Polyhedral Environment in Stereo Views: Representation and Extraction

Ronald Chung *Andrew Arengo*

Department of Mechanical and Automation Engineering
The Chinese University of Hong Kong, Shatin, Hong Kong
E-mail: rchung@cuhk.edu.hk

Abstract

A representation of polyhedral environment in a stereo pair of images is proposed and its three-dimensional recovery is presented. The polyhedral environment is represented as a number of image-to-image mappings in the form of matrices, each corresponding to a planar surface in the environment. Unlike a mere depth map, such a representation is segmented, in the sense that different surfaces in the environment correspond to different matrices and are separated. A mechanism is proposed to recover the representation even for scene that is not densely featured. Experimental results on typical polyhedral scenes are given.

1 Introduction

Reconstructing three-dimensional (3-D) information of a scene from its images, to the extent that surfaces and objects in it can be separated, is a basic goal of computer vision. Stereo vision is an important and well-studied vision cue for that. Surveys on stereo vision work over the years can be found in [1, 3].

To solve the reconstruction problem, the most important and the only scene-independent clue that stereo vision can exploit is the epipolar geometry. For any feature in one image, there exists in the other image a line named as the epipolar line on which the correspondence of the feature must be located. So long as the spatial relationship of the two cameras is known, the epipolar line is predictable, and the originally 2-D search in an image for stereo correspondence becomes a 1-D search along the corresponding epipolar line. However, even with such a useful constraint, matching features across the stereo images still has ambiguity along the epipolar lines. The classical approach of resolving the ambiguity is to rely on two assumptions: the surface-continuity or the surface-smoothness assumption, and the feature-ordering assumption. The first assumes that the scene is continuous or smooth, and the second assumes that features in the two images follow the same left-to-right order along the epipolar lines.

While the classical approach does give satisfactory results over smooth scenes, it has limited performance towards scenes with occlusions, for the two assumptions are not valid across occlusion boundaries. There have been extensions to the approach in the literature, which typically disable the two assumptions locally at selected places in the scene. However, the performance has been limited, owing to the mere fact that the disabling decisions are local.

In this paper, a representation of a polyhedral environment that is pictured in a stereo pair of images, like the building structures in the aerial views of an urban city or the corridor in the stereo views of inside a building, is proposed. The representation consists of a number of image-to-image mappings in the form of matrices, each corresponding to a planar surface in the environment. Unlike a mere depth map, the representation is segmented, in the sense that different surfaces in the environment correspond to different matrices and are separated. The essence of the representation is that it can be recovered through a simple mechanism even for environment that is not densely featured. For scene like a corridor, the representation is recovered not only with sparse depth estimates over the edges of walls and ceiling and floor, but also with the information that which line segments in the images form a wall or ceiling or floor in the environment. The representation can be used not only for depth recovery as required in autonomous navigation, but also for image transfer to arbitrary viewpoints as required in virtual reality applications.

2 Preliminaries on Homography

An image-to-image mapping was introduced recently by Faugeras [2] for three-view problems, [6, 5], namely the reprojection of a scene from two known views to a third view, and the recognition of an object in a third view using two fixed views as the reference. The mapping can be described as the following. All pairs of image positions $(\mathbf{p}_i, \mathbf{p}'_i)$ in two images, so long as they are projected from the same plane Π in 3-D, satisfy

$$\begin{bmatrix} \mathbf{p}'_i \\ 1 \end{bmatrix} \cong \mathbf{H}_\Pi \begin{bmatrix} \mathbf{p}_i \\ 1 \end{bmatrix} \qquad (1)$$

where $\cong$ denotes equality up to a scale, and $\mathbf{H}_\Pi$ is a 3×3 nonzero matrix. $\mathbf{H}_\Pi$ characterizes the correspondences between the images due to the plane Π, and is referred to as the homography (or the homography matrix) [2] induced by Π. It should be noted that epipoles $(\mathbf{e}, \mathbf{e}')$ also satisfy the above equation.

3 Homography-based Stereo

3.1 Theory

In this section it is outlined how homography can also be used in a two-view problem – stereo vision – for resolving correspondence ambiguity. Being a clue of relating image-to-image correspondences, homography can be used in place of surface-continuity and feature-ordering assumptions for resolving correspondence ambiguity, so long as the scene can be approximated as consisting of mainly planar surfaces. As to be explained later, the use of homography also offers many advantages over the use of the two assumptions.

