Content-Based 3D Mosaic Representation for Video of Dynamic 3D Scenes

Zhigang Zhu, Hao Tang, George Wolberg
Department of Computer Science, The City College of New York, New York, NY 10031
{zhu| tang| wolberg}@cs.ccny.cuny.edu

Jeffery R. Layne
Air Force Research Laboratory, WPAFB, Ohio 45433-7318, USA
Jeffery.Layne@wpafb.af.mil

Abstract

We propose a content-based three-dimensional (3D) mosaic representation for long video sequences of 3D and dynamic scenes captured by a camera on a mobile platform. The motion of the camera has a dominant direction of motion (as on an airplane or ground vehicle), but 6 degrees-of-freedom (DOF) motion is allowed. In the first step, a pair of generalized parallel-perspective (pushbroom) stereo mosaics is generated that captured both the 3D and dynamic aspects of the scene under the camera coverage. In the second step, a segmentation-based stereo matching algorithm is applied to extract parametric representation of the color, structure and motion of the dynamic and/or 3D objects in urban scenes where a lot of planar surfaces exist. Based on these results, the content-based 3D mosaic (CB3M) representation is created, which is a highly compressed visual representation for very long video sequences of dynamic 3D scenes. Experimental results will be given.

Keywords: Image fusion, multi-image registration, video surveillance, content-based video coding

1. Introduction

In this paper we address the problems of visual representation for large amount of video stream data, of three-dimensional (3D) urban scenes in particular, captured by a camera mounted on an airborne or a ground mobile platform. Applications include airborne or ground video surveillance for moving target extraction, automated 3D urban scene construction, airborne traffic monitoring, and image-based rendering. For these applications, hours of video streams may be generated every time the mobile platform performs a data collection task. The data amount is in the order of 100 GB per hour for standard 640*480 raw color images. The huge amount of video data not only poses difficulties in data recording and archiving but also is prohibitive for users to retrieve and to review. In the past, video mosaic approaches [1-4] have been proposed for video representation and compression, but most of the work is for panning (rotating) cameras instead of the moving (translating) cameras mostly used in the cases of airborne or ground mobile surveillance, where obvious motion parallax is the main characterization of the video sequences due to the self-motion of the sensors. Some work has been done in 3D reconstruction of panoramic mosaics [5,6], but usually the results are 3D depth maps instead of high-level 3D scene understanding for static and/or dynamic target extraction and indexing. Layered representations [7-9] have been studied for motion sequence representations; however, the methods are usually computationally expensive, and the outputs are typically motion segmentation represented by affine planes instead of true 3D information. Efficient, high-level, content-based, and very low bit-rate representations of 3D scenes and moving targets are in great demand.

2. Overview of Our Approach

We propose a content-based 3D mosaic representation (CB3M) for long video sequences of 3D and dynamic scenes captured by a camera on a mobile platform. The motion of the camera has a dominant direction of motion (as on an airplane or ground vehicle), but 6 DOF motion is allowed. In the first step, a pair of generalized parallel-perspective (pushbroom) stereo mosaics is generated that captured both the 3D and dynamic aspects of the scene under the camera coverage. Bundle adjustment techniques [14] can be used for camera pose estimation, sometimes integrated with the geo-referenced data from GPS and INS when available. A ray interpolation approach [10] is used to generate a pair of seamless parallel-perspective (pushbroom) stereo mosaics under the obvious motion parallax of a translating camera. The pair of stereo mosaics is a compact representation for a long video sequence for a 3D scene with independent moving targets. Therefore, the mosaics are dynamic pushbroom stereo mosaics.

However, the representation is still an image-based one without object content representation. Therefore, in the second step, a segmentation-based stereo matching algorithm [11] is applied to extract parametric representation of the color, structure and motion of the dynamic and/or 3D objects in urban scenes, where a lot of planar surfaces exist. In the algorithm, we use the fact that all the static objects obey the epipolar geometry, i.e. along the epipolar lines of pushbroom stereo. An independent moving object (moving on a road surface), on the other hand, either violates the epipolar geometry if the motion is not in the direction of sensor motion or exhibits unusual 3D structure – e.g., obviously hanging above the road or hiding below the road.

Based on the above two steps, the content-based 3D mosaic (CB3M) representation is created. This is a highly compressed visual representation for a very long video sequence of a dynamic 3D scene. For example, a real image sequence of a campus scene has 1000 frames of 640*480 color images. With its CB3M representation, a compression ratio of more than 2,000 is achieved. More importantly, the CB3M representation has object contents.

