, the misalignments of 1-2 pixels in (a) and ~4 pixels in (d) introduce not only visual seams but also about 10% errors in height estimations of the buildings.



(a)

(b)

(c)

(d)

Fig. 13. Depth from stereo mosaics. (a). Illustration of epipolar curves in stereo mosaics. (b) Histogram of mosaic displacement H(Δy) for the depth map of Fig. 4(c). (c). Curves of the x translation (t_{xl}) in the left mosaic. (d). The x displacements corresponding to the max and min y displacements in the stereo mosaics.

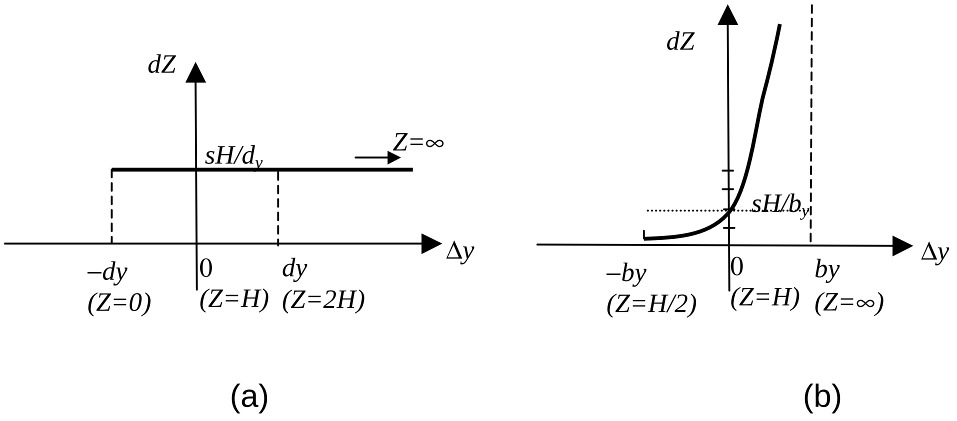


Fig. 14. (a) Depth resolution of stereo mosaics, and (b) Depth resolution of two-view stereo

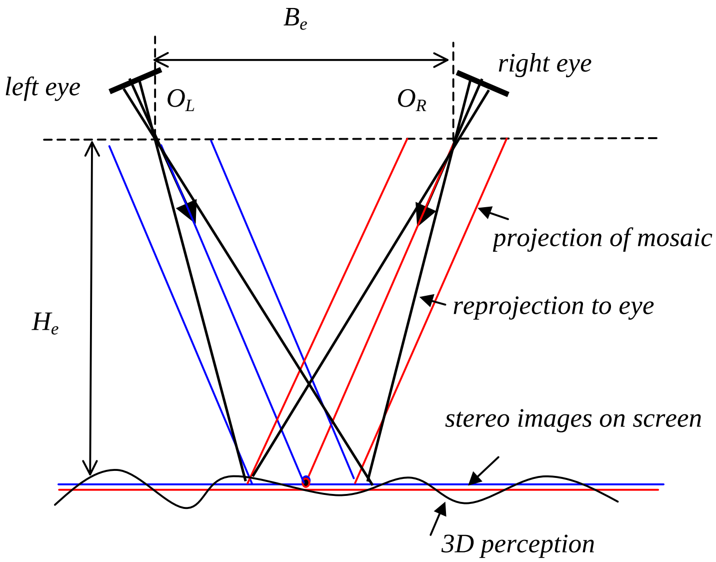
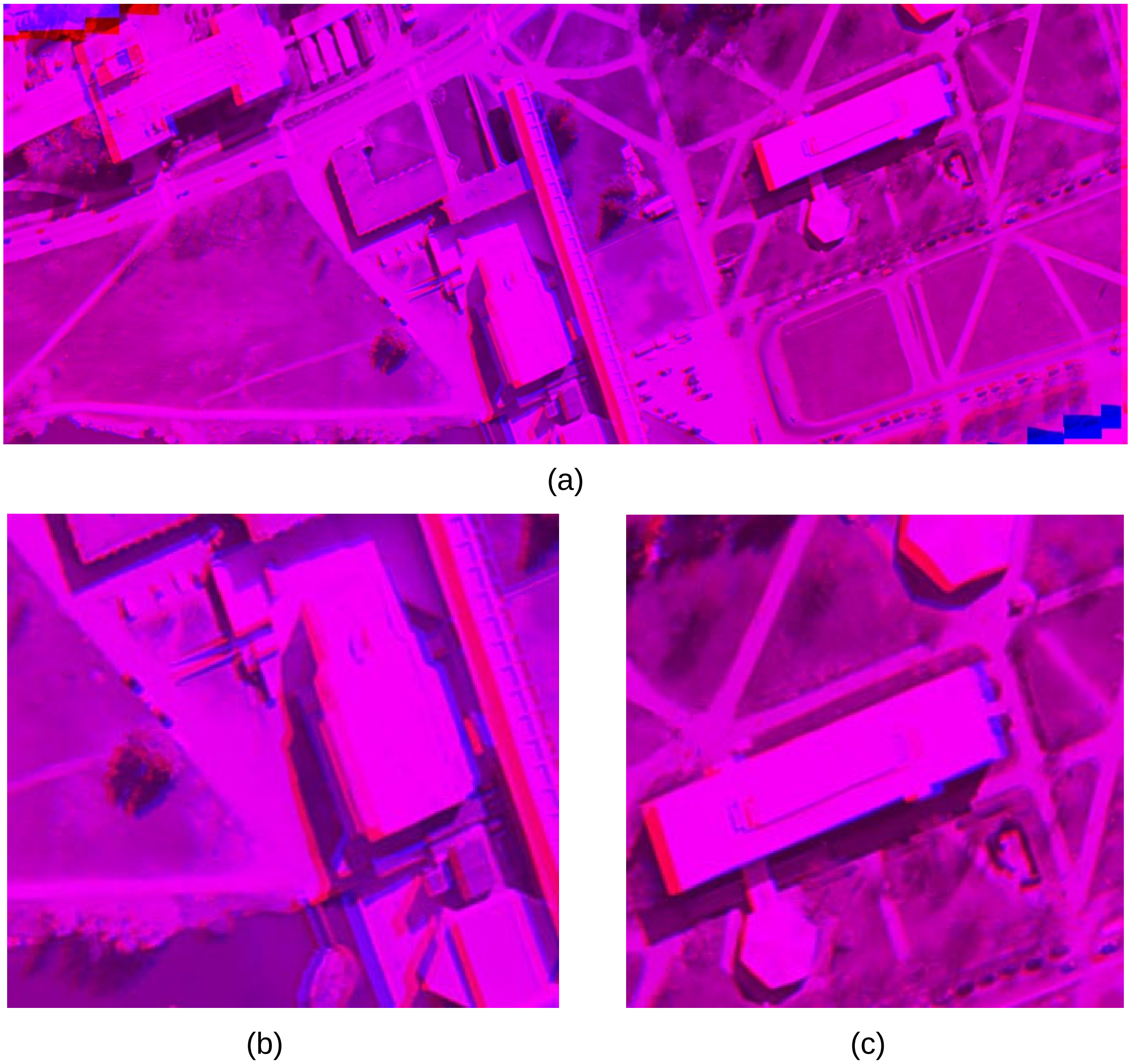


