Mosaics from Arbitrary Stereo Video Sequences

Nicolas Gorges[1], Marc Hanheide[1], William Christmas[2], Christian Bauckhage[1], Gerhard Sagerer[1], and Joseph Kittler[2]

[1] Bielefeld University, P.O. Box 100131, 33501 Bielefeld, Germany
{ngorges, mhanheid, cbauckha, sagerer}@techfak.uni-bielefeld.de
[2] University of Surrey, Guildford GU2 7XH, UK
{W.Christmas, J.Kittler}@eim.surrey.ac.uk

Abstract. Although mosaics are well established as a compact and non-redundant representation of image sequences, their application still suffers from restrictions of the camera motion or has to deal with parallax errors. We present an approach that allows construction of mosaics from arbitrary motion of a head-mounted camera pair. As there are no parallax errors when creating mosaics from planar objects, our approach first decomposes the scene into planar sub-scenes from stereo vision and creates a mosaic for each plane individually. The power of the presented mosaicing technique is evaluated in an office scenario, including the analysis of the parallax error.

1 Introduction and Motivation

Mosaicing techniques are recently used in various different applications, even though the common basis is always to represent a sequence of images of a given scene in one image. Thus, mosaicing provides a compact, non-redundant representation of visual information. Besides the compression benefits from avoiding redundancy in mosaics, the larger field of view of the integrated mosaic image serves as a better representation of the scene than the single image data, for instance for object recognition or scene interpretation.

But recent mosaicing techniques have restrictions. The main problem for building a mosaic of a non-planar scene is the occurrence of parallax effects as soon as the camera is moving arbitrarily. Parallax describes the relative displacement of an object as seen from different point of views. Each plane of the scene will move in a different relative speed in respect to each other and cause overlaps as soon as the camera center is moved. Therefore, the construction of only a single mosaic of the scene will not succeed. An avenue to deal with this problem is to control the motion of the camera and restrict it to rotation and zooming or compute mosaics on the basis of adaptive manifolds. Another possibility is to apply the mosaicing on (approximately) planar sub-scenes, which is the central assumption for the technique presented in this paper.

The mosaicing system provides visual information in terms of a *pictorial memory* as part of a cognitive vision system (CVS) which is applied in an office scenario[2]. This memory contains a compact visual representation of the scene.

The CVS uses two head-mounted cameras to access the visual outcome of the scene and a stereo video display for augmented reality[13], depicted in Fig. 1.

As the stereo camera-pair is located at the user's head, there is no control on the motion of the cameras. Thus, due to the parallax problem, it is not possible to create just one mosaic of the whole scene, but on almost planar parts. This restriction appears acceptable in an office environment since most of the objects (e.g. tables, walls,...) appear to have an rather planar nature. Therefore, we propose to decompose the scene into planes and than built a mosaic for each plane individually. The resulting set of mosaics provides the needed compact representation of the office scene.

Fig. 1. The setup

This motivation leads to the two central aspects of our mosaicing system. First, a decomposition of the scene into planar sub-scenes has to be computed from stereo information, as explained in detail in Sec. 3.1. Second, the planes have to be tracked during the sequence and for each of the detected planes separate mosaics are created by registering them to a reference frame. How this is done is described in Sec. 3.2. Results from image sequences obtained in the office scenario are discussed in Sec. 4.

2 Related Work

A lot of research has been done on applications of Mosaicing [9, 6] and improving their performance [14, 11, 10]. These approaches mainly focused on the conventional mosaicing method rather than on the restrictions. Most of these are linked with the occurrence of parallax effects. Approaches to make mosaicing invariant to any restrictions attempt to avoid parallax or use parallax explicitly. In order to overcome the restrictions for mosaicing, mosaics with parallax and layers with parallax [7] were introduced. In this case, additional information about the 3D structure is stored to take account of parallax and to make the construction of mosaic images more robust. Another approach [12] tries to present mosaicing as a progress of collecting strips to overcome most restrictions. The strip collection copes with the effects of parallax by generating dense intermediate views, but is still restricted to controlled translational parts in the motion of the camera.

Baker et al.[1] describe an approach to represent a scene as a collection of planar layers calculated from depth maps. But in contrast to our algorithm, the focus is mainly on approximating the 3D structure of the scene than on mosaics.