The rest of the paper is organized as the follows. The two important steps to prepare the CB3M representation will be summarized based on our previous work. First, a brief summary of the dynamic pushbroom stereo mosaic step is given in Section 3. Second, the 3D and motion content extraction step is summarized in Section 4. Then in Section 5, the content-based 3D mosaic representation is described. Experimental results are given for the CB3M representation construction Section 6 gives concluding remarks.

3. Dynamic Pushbroom Stereo Mosaics

First, we assume the motion of a camera is an ideal 1D translation, the optical axis is perpendicular to the motion, and the frames are dense enough. Then, we can generate two spatio-temporal images by extracting two columns of pixels (perpendicular to the motion) at the leading and trailing edges of each frame in motion (Fig. 1). The mosaic images thus generated are *parallel-perspective*, which have perspective projection in the direction perpendicular to the motion and parallel projection in the motion direction. In addition, these mosaics are obtained from two different oblique viewing angles of a single camera's field of view, so that a stereo pair of left and right mosaics captures the inherent 3D information. The geometry in this ideal case (i.e. 1D translation with constant speed) is the same as the linear pushbroom camera model proposed in [15]. Therefore we also call this representation *pushbroom stereo mosaics* (we drop the term "linear" since the linear constraint will be removed in the general case when the camera motion is not 1D translation).

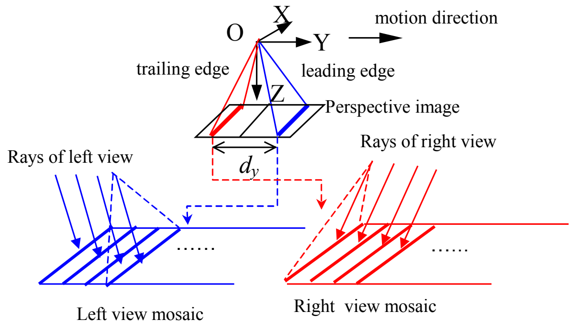

Fig. 1. Principle of the parallel-perspective pushbroom stereo mosaics

In real applications, there are two challenging issues. The first problem is that the camera usually cannot be controlled with ideal 1D translation and camera poses are unknown; therefore, camera orientation estimation (i.e., dynamic calibration) is needed. In our previous study on an aerial video application, we used external orientation instruments, i.e., GPS, INS and a laser profiler, to ease the problem of camera orientation estimation [10, 16]. More general approaches using bundle adjustment techniques [14] are under investigation for efficiently estimating camera poses of long image sequences. In this paper, we assume that the extrinsic and intrinsic camera parameters are known at each camera location. The second problem is to generate dense parallel mosaics with a sparse, uneven, video sequence, under a more general motion, and for a complicated 3D scene. To solve this problem, we have proposed a generalized stereo mosaic representation under constrained 6 DOF motion, and a *parallel ray interpolation for stereo mosaics* (PRISM) approach [10].

In principle, the PRISM approach needs to match all the points between the two overlapping slices of the successive frames to generate a complete parallel-perspective mosaic. In an effort to reduce the computational complexity, a fast PRISM algorithm [10] has been designed, based on the proposed PRISM method. It only requires matches between a set of point pairs in two successive images, and the rest of the points are generated by warping a set of triangulated regions defined by the control points in each of the two images. The proposed fast PRISM 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, or to perform pixel-wise dense matches to achieve true parallel-perspective (pushbroom) geometry.

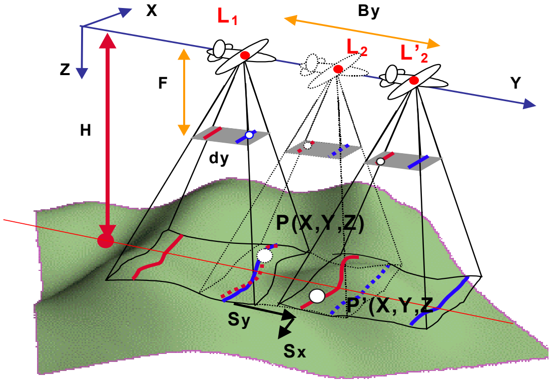

Fig. 2. Dynamic pushbroom stereo mosaics

Dynamic pushbroom stereo mosaics [11] are generated in the same way as with the static pushbroom stereo mosaics described above. Fig.2 illustrates the geometry. A 3D point *P(X,Y,Z)* on a target is first seen through the leading edge of an image frame when the camera is at location L_1. If the point P is static, we can expect to see it through the trailing edge of an image frame when the camera is at location L_2. The distance between leading and trailing edges is d_y (pixels), which denotes the constant "disparity" between this pair of images. However, if point P moves during that time, the camera needs to be at a different location L'_2 to see this moving point through its trailing edge. For simplifying equations, we assume that the motion of the moving point between two observations (L_1 and L'_2) is a 2D motion (S_x, S_y), which indicates that the depth of the point does not change over that period of time. Therefore, the depth of the moving point can be calculated as