If there is only a single planar surface Π in the scene, all correct stereo correspondences are captured by a constant 3×3 nonzero homography matrix $\mathbf{H}_\Pi$ under Equation (1). Suppose there are altogether P point features of Π which are visible in both images. Combining the equations from all correct pairings, it can be obtained that:

$$\mathbf{M}_\Pi \cdot \mathtt{vec}(\mathbf{H}_\Pi) = \mathbf{0} \qquad (2)$$

where $\mathbf{M}_\Pi$ is

$$\begin{bmatrix} \begin{bmatrix} \mathbf{p}_0 \\ 1 \end{bmatrix}^{\mathrm{T}}, 0,0,0, -\begin{bmatrix} \mathbf{p}_0 \\ 1 \end{bmatrix}^{\mathrm{T}} u'_0 \\ 0,0,0, \begin{bmatrix} \mathbf{p}_0 \\ 1 \end{bmatrix}^{\mathrm{T}}, -\begin{bmatrix} \mathbf{p}_0 \\ 1 \end{bmatrix}^{\mathrm{T}} v'_0 \\ \vdots \\ \begin{bmatrix} \mathbf{p}_i \\ 1 \end{bmatrix}^{\mathrm{T}}, 0,0,0, -\begin{bmatrix} \mathbf{p}_i \\ 1 \end{bmatrix}^{\mathrm{T}} u'_i \\ 0,0,0, \begin{bmatrix} \mathbf{p}_i \\ 1 \end{bmatrix}^{\mathrm{T}}, -\begin{bmatrix} \mathbf{p}_i \\ 1 \end{bmatrix}^{\mathrm{T}} v'_i \\ \vdots \\ \begin{bmatrix} \mathbf{p}_{(P-1)} \\ 1 \end{bmatrix}^{\mathrm{T}}, 0,0,0, -\begin{bmatrix} \mathbf{p}_{(P-1)} \\ 1 \end{bmatrix}^{\mathrm{T}} u'_{(P-1)} \\ 0,0,0, \begin{bmatrix} \mathbf{p}_{(P-1)} \\ 1 \end{bmatrix}^{\mathrm{T}}, -\begin{bmatrix} \mathbf{p}_{(P-1)} \\ 1 \end{bmatrix}^{\mathrm{T}} v'_{(P-1)} \end{bmatrix}$$

and $\mathtt{vec}(\mathbf{H}_\Pi)$ is the 9×1 column vector expanded from $\mathbf{H}_\Pi$. Equation (2) is a homogeneous system of $2P$ linear equations for the 9 unknowns in $\mathbf{H}_\Pi$, and under correct stereo correspondences nontrivial solution of $\mathbf{H}_\Pi$ should exist. The $2P \times 9$ matrix $\mathbf{M}_\Pi$ represents the matching between $\{\mathbf{p}_i\}$ and $\{\mathbf{p}'_i\}$ and is important. Hereafter it is referred to as the *correspondence matrix* for the planar surface Π. Note that the epipoles already constitute one point pair or two row vectors of $\mathbf{M}_\Pi$.

Suppose the total number of features of Π which are visible in both images, P, is such that $2P \geq 9$. To have non-trivial solution of $\mathbf{H}_\Pi$ Equation (2), the rank of $\mathbf{M}_\Pi$ should be such that

$$\mathtt{Rank}(\mathbf{M}_\Pi) < 9 \qquad (3)$$

This is a distinct property of the correct solution to the correspondence problem. In fact if it is known that there is only a single planar surface in the scene, the solution of $\mathbf{H}_\Pi$ in the system of Equation (2) should be unique up to a scaling factor, and the rank of $\mathbf{M}_\Pi$ should be exactly 8. Unless all correspondences are correct, such a rank property is generally not satisfied if $2P \geq 9$, and is unlikely to be satisfied if $2P \gg 9$.

