Compositing a bird's eye view mosaic

Robert Laganiere
School of Information Technology and Engineering
University of Ottawa
Ottawa, Ont K1N 6N5

Abstract

This paper describes a method that allows the composition of an overhead view mosaic of a worksite from the views available. This global view of the site will help a teleoperator to adequately specify the actions to be performed in the realization of a given task. The idea proposed here is to obtain the overhead view by combining two known techniques, namely mosaicing and image rectification.

1 Introduction

In applications like mining, excavation, forestry, human operators have to perform difficult tasks in harsh worksite with large and slow-moving mechanical units. Telerobotics becomes then an attractive solution that allows to remove the human operator from the worksite, leading to a safer and more efficient operation. In such a teleoperation framework, there will be usually several cameras located at the worksite that provide the user with visual information. It is expected that from the multiple views available of the worksite, the teleoperator will be able to make decisions about the actions that must be taken in the execution of a given task. However, the integration of this multiple source of visual information in a real context can be difficult to perform. The human operator must indeed be able to build a mental model of the remote site from the different images available.

The goal of this paper is to describe a solution that consists in the composition of a bird's eye view mosaic of a worksite from the available views. This global view will serve as a virtual map of the site used to help the user to adequately specify the commands to be sent to the vehicle under his control. The idea proposed here is to obtain this overhead view by combining two known techniques, namely *mosaicing* and *image rectification*.

Since the overhead view is obtained from a set of images where each image contains a limited portion of the site, these must be assembled together to form a larger composition. Mosaicing is the technique that allows the production of a large image from smaller ones as long as they can be related to each other by a global mapping [1]. In the case of a camera rotated about its center, this mapping can be based on affinity, quadratic [2] or homographic transformations [3]. It becomes then possible to compose a panoramic mosaic by registering each image with respect to a given reference frame. The parameters of the transformation modeling the mapping between views are estimated by a feature-based approach [4], or using direct minimization of pixel intensity difference [5].

Image rectification consists in the reprojection of an image onto a new projection plane. This technique has been extensively used in many area of computer vision such as stereo matching or image interpolation. In our work, the interesting case is when a plane is the observed structure. The image of this plane can then be rectified by a projective warping to one that corresponds to a fronto-parallel view of the plane under observation. When the world coordinate of 4 points on the plane are known, then it becomes possible to make Euclidean measurement [6]. Plane image rectification has been successfully used in several works, in areas such as 3D reconstruction [7][8], metrology [9] or augmented reality [10].

However, the present context under which these techniques are used introduces peculiar problems. Indeed a strategy to assemble the global view must be devised and the way these images will be combined together must also be determined. Finally, the methods to obtain the overhead view transformation and the inter image have to be defined.

The rest of this paper is organized as follows: Section 2 reviews the basic projective relations, Section 3 presents the image plane to world plane transformation and Section 4 explains the procedure to follow in the composition of the bird's eye view mosaic. Section 5 is a conclusion.

2 The projective model

Under the pinhole model, the camera performs a perspective projection of a 3D point onto an image point located on a retinal plane. Using homogeneous coordinates, the projective relation between a 3D point $\mathbf{P} = [X, Y, Z, 1]^T$ and its image $\mathbf{p} = [x, y, 1]^T$ can be expressed as:

$$\mathbf{p} = \mathbf{MP} \tag{1}$$

The 3×4 matrix $\mathbf{M}$ is the projection matrix. It relates world points to image points according to the camera location with respect to the reference frame, represented by a rotation matrix $\mathbf{R}$ and a translation vector $\mathbf{T}$. The intrinsic parameters of the camera are represented by the following matrix:

$$\mathbf{C} = \begin{bmatrix} f_x & 0 & c_x \\ 0 & f_y & c_y \\ 0 & 0 & 1 \end{bmatrix} \tag{2}$$

where f_x and f_y correspond to the focal length respectively measured in horizontal and vertical pixel units. The position (u_0, v_0) is the principal point where the optical axis pierces the image plane. This model assume an orthogonal grid. With these definitions, the projection matrix can be written as:

$$\mathbf{M} = \mathbf{C} \begin{bmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{bmatrix} \begin{bmatrix} \mathbf{R} & \mathbf{T} \\ \mathbf{0}_3^T & 1 \end{bmatrix} \tag{3}$$