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

where F is the focal length of the camera and B_y is the distance of the two camera locations (in the y direction). Mapping this relation into stereo mosaics following the notation in [10], we have

$$Z = H(\frac{d_y + \Delta y - s_y}{d_y}) \tag{2}$$

and

$$(S_x, S_y) = (Z\frac{s_x}{F}, H\frac{s_y}{F}) = (Z\frac{\Delta x}{F}, H\frac{s_y}{F}) \tag{3}$$

where H is the depth of plane on which we want to align our stereo mosaics, (Δx, Δy) is the *visual motion* of the moving 3D point P, which can be measured in the stereo mosaics. The vector (s_x, s_y) is the target motion represented in stereo mosaics. Obviously, we have $s_x = \Delta x$.

The above analysis only shows the geometry of a moving camera with 1D translational motion. In fact, a pair of generalized stereo mosaics can be generated when the camera undertakes a constrained 6 DOF motion. Details of the representation and algorithms can be found in [10]. Fig. 3 shows a red-blue stereo image with a pair of the dynamic pushbroom stereo mosaics generated from a video sequence with about 1000 frames, where the camera had obvious motion in the x direction as well as the y direction. In this figure, for stereo viewing with a pair of red-blue stereo glasses, the left mosaic is in the green and blue channels, and the right mosaic is in the red channel of a single RGB color bitmap. Visual displacements due to 3D structures and independent object motion can be observed if close-up views are shown.

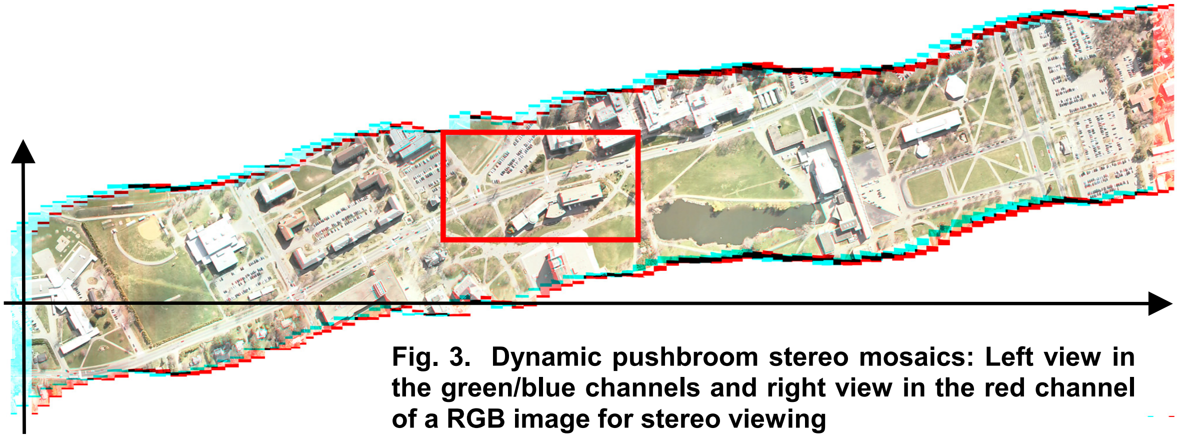

Fig. 3. Dynamic pushbroom stereo mosaics: Left view in the green/blue channels and right view in the red channel of a RGB image for stereo viewing

4. 3D and Motion Content Extraction

We have the following interesting observations about the dynamic pushbroom stereo geometry for 3D and moving target extraction.

(1) *Stereo fixation*. For a static point (i.e. $S_x = S_y = 0$), the visual motion of the point with a depth H is (0,0), indicating that the stereo mosaics thus generated fixate on the plane of depth H. This fixation facilitates the stereo matching and the detection of moving targets on that plane.

(2) *Motion accumulation*. For a moving point ($S_x \neq 0$ and/or $S_y \neq 0$), the motion between two observations accumulates over a period of time due to the large distance between the leading and trailing edges in creating the stereo mosaics. This will increase the discrimination of slow moving objects viewed from a relatively fast moving aerial camera.