3 Mosaics of planar sub-scenes

Constructing mosaics from image sequences consists of computing a transformation from the coordinates of the current image to a reference system, warping the current image to the reference frame and integrating new pixel data into the mosaic. The warping function can easily be computed if the images were

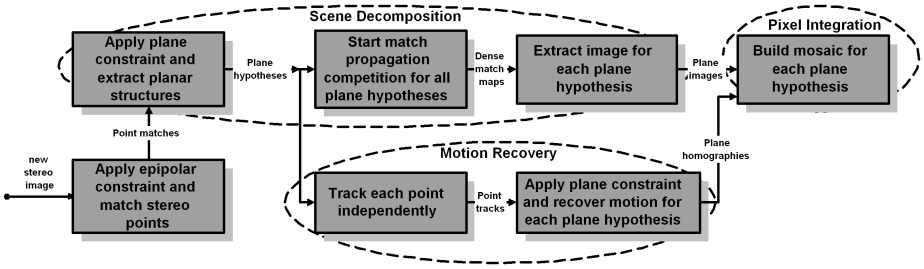


Fig. 2. System overview

acquired by a camera rotating about its fixed center or if the scene is planar. Under these restrictions, however, mosaicing is not suitable for all applications.

For the more general case where the scene is not completely planar and the camera center is moving, a single transformation will not exist. But if the scene is partial planar, there will be several warping functions each of them relating different views of corresponding planar regions. This motivates to build not one but several mosaics: one for each planar sub-scene. Given stereo image data, mosaicing then becomes a three step procedure:

1. **Scene Decomposition:** Segment the current stereo image pair into pixel regions depicting coplanar areas of the scene.
2. **Plane Motion Recovery:** Recover motion of planar regions in order to calculate warping functions.
3. **Planar Mosaic Construction:** Expand mosaics and integrate warped planar regions.

Fig. 2 gives an overview of this concept and the computational modules of the framework introduced here. Next, this framework shall be presented in detail.

3.1 Scene Decomposition

Since stereo data is available due to the design of the used AR gear, identifying planes in a scene is accomplished by means of the following four steps:

1. **Local Coplanar Grouping:** Starting with extracted key points from a pair of images (e.g. by using the Harris detector [4]) and computing their correspondences using epipolar geometry, a plane hypothesis is represented by a *local* group of point matches forming a planar patch.
2. **Coplanar Grouping - Extension of local patch:** Point matches outside the local patch are added to the plane if they satisfy the plane model.
3. **Constrained Plane Propagation:** From a set of point matches, the plane is now extended to pixel regions which satisfy the plane model. The result is a dense match map of a plane which displays textured regions of the plane.
4. **Second plane propagation - A map of the plane:** Finally regions with less texture are assigned to the next neighboring textured region. The result is a boolean map which tells whether a pixel is part of the plane or not. Conjuncting this map with the current image of the scene, yields a pixel representation of the plane which is suitable for mosaicing.

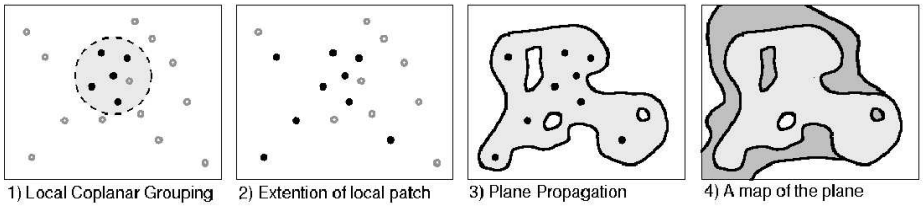


Fig. 3. Identification of pixels belonging to a plane using matched stereo-key-points.

Fig. 3 illustrates this method. It shows the evolution of a single plane given a set of key points. In the first and second step, the plane is represented by a set of point matches. Black points indicate inlier points while pale points represent outliers. Note that the first two steps make use of a feature-based representation while the final steps result in an image-based representation of a plane.