3 The image plane to world plane transformation

When the structure under observation is a plane, a simpler formulation becomes available. Since the world coordinate system can be set anywhere, it can be conveniently positioned on the plane, such that this latter has zero Y coordinate. This choice reduces the projection matrix to:

$$\begin{aligned} \begin{bmatrix} x \\ y \\ 1 \end{bmatrix} &= \begin{bmatrix} m_{11} & m_{12} & m_{13} & m_{14} \\ m_{21} & m_{22} & m_{23} & m_{24} \\ m_{31} & m_{32} & m_{33} & m_{34} \end{bmatrix} \begin{bmatrix} X \\ 0 \\ Z \\ 1 \end{bmatrix} \\ &= \begin{bmatrix} m_{11} & m_{13} & m_{14} \\ m_{21} & m_{23} & m_{24} \\ m_{31} & m_{33} & m_{34} \end{bmatrix} \begin{bmatrix} X \\ Z \\ 1 \end{bmatrix} \\ &= \mathbf{H} \begin{bmatrix} X \\ Z \\ 1 \end{bmatrix} \end{aligned} \tag{4}$$

This 3×3 matrix is a homography $\mathbf{H}$ that represents the projective relation between the world plane and the corresponding image point. Again the equality above is up to a scale factor and consequently the homography matrix has 8 degrees of freedom. Since this matrix can be inverted, it follows that if $\mathbf{H}$ is completely known, the world coordinates of any image point on the plane can be recovered. Without loss of generality, we can assume that the optical axis of the camera is aligned with the Z axis. The location of the camera can therefore be specified by its height h and its tilt and swing angles respectively θ and ϕ. The rotation matrix is

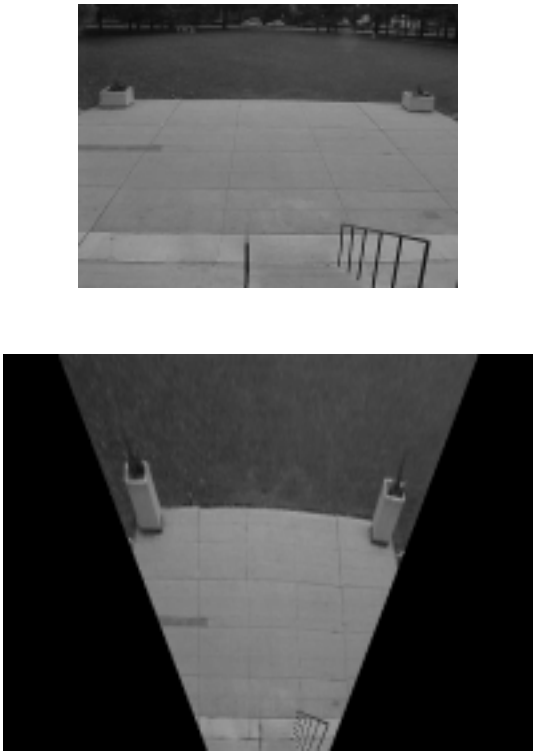

Figure 1: The overhead view obtained from an image plane to world plane homography transformation.

therefore in this case:

$$\mathbf{R} = \begin{bmatrix} \cos\phi & \sin\phi & 0 \\ -\sin\phi & \cos\phi & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} 1 & 0 & 0 \\ 0 & \cos\theta & \sin\theta \\ 0 & -\sin\theta & \cos\theta \end{bmatrix} \tag{5}$$

This leads us to the following homography:

$$\mathbf{H} = \mathbf{C} \begin{bmatrix} \cos\phi & \sin\phi\sin\theta & h\sin\phi\sin\theta \\ -\sin\phi & \cos\phi\sin\theta & -h\cos\phi\sin\theta \\ 0 & \cos\theta & h\cos\theta \end{bmatrix} \tag{6}$$

This means that if the position of a calibrated camera with respect to the plane is known, then one can reconstructed the overhead view of the plane. This is the approach we used. In our teleoperation framework, we assume that the cameras have been calibrated and that their tilt and swing angles as well as their height with respect to the observed plane are all known. Note that if no metric information is needed, the height of the camera can be arbitrarily fixed to 1. Figure 1 shows the overhead view obtained from one view of a site as given by the homographic transformation (6) with $\theta = 35^o$ and $\phi = 0^o$.

