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

Zhigang Zhu, Hao Tang and George Wolberg, City College of New York, New York, NY 10031, USA
Jeffery R. Layne, Air Force Research Laboratory, WPAFB, Ohio 45433-7318, USA

Introduction
In this paper we address visual representation problems for large amount of video streams, of 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, 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, and the huge amount of video data is prohibitive for users to review. In the past, video mosaic approaches [1-4] have been proposed for video compression, but most of the work is for a rotating camera instead of a translating camera in the cases of airborne or ground mobile surveillance where obvious motion parallax is the main characteristics of the video sequences. 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 greatly demanded.

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 adjustments are 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 seamless parallel-perspective stereo mosaics under obvious motion parallax of a translating camera. The pair of stereo mosaics is a compact representation for a long video sequence (as in Fig.1 for a 1000-frame video). However, the representation is still an image-based one without 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 – obviously hanging above the road or hiding below the road.

(Summary)
Note: Left view in the green/blue channels and right view in the red channel of a RGB image for stereo viewing

Fig. 1. Dynamic pushbroom stereo mosaics

Fig. 2 Content-based 3D mosaic representation: results for a portion of the stereo mosaics marked in Fig. 1: (a) left color mosaic; (b) left color labels; (c) region boundaries; (d) depth map of static regions; (d) moving targets (motion: blue to red). Note how close the color label image to the original color image is.
**Segmentation-based stereo matching approach**

The proposed segmentation-based stereo matching approach 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. 2a), by segmenting it into homogeneous color regions that are treated as planar patches (Figs. 2b and 2c). 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, between two original color mosaics. The basic idea is to only match those pixels that belong to each region (patch) between two 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. 2e).

The proposed algorithm has the following four steps. (1) **Stereo Matching.** After segmenting the left image using the mean-shift method, local matching is applied based on each natural matching primitive in every patch. Moreover, matches are refined by the performing planar fitting, neighborhood supporting 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 on 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), moving objects in the same motion direction will exhibit wrong 3D characteristics (hanging above roads or hiding below roads). (4) **Motion extraction.** Search matches for outliers (which could be moving objects) with a 2D and larger search range, or along the road directions (if available).

**CB3M: Content-based 3D mosaic representation**

The content-based 3D mosaic (CB3M) representation is a highly compressed visual representation for a very long video sequence of a dynamic 3D scene. It fits into the MPEG-4 standard [13], in which a scene is described as a composition of several Video Object(VOs), encoded separately. A CB3M standard [13], in which a scene is described as a composition of several Video Object(VOs), encoded separately. A CB3M representation is a set of VO primitives (patches) that are defined as CB3M = \{VO\_i, i =1, …, N\} = \{ (c_i, b_i, n_i, m_i), i =1, …, N\} where (1) N is the number of natural regions; (2) c_i is the color (3 bytes) of the ith region (before region merging); (3) b_i is the 2D boundary in the left mosaic of the ith region chain coded as b_i = \{(x_0, y_0), b_1, b_2, …, b_K\}, where the starting point (x_0,y_0) has 4 bytes, each chain code has 3 bits, K is the number of boundary points and K = Σ K_i is the total for all regions; (4) n_i = (n_{x}, n_{y}, n_{z}, d) represents the plane parameters of the region in 3D; and (5) 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 23N+3K/8+N_{x}4L bytes when each of the motion and structure parameters needs 4 bytes and the number of moving regions is N_{x}(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 a real image sequence of a campus scene, 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. 1) is 4448*1616*2, which has 41MB (without compression and with more than half empty space due to 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 N = 20,636 and K = 1,009,247. The total amount of data is 888 KB (with a header but without coding the motion), and is 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!

Experimental results will be given for the CB3M representation construction and video demos will be shown during the presentation.

**Acknowledgements:** This work is supported by the Air Force Research Lab under the RASER program (Award No. FA8650-05-1-1853) and an AFRL/HECB Grant (No. F33615-03-1-63-83), 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).

**References**


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


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