The feature-based steps of this method were introduced for the following reason: It is known that two images embedded in the same plane π are related by a 2D projective transformation (homography) H and that a homography is uniquely defined by four point matches (cf. [5]). However, after extracting key points from stereo images, any four matched points will define a plane. An important issue is thus to distinguish *virtual planes* from physical ones, which is done as follows:

A *plane hypothesis* is defined as a pair (M_i, H_i) where M_i is a set of point matches and H_i a corresponding homography representing the plane model. The set of all point matches is denoted as M. The *dominant plane* $\pi_{dominant}$ of a scene is defined as the plane hypothesis which incorporates the largest amount of point correspondences, i.e.

$$\pi_{dominant} = \underset{\pi_i}{\operatorname{argmax}} ||M_i||.$$

Plane candidates π_i are found by coplanar grouping of point matches using RANSAC [3]. By choosing the actually dominant plane hypothesis $\pi_{dominant}$ and removing its point matches from M, we try to find the next dominant plane of the scene similarly until no new planes can be found or the maximum number of planes is reached. The result is a rough scene decomposition represented by a set of plane hypotheses.

In order to avoid the extraction of virtual planes, we apply a *local planarity constraint*. By restricting the choice of the four points to random local image areas and fitting plane hypotheses to this patches, it is granted that extracted planes are at least locally planar. Then, local plane hypotheses are evaluated with respect to the total number of key points. The hypothesis that accords with most global matches is chosen for further processings. Fitting planes to local patches also allows to measure the *planarity* of planes: if the relation of outlier to inlier points is below a certain threshold hypotheses are rejected.

Since planar surfaces in a scene may contain holes and as there might be regions in the scene for which we do not have enough information to assign them

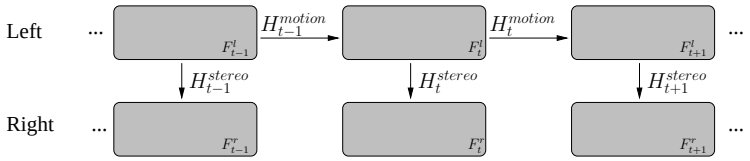


Fig. 4. Homographies between the stereo frame sequence of a plane

to a plane, we apply a pixel-based plane growing method to embed local discontinuities. Based on the algorithm described in [8], we suggest an image-based propagation process which densifies the plane hypotheses. This resembles classical region growing methods for image segmentation. Instead of a homogeneity criterion, normalized cross correlation between point matches is used for region expansion. Starting from a set of matches with high textureness, the algorithm densifies the matches to regions with less textureness. Expansion stops in regions which diverge from the reference homography or have no texture. This restricts the propagation to regions which can be approximated by the plane hypothesis.

So far, only the propagation of a single plane has been considered. Given a set of plane hypotheses, the idea is to start a competition between these planes. Therefore, each plane hypothesis is also associated with the best correlation score among all its point matches. Then, only the plane π_i with the best point match $p_{best}(a, b)$ is allowed to start a single propagation step. Therefore, the neighborhood $N(a, b)$ of point match p_{best} is densified. The chosen plane provides its next best point match and the next iteration begins. The propagation stops if none of the planes has a point match left to be processed.

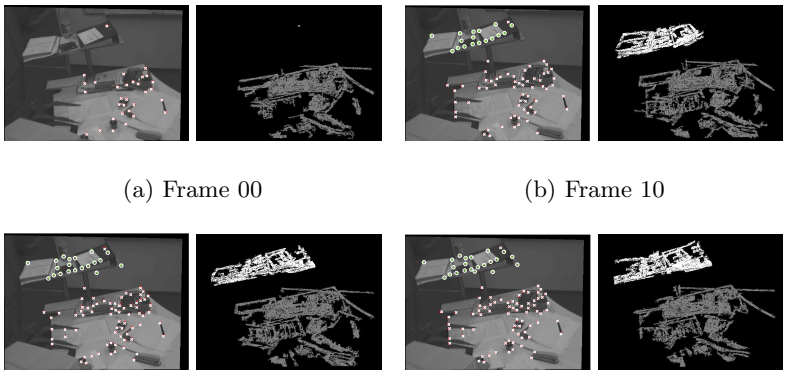
3.2 Plane Motion Recovery and Planar Mosaic Construction