Fig. 15. Stereoscopic viewing geometry



(a)

(b)

(c)

Fig. 16. Red-blue overlaid image of campus stereo mosaics. (a) A wide FOV . (b) and (c) zoom FOVs.

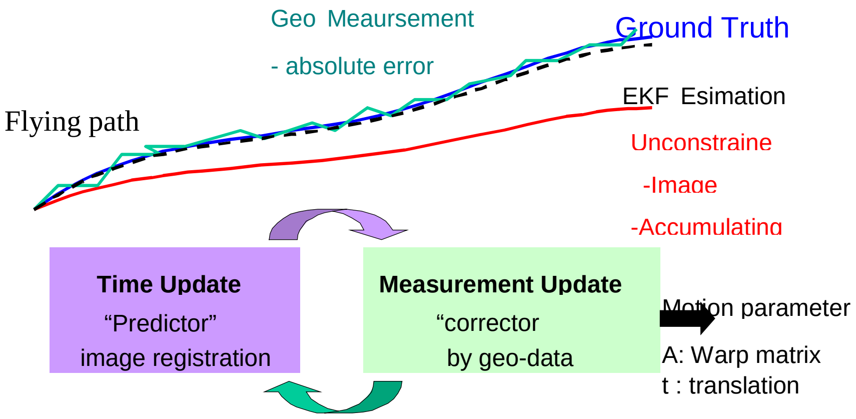


Fig. A-1. Motion Refinement by an Extended Kalman Filter

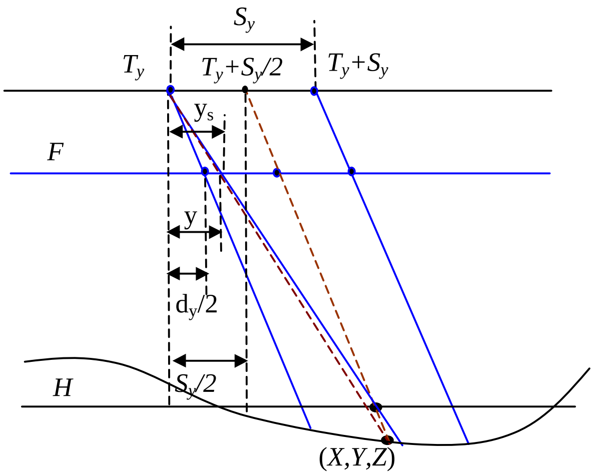


Fig. A-2. Registration error due to direct paste