It is important to note that the transformation (6) is valid only for points lying on the observed plane. Other objects

visible in the image will therefore be distorted by the reprojection; this is a consequence of the planar parallax geometry [11]. Also, the portion of the image located above the vanishing line of the plane must not be considered when generating the overhead view since these points above the line of horizon do not belong to the plane. The equation of this line can easily be obtained from the homography matrix by noting that points on the vanishing line correspond to world points located at infinity (represented by a homogeneous vector having a 0 coordinate for its last element). Therefore, the set of image points belonging to the line of horizon are the ones that satisfy the following equation:

$$\mathbf{p} = \mathbf{H} \begin{bmatrix} X \\ Z \\ 0 \end{bmatrix} \tag{7}$$

It follows that the last row of the homography matrix gives the equation of the vanishing line of the corresponding plane. Any image must therefore be rectified only in the portion that lies under this line.

Finally, it should be mentioned that in order to estimate the image to world plane homography one can use known information about the visible patterns on the plane (such as parallel lines or known angles between lines). This is the solution presented in [12]. We do not retain this solution for the proposed application, since the site of operations is an unstructured environment (for example an open-pit mine).

4 Compositing the mosaic

In a large environment, a single image will only capture a small portion of the worksite. Several images taken from different points of view, possibly using several cameras will therefore be required. In order to obtain a global view of the site, all these images must be assembled together to form the bird's eye view mosaic.

It will be possible to compose such a mosaic if the images to be assembled can be related to each other through a global mapping. It directly follows from the homographic transformation (6) that the same kind of relation also exists between the two images of a given plane. In this latter case, the homography allow to map an image point $\mathbf{x}^i$ in view i to its corresponding point $\mathbf{x}^j$ in image j, i.e.:

$$\mathbf{x}^j = \mathbf{H}_{ij}\mathbf{x}^i \tag{8}$$

One of the images at our disposal will therefore serves as the reference frame. This image will be rectified using the homography $\mathbf{H}_{0W}$as computed from the available camera parameters. The other images will then be aligned to this overhead image through the estimated homographies between these images and the reference frame, e.g. $\mathbf{H}_{10}$. Each image of the plane is therefore reprojected on its fronto-parallel view using the homography $\mathbf{H}_{iW} = \mathbf{H}_{0W}\mathbf{H}_{i0}$. If for a given view, it is not possible to estimate the homography transformation with respect to the reference frame, then the $\mathbf{H}_{iW}$ mapping will simply be obtain by the appropriate concatening of the computed homographies.

4.1 Homography estimation

As we have seen, it becomes possible to compose an overhead mosaic once the homographic transformation between the available views have been estimated. We used here a feature based approach. When several image points of a plane have been detected and matched, the value of the elements of $\mathbf{H}$ can be computed. For each pair of corresponding points, it is possible to extract two independent linear equations from (8), by rewriting it as:

$$\mathbf{x}' \times \mathbf{H}\mathbf{x} = 0 \tag{9}$$

Since the matrix $\mathbf{H}$ has 8 DOF, the element h_{33} can be arbitrarily sets to 1, giving 8 unknowns. Consequently, a total of 4 point correspondences is required to determine $\mathbf{H}$.

Obviously, in a context of non-perfect data, many more points should be used. Then, $\mathbf{H}$ would be estimated by a minimization scheme. This is usually done by defining the 9×1 vector $\mathbf{h}$ made of the 9 unknown elements of matrix $\mathbf{H}$. With N point correspondences, it is possible to extract $2N$ linear constraints from (9). This results in a system of the form:

$$\mathbf{B}\mathbf{h} = 0 \tag{10}$$

We then have to solve the following problem

$$min_{\mathbf{h}}||\mathbf{B}\mathbf{h}||^2 \; subject\, to \; ||\mathbf{h}|| = 1 \tag{11}$$

The solution is then the eigenvector of matrix $\mathbf{B}^\mathbf{T}\mathbf{B}$ that corresponds to the smallest eigenvalue [13]. To obtain a more stable linear system, the coordinates of the point correspondences are normalized, as explained in [14], where a similar method is used to compute the fundamental matrix.

4.2 Combining the overhead views

When compositing the overhead view, one must determine the way the overlapping portions of the reprojected images should be combined. This texture mapping is achieved using the world to image plane mapping so that a pixel in the destination mosaic is back-projected to its corresponding point in the source image. The color value is then obtained by bi-linear interpolation. In the case of panoramic mosaic, a blending region is usually defined in order to avoid the step change in intensity along the joint between two images due to camera automatic gain control or to differences in illumination conditions. A blending function is therefore defined which will weight the intensity of the pixels from each source images to ensure a smooth (and invisible) image transition [15].