For the construction of a planar mosaic the homographies H^{motion} between the different frames has to be computed to recover the motion of the camera. The motion of the feature points, that have been established in the decomposition stage, are also used to compute these homographies. Thus, the motion recovery performed for each plane can be divided into two steps:

1. **Tracking of plane points:** Given a set of points on a plane, each point is tracked independently. Assuming that a point is moving with constant velocity, a linear first order prediction of a point is used.
2. **Recovering plane motion:** The resulting point tracks $T_t = (p^i_{t-1}, p^i_t)$ are supposed to lie on the same plane. For two views of a plane, there exists a homography H^{motion}_{t-1} (see Fig. 4) which relates p^i_{t-1} to p^i_t. Again, RANSAC is used for a robust estimation of this homography.

Furthermore, the tracked plane has to be updated in terms of integrating new points and removing the ones gone out of sight. Therefore the homography H^{stereo}_t is recomputed and new points are added if they fulfill the planarity constraint.

Based on the interframe homographies H^{motion}_t all plane images are warped to the reference frame F^l_1 of the mosaic. The integration computes the median of the warped frames to determine the value of the resulting mosaic pixel.



(a) Frame 00 (b) Frame 10

(c) Frame 20 (d) Frame 30

Fig. 5. Decomposition of the scene in two planes: Each left images displays the tracked feature points. Respectively, on the right, the textured regions of the planes are shown.

(a) Initial frame (b) Plane image (c) Final Mosaic

Fig. 6. An example for occlusion elimination (detail view)

4 Results

The focus of the evaluation is on the quality and the consistency of the mosaics as they are the final result of the presented procedure. The integration of new pixel data into the mosaic strongly depends on the preprocessing steps, namely *Scene Decomposition*, and *Plane Tracking*. Especially the scene decomposition plays an important role as plane tracking is based on its results. Errors occurring in this processing step are spread to all the following stages and result in erroneous mosaics.

Fig. 5 presents the result of the scene decomposition of a sequence in the office. The decomposition has been limited to two dominant planes to ease the evaluation. The desk has two planes which both are detected correctly. The tracked feature points are highlighted in each frame (left images) and the propagated planes are shown in different gray shadings (right images). Note, that in frame 00 only one plane is detected, but ten frames later further points and another plane is added and tracked from now on. Another positive effect of only integrating image parts that belong to the same plane into the mosaics is depicted in Fig. 6. Because the parcel in the foreground of the scene does not belong to the same plane as the table with the journal, it is omitted from the

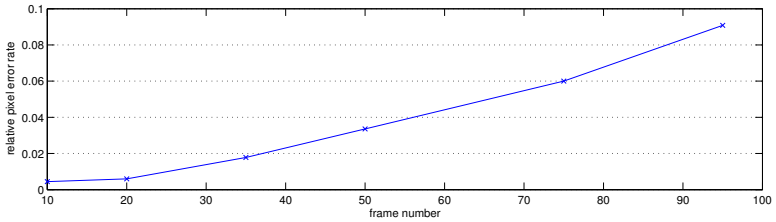


Fig. 7. Evolution of the relative parallax error over a sequence

(a) Frame 10

(b) Frame 95

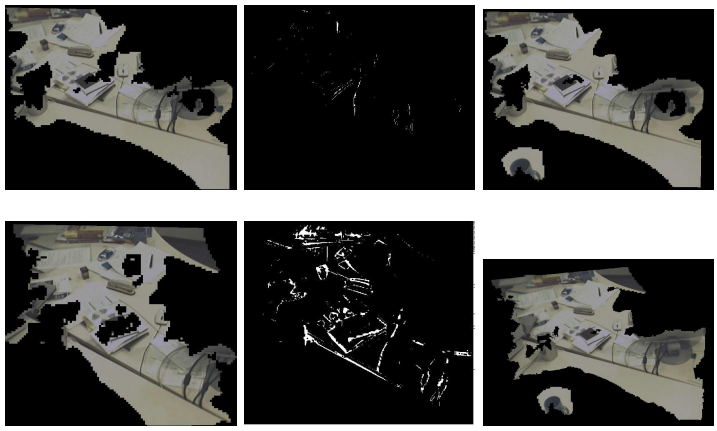


Fig. 8. The parallax error (center) is computed as difference between the single image of the tracked plane (left) and the warped mosaic (right). Errors appear as white dots.

mosaic and the occlusion is eliminated. This allows to create complete views of partially occluded objects in the scene.