Though inadequate to resolve all correspondence ambiguity, the epipolar constraint plus reasonable bounds of the disparity gradient are often enough to resolve the ambiguity of a few correspondences. One mechanism of solving the stereo correspondence problem is therefore the following. Correspondences unique under the epipolar constraint are first extracted. Should such initial correspondences exceed three point correspondences or the equivalent, they together with the known epipoles allow 8 or more row vectors of $\mathbf{M}_\Pi$ to be available. Such initial row vectors are a subset of the row vectors of $\mathbf{M}_\Pi$, and the matrix they form in the same manner as $\mathbf{M}_\Pi$ is hereafter referred to as the *initial correspondence matrix* ${}^i\mathbf{M}_\Pi$. With ${}^i\mathbf{M}_\Pi$, $\mathbf{H}_\Pi$ can be determined up to a scaling factor from the nullspace of ${}^i\mathbf{M}_\Pi$. Through Equation (1), such an $\mathbf{H}_\Pi$ can then be used to extrapolate all other correspondences due to the planar surface.

The initial correspondences need not be over three points. If the surface boundary is visible partially or entirely, often junctions are found along it, and the stereo correspondence of a single junction with two branches is already equivalent to three point correspondence. A junction with two branches is hereafter referred to as an L-junction. L-junctions are used instead of points in this work, as they are more distinct and more sparse and thus more likely to have unique correspondences under the epipolar constraint.

Yet, when extended to a more realistic case where there are multiple surfaces $\{\Pi\}$ in the scene, the above solution mechanism has a complication. It is not unreasonable to assume three initial point correspondences or one initial L-junction correspondence per surface to be available from the epipolar constraint. However, such initial correspondences over different surfaces are all mixed together, as the surfaces are not segmented in the images. In other words, the initial correspondence matrix ${}^i\mathbf{M}_\Pi$ for individual surface Π in the scene is not explicitly available. Rather, the row vectors of matrices $\{{}^i\mathbf{M}_\Pi\}$ for different surfaces $\{\Pi\}$ are available as row vectors in a single matrix ${}^i\mathbf{M}$, in no particular order. A mechanism is therefore needed to sort out which of the known correspondences or the row vector pairs are from which surfaces, i.e., to segment ${}^i\mathbf{M}$ into ${}^i\mathbf{M}_\Pi$'s due to different surfaces Π's. Below a procedure is described that serves that purpose.

3.2 Initial Homography Estimation

If the initial stereo correspondences are all unrelated, like in the form of individual point correspondences, the homographies contained in them have to be estimated through a subspace clustering process.

However, this work assumes the context of polyhedral scenes like outside or inside building structures, and often L-junctions are found in the images. Each L-junction correspondence is the stereo correspondences of two line segments and a point or equivalently three points, and it is just enough to define with the epipole pair a homography. The problem then becomes how to group the homographies which are defined by the initial L-junction correspondences into different sets, each set being a collection of homographies which are alike enough (meaning that they are the same surface of the environment), and different sets having homographies different enough.

A simple solution to the above is the nearest-neighbor clustering algorithm [4]. The algorithm requires a measure of inter-homography distance and a threshold to decide whether any two homographies should be treated as from the same surface or not. In the implemented system, the inter-homography distance is defined not in the homography space, but in the stereo images directly, so as to put emphasis on the aspect of image-to-image mapping a homography represents. More precisely, to find the distance between two homographies, the image features defining one homography are pushed through the other homography and the mapping errors so resulted are noted. The inter-homography distance is defined as the maximum error between the measured position and the mapped position of image feature. The threshold used in the system is $\sqrt{10}$ pixels. That is, for two homographies to be regarded as from the same surface, the mapped position of an image feature should not be more than $\sqrt{10}$ pixels away from its measured position.

Once the initial homographies are clustered into sets of similar ones, a homography is further defined for each set as a whole using the *least-squares method*. Such homographies are then used to extrapolate feature correspondences in the images, and be confirmed by them if the extrapolations are supported by image measurements.

3.3 System Overview

A polyhedral environment, or any environment made from planar surfaces, when observed through stereo imaging can therefore be represented as a collection of homographies. Such a representation can be recovered using a mechanism as outlined below.

An overview of the recovery mechanism is shown in Figure 1. Line segments are first extracted from the images through edge detection and line fitting processes. L-junctions are then hypothesized from the line segments through a corner detection process. L-junctions which have unique correspondences under the epipolar constraint are identified. The unique correspondences are then supplied to the subspace clustering process described in Section 3.2, which extracts the homographies present in the stereo images. For any planar surface in the scene, as long as one L-junction correspondence over it is initially available, the associated homography can be estimated. With more than one initial L-junction correspondence, the homography is even confirmed.