In our case, since the mosaic is composed from rectified images, the pixels from which the color values will be read are, in fact, irregularly sampled. Therefore, when the value of a pixel in the destination mosaic must be determined, all rectified images for which the corresponding point is visible are considered. Among these images, the one that has the best resolution at the current point will be selected.

In order to determine which image has the best resolution at a given point, we will defined what we called the *instantaneous sampling rate*. The homography transformation that maps a world plane $[X, Y, 1]^T$ to an image point $[x, y, 1]^T$ according to a given source image can be rewritten as:

$$X = h_X(x,y) = \frac{h_{11}x + h_{12}y + h_{13}}{h_{31}x + h_{32}y + h_{33}} \quad (12)$$

$$Y = h_Y(x,y) = \frac{h_{21}x + h_{22}y + h_{23}}{h_{31}x + h_{32}y + h_{33}} \quad (13)$$

The sampling rate in the X direction at this point is simply the number of source image pixels between the point $[X, Y, 1]^T$ and the point $[X + \Delta X, Y, 1]^T$ divided by the distance between these two points. It then follows that the instantaneous horizontal sampling rate is given by:

$$s_X = \sqrt{\left(\frac{\delta h_X(x,y)}{\delta x}\right)^2 + \left(\frac{\delta h_X(x,y)}{\delta y}\right)^2} \quad (14)$$

and similarly for the instantaneous vertical sampling rate:

$$s_Y = \sqrt{\left(\frac{\delta h_Y(x,y)}{\delta x}\right)^2 + \left(\frac{\delta h_Y(x,y)}{\delta y}\right)^2} \quad (15)$$

From these definitions, we simply defined the instantaneous sampling density as:

$$s = s_X s_Y \quad (16)$$

and the image having the highest sampling density at the current point will therefore be selected as source image.

Figure 2 shows the result obtained in the composition of a bird's eye view mosaic using 3 images. The center view is the reference frame, its corresponding overhead view transformation being the one shown in Figure 1. The left to center and right to center homographies have been used to combine the views.

5 Conclusion

A method that allows the composition of a bird's eye view mosaic of a worksite has been proposed in this paper. The method is based on the image rectification of a reference frame to the fronto-parallel view of the observed planar surface. This transformation can be determined from the geometry of the camera system. The other views can then be combined to the reference through the estimation of the homography that relates each views of the plane.

We believe that the resulting virtual map of the site of operations can be usefully integrated in a teleoperation framework where the user will be able to remotely specify actions to be taken in the execution of a given task. The bird's eye view mosaic will indeed provide the user with an integrated global view of the site and will give him a mean by which the command to be sent can be more accurately specified.

Acknowledgments

The author wishes to thank A.-L. Henry and H. Hajjdiab for their help in the software implementation of this work. This work was supported in part by the IRIS Network of Centres of Excellence and by the Natural Science and Engineering Research Council of Canada.

References

[1] L.G. Brown, A Survey of Image Registration Techniques, *ACM Computing Surveys*, 24:4, 1992.

[2] R. Kumar, P. Anandan, M. Irani, J. Bergen, K. Hanna, Representation of Scenes from Collections of Images, *ICCV Workshop on the Representation of Visual Scenes*, 1995.

[3] D. Capel, A. Zisserman, Automated Mosaicing with Super-resolution Zoom, *Proc. of CVPR*, 885-891, 1998.

[4] T.-J. Cham, R. Cipolla, A Statistical Framework for Long-Range Feature Matching in Uncalibrated Image Mosaicing, *Proc. of CVPR*, 443-447, 1998.

[5] M. Irani, P. Anandan, S. Hsu, Mosaic based Representations of Video Sequences and their Applications, *Proc. of ICCV*, 605-611, 1995.

[6] A. Criminisi, I. Reid, A. Zisserman, A Plane Measuring Device, *Image and Vision Computing*, Vol. 17, 625-634, 1999.

[7] D. Liebowitz, A. Criminisi, A. Zisserman, Creating Architectural Models from Images, *Proc. EuroGraphics*, 1999.

[8] C. Baillard, A. Zisserman, Automatic Reconstruction of Piecewise Planar Models from Multiple Views, *Proc. of CVPR*, 559-565, 1999.

[9] A. Criminisi, I. Reid, A. Zisserman, Single View Metrology, *Proc. of ICCV*, 434-442, 1999.

[10] M. Jethwa, A. Zisserman, A. Fitzgibbon, Real-time Panoramic Mosaics and Augmented Reality, *Proc. of BMV*, 1999.