If the decomposition of the scene into planes would be perfect, one would expect no parallax error in the mosaic. But due to the just approximately planar nature of extracted sub-scenes, errors will occur, especially at the borders of flat objects (e.g. a flat book lying on the table) as well as at the edges of extracted planes. We calculated the *relative parallax error* $\epsilon = \delta/s$ to evaluate these effects, which is defined as the amount of pixel differences δ the tracked plane of the frame and the so far integrated mosaic, normalized by the size s of the plane measured in pixels. For calculating that difference the mosaic is warped into the actual frame. In Fig. 7 the evolution of this error measure is plotted over the whole sequence which is partially shown in Fig. 8. As expected, the parallax error rate increased while the mosaic is growing, but even in the last frame 95, errors only occur at the edges of the objects, as can be seen in the center image of Fig. 8(b). The computation of the mosaics (tracking, and updating the homographies) can be performed in real-time after the initialization or the update respectively of the planes is done.

5 Conclusion

We presented an unique approach to create mosaics for arbitrarily moving head-mounted cameras. The three stage architecture first decomposes the scene into approximated planes using stereo information, which afterwards can be tracked and integrated to mosaics individually. This avoids the problem of parallax errors usually occurring from arbitrary motion and provides a compact and non-redundant representation of the scene. Furthermore, creating mosaics of the plane allows to eliminate occlusion, since objects blocking the sight on a plane are not integrated. This can for instance help object recognition systems and further scene interpretation in the Cognitive Vision System, this approach is part of. The proposed robust decomposition and tracking algorithms allow to apply the system in real office scene with common cameras.

Acknowledgements

This work was partly funded by VAMPIRE (IST-2001-34401) and ECVision.

References

1. S. Baker, R. Szeliski, and P. Anandan. A layered approach to stereo reconstruction. In *Proc. CVPR*, pages 434–441, 1998.
2. C. Bauckhage, M. Hanheide, S. Wrede, and G. Sagerer. A Cognitive Vision System for Action Recognition in Office Environments. In *Proc. CVPR*, 2004. to appear.
3. M.A. Fischler and R.C. Bolles. Random sample consensus: A paradigm for model fitting with applications to image analysis and automated cartography. *Communications of the ACM*, 24(6):381–395, 1981.
4. C. Harris and M. Stephens. A combined corner and edge detector. In *Proc. Alvey Vision Conference*, pages 147–151, 1988.
5. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge University Press, 2001.
6. B. Hu and C. Brown. Interactive indoor scene reconstruction from image mosaics using cuboid structure. In *Proc. Motion & Video Computing*, pages 208–213, 2002.
7. R. Kumar, P. Anandan, M. Irani, J. Bergen, and K. Hanna. Representation of scenes from collections of images. In *Proc. IEEE Workshop on Representations of Visual Scenes*, pages 10–17, 1995.
8. M. Lhuillier and L. Quan. Robust dense matching using local and global geometric constraints. In *Proc. ICPR*, pages 968–972, 2000.
9. P. Anandan M. Irani and S. Hsu. Mosaic based representations of video sequences and their applications. In *Proc. ICCV*, 1995.
10. B. Möller, D. Williams, and S. Posch. Robust image sequence mosaicing. In *Proc. of 25th DAGM Symposium*, LNCS 2781, pages 386–293, 2003.
11. S. Peleg and J. Herman. Panoramic mosaics by manifold projection. In *Proc. ICPR*, pages 338–343, 1997.
12. S. Peleg, B. Rousso, A. Rav-Acha, and A. Zomet. Mosaicing on adaptive manifolds. *IEEE PAMI*, 22:1144–1154, 2000.
13. H. Siegl, M. Brandner, H. Ganster, P. Lang, A. Pinz, M. Ribo, and C. Stock. A mobile augmented reality system. In *Exhibition Abstracts of Int. Conf. Computer Vision Ssytems*, pages 13–14, 2003.
14. I. Zoghlami, O. Faugeras, and R. Deriche. Using geometric corners to build a 2d mosaic from a set of images. In *Proc. CVPR*, pages 420–425, 1997.