Through Equation (1), the homographies can then serve as mappings to extrapolate correspondences of all other features in the two images. For any feature

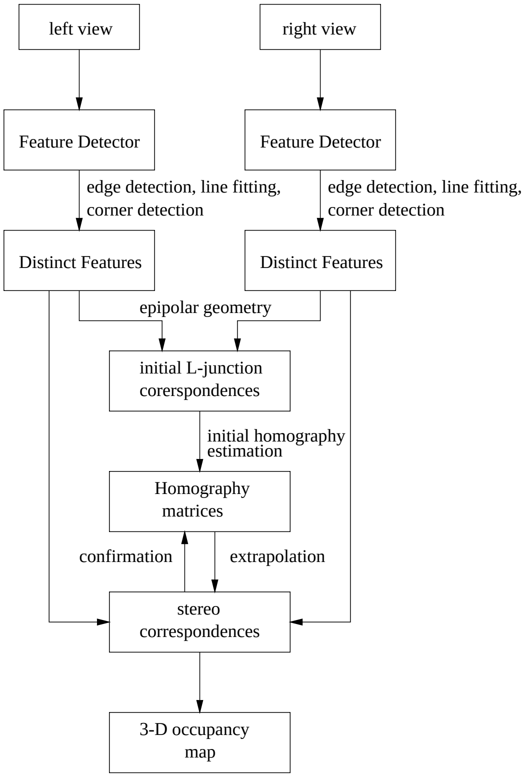


Figure 1: Overview of the recovery mechanism.

in one image, the identified homographies can be used in turn to predict its correspondence in the other image. In the implemented system only line segments are involved in this extrapolation process. Once the correspondence extrapolation and confirmation of homographies are completed, the representation of the environment as a number of homographies is available. The representation comes with a segmentation of the environment, i.e., different planar surfaces in the environment correspond to different matrices in the representation, and it is known which features in the images are contained in which matrix. This representation can be used to generate an occupancy map of the environment over the space around the detected line segments.

In the proposed system homographies are used in place of the surface-continuity and feature-ordering assumptions in resolving the correspondence ambiguity along epipolar lines. This has a number of advantages. Since smoothness is not assumed for the scene, occlusion is allowed and it affects the correspondence process only in the form of additional homographies. Moreover, as soon as the homographies are extracted, not only the features are matched, there is also a decomposition of the scene into various surfaces. Such an approach is particularly suitable for scenes which are composed mainly of planar surfaces.

4 Experiments

The stereo vision system has been implemented and tested with various image datasets. Two sets of real image experiments are presented here. Both sets of image data are obtained from the Department of Computer Science of University of Massachusetts at Amherst. The epipolar geometry of each stereo pair is known. The focal lengths and the baseline width of the stereo geometry are unknown though, and they are assumed with arbitrary values in the presentation of the results.

For all image data sets, line segments are first extracted from images through Canny edge detector and the line-fitting subsystem of Nevatia-Babu LINEAR package. A simple corner detector is then applied to the line segments, examining if any two lines segments have their end points nearby and have their orientations different enough, thereby proposing an L-junction at the intersection of the line segments if they do.

Figure 2 shows stereo images of a corridor with two walls, accompanied with the extracted L-junctions. Initial L-junction correspondences unique under the epipolar constraint allowed two homographies to be identified. As shown in Figure 3, one is for the wall on the left and another one for the wall on the right. Careful inspection of the stereo images can tell that there are altogether three vertical walls. However, the two on the left are so close to each other that they were indistinguishable under the thresholds used in the system. It is expected that if the scene is viewed at a closer range, the image resolution would allow the two to be separated. The homography matrices estimated are then used to extrapolate other stereo correspondences and be confirmed by them. The reconstruction result is illustrated with a reprojection of the environment from an oblique angle in Figure 4.

Figures 5, 6, and 7 show results over another set of image data. The stereo images are those of a hallway. The hallway consists mainly of five surfaces, two horizontal surfaces being the ceiling and floor, two vertical surfaces being the left and right vertical wall, and one being the end of the hallway, i.e., the exit. This hallway is so sparsely featured and the contrast of the imaging is so weak that almost no feature can be found except on the surface of the exit. The L-junction correspondences on the four horizontal and vertical surfaces are shown in the top of Figure 6. Even with as few as one L-junction correspondence over each of these four surfaces, they are enough to estimate the corresponding homographies. The exit has more features detected, as shown in the bottom of Figure 6. All the five sur-