[11] A. Criminisi, I. Reid, A. Zisserman, Duality, Rigidity and Planar Parallax, *Proc. of ECCV*, 846-861, 1998.

[12] D. Liebowitz, A. Zisserman, Metric Rectification for Perspective Images of Planes, *Proc. of CVPR*, 1998.

[13] M. Lourakis, S. Orphanoudakis, Visual Detection of Obstacles Assuming a Locally Planar Ground, *Proc. 3rd Asian Conference on Computer Vision*, Hong-Kong, pp. 527-534 1998.

[14] R. Hartley. In Defense of the Eight-Point Algorithm, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol. 19, pp. 133-135, 1997.

[15] S. Peleg, J. Herman, Panoramic Mosaics by manifold projection, *Proc. of CVPR*, 1997.

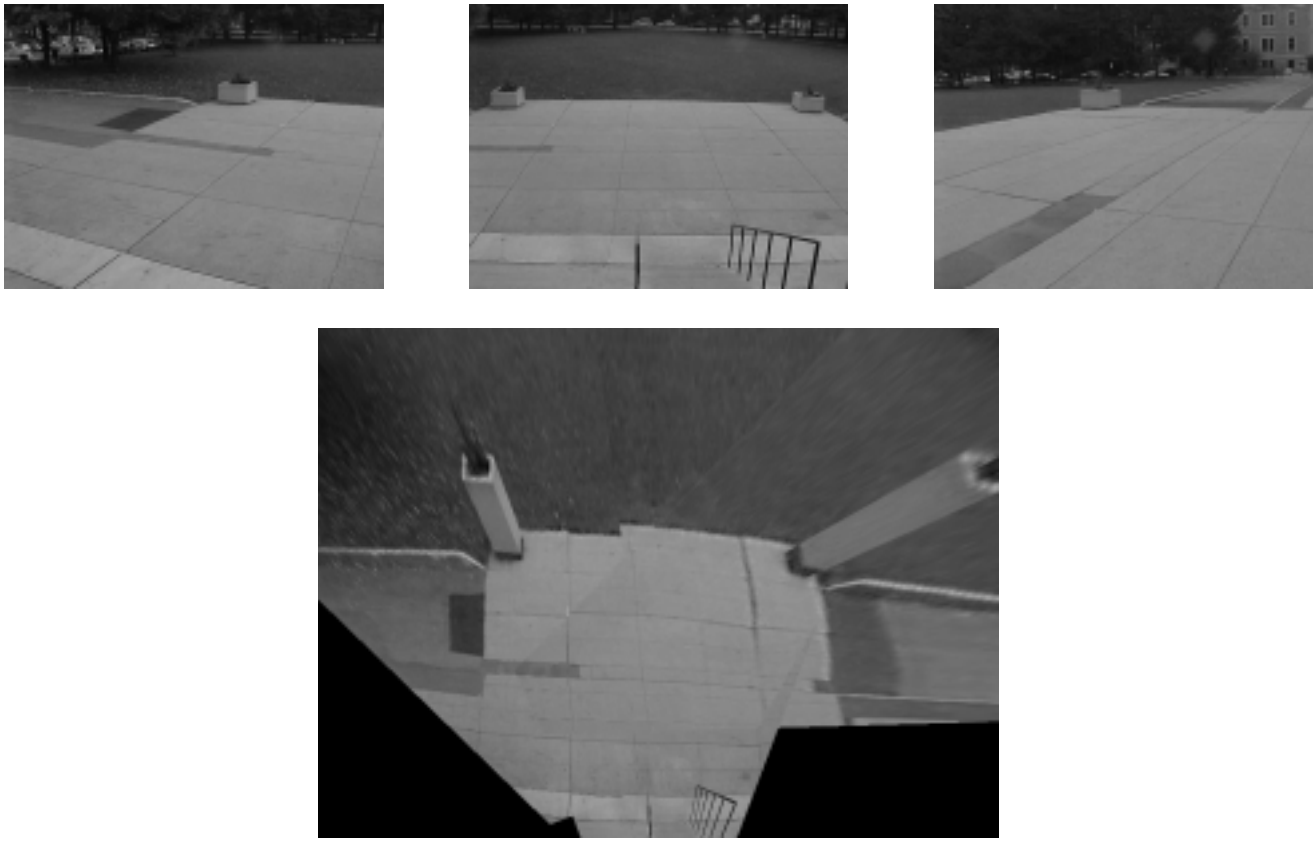

Figure 2: The bird's eye view mosaic composited from the three shown views.