(3) *Epipolar constraints.* In the ideal case of 1D translation of the camera (with which we present our dynamic pushbroom stereo geometry in this paper), the correspondences of static points are along horizontal epipolar lines, i.e. $\Delta x = 0$. (As a generalization, an epipolar curve geometry under 3D camera motion is given in [10].) Therefore, for a moving target P, the visual motion with nonzero Δx (i.e., the visual motion in the x direction) will identify itself from the static background in the general case, which implies that the motion of the target in the x direction is not zero (i.e., $S_x \neq 0$). In other words, the correspondence pair of such a point will violate the epipolar line constraint for static points (i.e. $\Delta x = 0$).

(4) *3D constraints.* Even if the motion of the target happens to be in the direction of the camera's motion (i.e., the y direction), we can still discriminate the moving target by examining 3D anomalies. Typically, a moving target (a vehicle or a human) moves on the flat ground surface (i.e., road) over the time period during which it is observed through the leading and trailing edges of video images with a limited field of view. We can usually assume that the moving target share the same depth as its surroundings, given that the distance of the camera from the ground is much larger than the height of the target. (The method to deal with 3D structure of 1 moving target is discussed in [11].) A moving target in the direction of camera movement, when treated as a static target, will show 3D anomaly - either hanging up above the road (when it moves to the opposite direction, i.e., $S_y < 0$), or hiding below the road (when it moves in the same direction, i.e., $S_y > 0$).

After a moving target has been identified, the motion parameters of the moving target can be estimated. We first estimate the depth of its surroundings and apply this depth Z to the target, then calculate the object motion s_y using Eq. (2) and (S_x, S_y), using Eq. (3), given the visual motion ($\Delta x, \Delta y$) measured in the stereo mosaics.

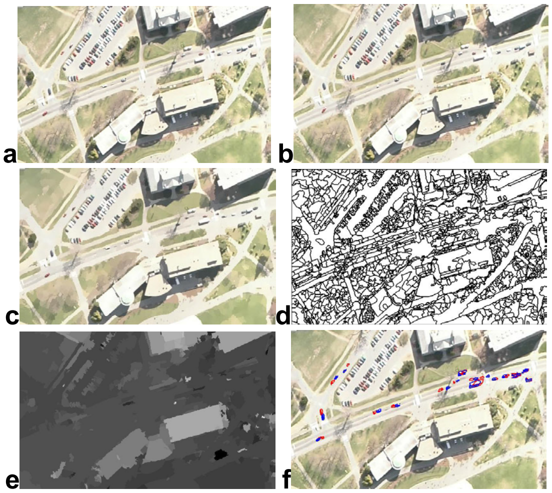

Fig. 4. Content-based 3D mosaic representation: results for a portion of the stereo mosaics marked in Fig. 3: (a) left color mosaic; (b) right color mosaic; (c) and (d) left color labels and region boundaries; (e) depth map of static regions; (f) moving targets (motion: blue to red). Note how close the color label image to the original color image is.

Based on these observations, the proposed segmentation-based stereo matching approach [11] integrates the estimation of 3D structure of an urban scene and the extraction of independent moving objects from a pair of dynamic pushbroom stereo mosaics in a unified framework. The algorithm starts with the left mosaic (see a portion in Fig. 4a), by segmenting it into *homogeneous* color regions that are treated as *planar* patches (Figs. 4c and 4d). We apply the mean-shift-based approach [12] for color segmentation. Then the stereo matching is performed based on these patches, called *natural matching primitives* [11], between two original color mosaics (Figs. 2a and 2b). The natural matching primitives are named since they are based on the real shapes of objects in the natural scenes. The basic idea is to only match those pixels that belong to each region (patch) between two color images in order to both produce sharp depth boundaries for man-made targets (Fig. 2d) and to facilitate the searching and discrimination of the moving targets (each covered by one or more homogeneous color patches) (Fig. 2f).

As a summary, the *natural matching* algorithm has the following four steps.

(1). *Stereo Matching*. After segmenting the left image using the mean-shift method, stereo matching is performed on each natural matching primitive, i.e. the selected "interest points" along the boundary of every patch. The following three sub-steps are performed: planar fitting for refining the local matches of interest points, neighborhood supporting for correcting possible errors of the local matches, and region merging so that those neighboring regions with the same planar parameters will be grouped into one larger, physically meaningful region.

(2). *Epipolar test*. Using pushbroom epipolar geometry in stereo matching, correct matches are found for the static objects, but moving objects will be those "outliers" without correct matches along epipolar lines.