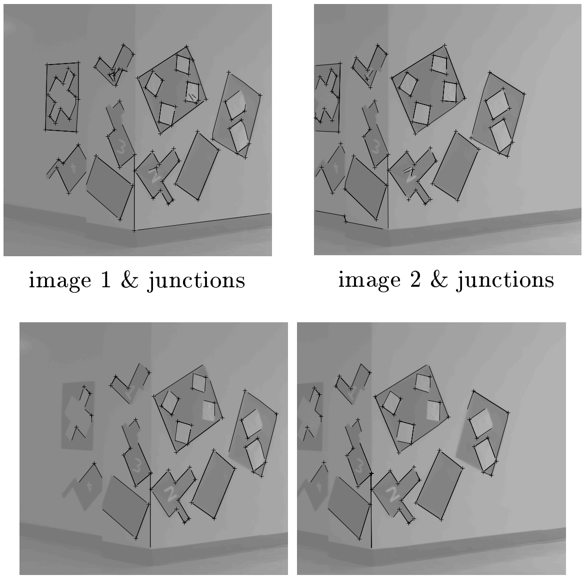
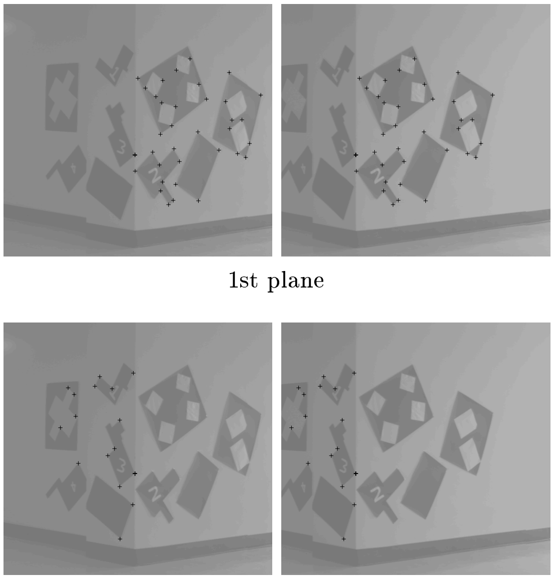


image 1 & junctions

image 2 & junctions

initial L-junction correspondences

Figure 2: Stereo images of a corridor and preliminary processings.



1st plane

2nd plane

Figure 3: Extracted homographies (corridor).

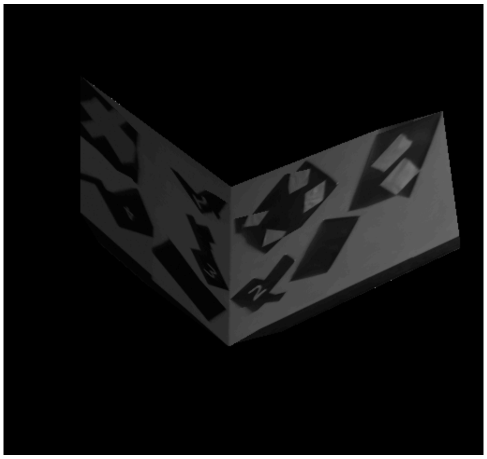


Figure 4: Bird's eye view according to reconstruction results (corridor).

faces were recovered at the end of the clustering process. There are tiny surfaces close and parallel to the surface of the exit, but under the thresholds of the implemented system they are indistinguishable. Again, it is expected that when the robot gets closer to the exit, the difference between their homographies and the homography of the exit will be significant enough for them to be isolated. This pair of stereo images is so sparsely featured that it presents great difficulty to generic stereo vision to recover a dense occupancy map of the environment. Yet, with the proposed representation and the recovery mechanism, how many surfaces there are and where they are positioned can be estimated. To illustrate the performance of the system, a side view of the reconstructed environment is shown in Figure 7, which is a reprojection from a small elevation angle. Part of the exit surface and the floor are occluded by the left wall in this elevated view.