(3). *3D anomaly test*. After ground surface fitting (and road detection if possible), moving objects in the same motion direction as the camera's will exhibit wrong 3D characteristics (e.g., hanging above roads or hiding below roads).

(4). *Motion extraction.* Searching matches for outliers (which are candidates for moving objects) with a 2D and larger search range, or along the road directions (if available).

5. CB3M: Content-Based 3D Mosaics

The proposed content-based 3D mosaic (CB3M) representation is a highly compressed visual representation for a very long video sequence of a dynamic 3D scene. It could fit into the MPEG-4 standard [13], in which a scene is described as a composition of several Video Objects (VOs), encoded separately.

A CB3M representation is a set of VO primitives (patches) that are defined as

$$\mathbf{CB3M} = \{VO_i,\ i = 1, \ldots, N\} = \{\ (c_i, \mathbf{b}_i, \mathbf{n}_i, \mathbf{m}_i),\ i = 1, \ldots, N\} \qquad (4)$$

where

(1) N is the number of VOs, i.e., natural patches (regions);

(2) c_i is the color (3 bytes) of the *ith* region;

(3) $\mathbf{b_i}$ is the 2D boundary of the *ith* region in the left mosaic, chain-coded as $\mathbf{b_i} = \{(x_0, y_0), b1, b2\ , \ldots\ b_{Ki}\}$, where the starting point (x_0,y_0) has 4 bytes, and each chain code has 3 bits. K_i is the number of boundary points and $K = \sum K_i$ is the total for all regions;

(4) $\mathbf{n_i} = (n_x, n_y, n_z, d)$ represents the plane parameters of the region in 3D, 4 bytes for each parameter; and

(5) $\mathbf{m_i}$ represents the L motion parameters of the region if in motion (e.g. L =2 for 2D translation on the ground).

Therefore the total data amount is

$$\begin{aligned} & N_{color} + N_{boundaryr} + N_{structure} + N_{motion} \\ & = 3N + (4N+3K/8) + 4*4N+4L*N_m \\ & = 23N+3K/8+4LN_m \text{ (bytes)} \qquad (5) \end{aligned}$$

when each of the motion and structure parameters needs 4 bytes. In the above equation, N_m is the number of moving regions (which is much smaller than the total region number N).

The compression of a video sequence comes from two steps: stereo mosaicing and then content extraction. For the real image sequence of a campus scene we discussed before, we have 1000 frames of 640*480 color images, so the data amount is 879 MB. The size of pair of the stereo mosaics (Fig. 3) is 4448*1616*2, which has 41MB (without compression and with more than half empty space due to the fact that the mosaics go in a diagonal direction). The two mosaics in high-quality JPEG format only have 2*560 KB; therefore, a compression ratio of about 800 is achieved for the stereo mosaics (the first step).

Then after color segmentation 3D planar fitting and motion estimation, we obtained the CB3M representation of the video sequence, with the total number of the natural regions N = 20,636 and the total number of boundary points K = 1,009,247. The total amount of data in its CB3M representation is 888 KB (with a header but without coding the motion). This real file size is consistent with the estimation of data amount using Eq. (5), which is about 833 KB. The data amount is reduced to 398 KB with a simple lossless Winzip on the CB3M data; therefore, the compression ratio is about 2261. More importantly, the CB3M representation has object contents which can be used for object indexing, retrieval and image-based rendering.

6. Conclusions

In this paper we propose a content-based 3D mosaic representation (CB3M) for long video sequences of 3D and dynamic scenes captured by a camera on a mobile platform. In real applications, the motion of the camera should have a dominant direction of motion (as on an airplane or ground vehicle), but 6 DOF motion is allowed. In the first step, a pair of generalized parallel-perspective (pushbroom) stereo mosaics is generated that captures both the 3D and dynamic aspects of the scene under the camera coverage. In the second step, a segmentation-based stereo matching algorithm is applied to extract parametric representation of the color, structure and motion of the dynamic and/or 3D objects, and to represent them as planar surface patches.

The content-based 3D mosaic (CB3M) representation is a highly compressed visual representation for very long video sequences of dynamic 3D scenes. It could fit into the MPEG-4 standard, in which a scene is described as a composition of several Video Objects (VOs), encoded separately. The compression of a video sequence comes from both steps: stereo mosaicing and then content extraction. For the real image sequence of a campus scene discussed in the paper, with 1000 frames of 640*480 color images, a compression ratio of more than 2,000 is achieved. More importantly, the CB3M representation has object contents represented. The CB3M representation presented in this paper, however, it still in its conceptual level, and many details and practical issues have not been considered. First, more experiments are needed with both simulated and real video sequences to evaluate the coding and compression capabilities of this representation. Second, in order to use the CB3M representations for real applications, further enhancements are also needed. For example, the neighboring regions, which have been extracted in the second steps, and which are important in object recognition and occlusion handling in image rendering, are not represented in the current model. Developing higher-level representations that group the lower-level natural patches for physical objects may also be very useful for many applications.

7. Acknowledgements

This work is supported by the Air Force Research Lab under the AFRL/SN RASER program for Multi-Sensor Registration (Award No. FA8650-05-1-1853) and an AFRL/HECB Grant for Multimedia Surveillance (Award No. F33615-03-1-63-83). The work is also supported by Army Research Office through Grant No. W911NF-05-1-0011 and by New York Institute for Advanced Studies. Some data were captured in the UMass Computer Vision Lab under an NSF grant (No EIA-9726401). The U.S. Government is authorized to reproduce and distribute reprints for Governmental purposes notwithstanding any copyright notation thereon. However, the views and conclusions contained herein are those of the authors and should not be interpreted as necessarily representing the official policies or endorsements, either expressed or implied, of the U.S. Government. Thanks are also given to Mr. Robert Hill at the City College of New York for proofreading the manuscript.

8. References

[1]. M. Irani, P. Anandan, J. Bergen, R. Kumar, and S. Hsu, Mosaic representations of video sequences and their applications. *Signal Processing: Image Communication*, vol. 8, no. 4, May 1996.

[2]. S. Hsu, P. Anandan, Hierarchical representations for mosaic based video compression, In *Proc. Picture Coding Symp.*, 395-400, March 1996.

[3]. F. Odone, A. Fusiello and E. Trucco, Robust motion segmentation for content-based video coding, *the 6th Conference on Content-based Multimedia Information Access*, College de France, 2000: 594-601.

[4]. W. H. Leung and T. Chen, Compression with mosaic prediction for image-based rendering applications*, IEEE Intl. Conf. Multimedia & Expo*., New York, July 2000.

[5]. Y. Li, H.-Y. Shum, C.-K. Tang, R. Szeliski, Stereo reconstruction from multiperspective panoramas. *IEEE Trans. on PAMI*, 26(1), 2004: pp 45-62.

[6]. C. Sun and S. Peleg, Fast Panoramic Stereo Matching using Cylindrical Maximum Surfaces, *IEEE Trans. SMC Part B*, 34, Feb. 2004: 760-765.

[7]. J. Xiao and M. Shah, Motion layer extraction in the presence of occlusion using graph cut, In *Proc. CVPR'04.*

[8]. Y. Zhou and H. Tao, A background layer model for object tracking through occlusion," In *Proc. ICCV'03*: 1079-1085.

[9]. Q. Ke and T. Kanade, A subspace approach to layer extraction, In *Proc. IEEE CVPR'01.*

[10]. Z. Zhu, E.. Riseman, A. Hanson, Generalized Parallel-Perspective Stereo Mosaics from Airborne Videos, *IEEE Trans. PAMI*, 26(2), Feb 2004, 226-237.

[11]. Z. Zhu, H. Tang, B. Shen, G. Wolberg, 3D and Moving Target Extraction from Dynamic Pushbroom Stereo Mosaics, *IEEE Workshop on Advanced 3D Imaging for Safety and Security* (with CVPR'05), June 25, 2005, San Diego, CA, USA.

[12]. D. Comanicu and P. Meer, Mean shift: a robust approach toward feature space analysis. *IEEE Trans. PAMI*, May 2002.

[13]. R. Koenen, F. Pereira and L. Chiariglione, MPEG-4: Context and objectives. *Signal Processing: Image Communications*, 9(4), 1997:295-300.

[14]. B. Triggs, P. McLauchlan, R. Hartley and A. Fitzgibbon, Bundle Adjustment -- A Modern Synthesis, In *Vision Algorithms: Theory and Practice*, Lecture Notes in Computer Science, vol 1883, pp 298--372, 2000, eds. B. Triggs, A. Zisserman and R. Szeliski", Springer-Verlag.

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

[16]. Z. Zhu, E. M. Riseman, A. R. Hanson and H. Schultz, An Efficient Method for Geo-Referenced Video Mosaicing for Environmental Monitoring. *Machine Vision Applications Journal*, 2005 (to appear).