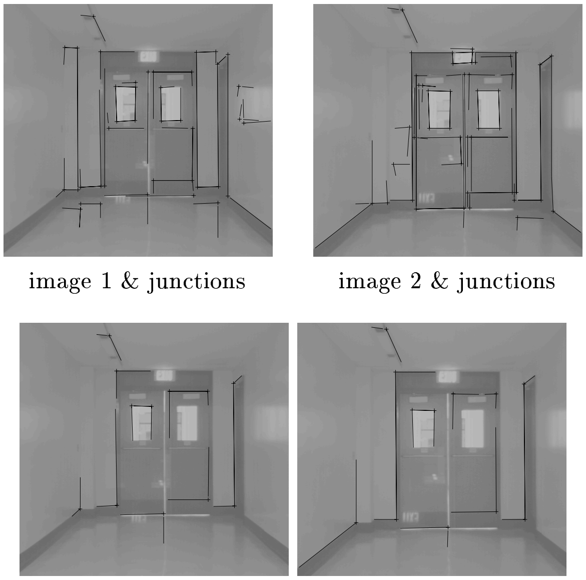
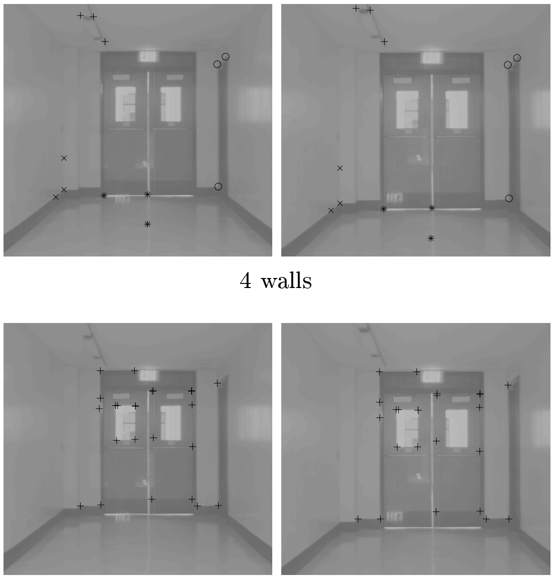


image 1 & junctions

image 2 & junctions

initial L-junction correspondences

Figure 5: Stereo images of a hallway and preliminary processings.



4 walls

exit

Figure 6: Five extracted homographies (hallway).

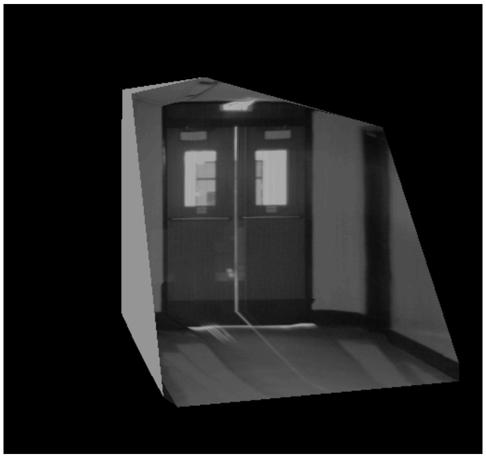


Figure 7: Side view according to reconstruction (hallway).

5 Conclusion

A new representation of polyhedral environment is proposed and its recovery is presented. Experimental results on real images has been encouraging. Future work will include further experiments over environments with occlusions.

6 Acknowledgments

The courtesy of the Department of Computer Science of the University of Massachusetts at Amherst in supplying the image data sets is much appreciated. The work described in this paper (or the equipment/facility) was partially supported by a grant from the Research Grants Council of Hong Kong Special Administrative Region, China (RGC Ref. No. CUHK 4169/98E).

References

[1] U. R. Dhond and J. K. Aggarwal. Structure from stereo–A review. *IEEE Transactions on Systems, Man & Cybernetics*, 19(6):1489–1510, November/December 1989.

[2] O. Faugeras. Stratification of three-dimensional vision: projective, affine, and metric representations. *Journal of the Optical Society of America - A*, 12(3):465–484, March 1995.

[3] G. A. Jones. Constraint, Optimization, and Hierarchy: Reviewing Stereoscopic Correspondence of Complex Features. *Computer Vision and Image Understanding*, 65(1):57–78, January 1997.

[4] S. Y. Lu and K. S. Fu. A Sentence-to-Sentence Clustering Procedure for Pattern Analysis. *IEEE Transactions on Systems, Man and Cybernetics*, 8(5):381–389, May 1978.

[5] Q.-T. Luong and T. Vieville. Canonic Representations for the Geometries of Multiple Projective Views. *Computer Vision and Image Understanding*, 64(2):194–229, September 1996.

[6] A. Shashua. Algebraic Functions for Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 17(8):779–789, August 1995.