The relationship between photogrammetry and computer vision

J.L. Mundy
GE Corporate Research and Development
Schenectady, NY 12309

Abstract

The relationship between photogrammetry and computer vision is examined. This paper reviews the central issues for both computer vision and photogrammetry and the shared goals as well as distinct approaches are identified. Interaction in the past has been limited by both differences in terminology and in the basic philosophy concerning the manipulation of projection equations. The application goals and mathematical techniques of both fields have considerable overlap and so improved dialog is essential.

1. INTRODUCTION

1.1 Motivation

The goal of this paper is to clarify the relationship between the disciplines of photogrammetry and computer vision. In recent years, a number of applications areas have evolved linking the two fields, such as image-based cartography, aerial reconnaissance and simulated environments. On the other hand, there has been almost no genuine exchange of ideas between the two fields. It is hoped that the following discussion will help illuminate the difficulties and suggest avenues for progress.

1.2 Two illustrations

The situation is well illustrated by an exchange several years ago between a computer vision researcher and an experienced photogrammetrist. The discussion turned on the number of possible solutions for the problem of camera pose determination for a triangle under perspective. This example is of considerable theoretical interest to the computer vision community but is of practically no interest to photogrammetrists. It is well known that, in general, there are multiple solutions for the six degrees of freedom for the pose of the camera with respect to the coordinate frame of the triangle. However, the photogrammetrist refused to even admit the possibility that there could be more than one solution. Both parties quit the exchange in frustration and without any mutual benefit.

Another example concerns image formation by projection through a single point in space. This projection is called a *pinhole* model by computer vision researchers since the central projection model is realized exactly when the camera is implemented with an infinitesimal pinhole lens. Curiously, this term is found to be grating to the ears of a photogrammetrist who prefers the terms *perspective* or *central projection.* Perhaps the term suggests an inexpensive camera which would never be employed in exacting mapping applications. In any case, many of the barriers to collaboration between the two fields arise from such clashes of terminology.

1.3 The root of the problem

Photogrammetry is a mature subject with well established problem descriptions and solutions. As a consequence, photogrammetrists are generally not very tolerant of results couched in alternative terminology and with somewhat different goals. This is not to say that there is any deficiency of stubbornness and dogma on the computer vision side. The IU community is largely unaware of much of the historical photogrammetry literature. Many results developed over the last decade by IU researchers were already known early in this century by photogrammetrists and existing photogrammetric theory still has much to offer to the problems of object recognition and scene modeling. At the same time, results from IU can help to advance photogrammetry both in the discovery of completely new approaches as well as the automation of control point correspondence and complex feature extraction.

In order to clarify the problem, a number of key questions have to be addressed.

- What are the goals of computer vision?
- What are the goals of photogrammetry?
- What are shared goals?
- What are distinct differences?

At the outset, we should define the terminology to be used. Both disciplines are concerned primarily with the *pinhole* or *perspective* camera, and the mapping from points in the 3-D world (object points) to 2-D image points. In photogrammetry, this is usually expressed in terms of the collinearity equations, whereas in computer vision this mapping is usually expressed (equivalently) as a linear mapping of homogeneous coordinates. In particular, a point in the 3D world is expressed as a 4-vector $\mathbf{x} = (x, y, z, t)^\top$ representing the point $(x/t, y/t, z/t)$ and an image point is expressed as $\mathbf{u} = (u, v, w)^\top$ representing the point $(u/w, v/w)$ in Euclidean coordinates. Two homogeneous vectors that differ by a constant scale factor represent the same point. The mapping from the world to image space may then be succinctly expressed by a 3×4 matrix, M, called the *camera matrix* such that $\mathbf{u} = M\mathbf{x}$. Provided that the camera centre is not at infinity, an arbitrary such 3×4 matrix may may be factored as a product

$$M = KE \tag{1}$$

where K is an upper triangular 3×3 matrix and E is a 3×4 matrix representing a Euclidean (rigid) coordinate transformation. The matrix E encodes the *exterior orientation* of the camera, whereas K represents the *calibration* of the camera. The matrix M, and hence also K are determined only up to a scale factor, K has 5 degrees of freedom. The usual parameters of *interior orientation*, namely principal point offset and focal length (or magnification) occur as entries in the matrix K. The other calibration factors are pixel skew and pixel aspect ratio, which may also be read from K. In photogrammetric applications these are often ignored, resulting in a calibration matrix K of the form

$$\begin{pmatrix} k & 0 & p_u \\ 0 & k & p_v \\ 0 & 0 & 1 \end{pmatrix} \tag{2}$$

where k is the magnification factor[1] as and (p_u, p_v) are the coordinates of the principal point.

There are advantages to allowing full calibration matrices however.

2. THE GOALS OF COMPUTER VISION

The field of computer vision[2] has evolved under the central theme of achieving human-level capability in the extraction of information from image data. There are many and diverse applications of computer vision since much of human experience is associated with images and with visual information processing. Below we discuss three major technical goals and provide a brief discussion of issues that will be important to the subsequent analysis.

[1] Some authors use the focal length $f = 1/k$ instead.

[2] The technical discipline of computer vision is also often called image understanding.

Figure 1: The operational structure of object recognition algorithms.

2.1 Object recognition

The desired outcome is for a recognition algorithm to arrive at the same class for an object as that defined by the human conceptual framework. For example, a long term goal of computer vision with respect to aerial reconnaissance applications is *change detection*. In this case, the changes from one observation to the next are meant to be *significant* changes, i.e. significant from the human point of view. Thus, in order to define only significant change it is essential to be able to characterize human perceptual organization and representation.

Current object recognition algorithms operate according to the data flow illustrated in Figure 1 Image features are extracted from the image intensity data such as:

- regions of uniform intensity,
- boundaries along high image intensity gradients,
- curves of local intensity maxima or minima, e.g. line features,
- other image intensity events defined by specific filters, e.g. corners.

These features are processed further to extract high level measurements. For example, a portion of a step intensity boundary may be approximated by a straight line segment and the parameters of the resulting line are used to characterize the boundary segment. As another example, an image region can be characterized by statistical parameters such as intensity mean and standard deviation, as well as geometric properties of the region, such as the aspect ratio of a rectangular box enclosing the region.

A next key step in recognition is then formation of a model for each class. Some recognition algorithms store the feature measurements for a particular object, or a set of object instances for a given class, and then use statistical classification methods to classify a set of features in a new image according the stored feature measurements. If these measurements are view dependent, the resulting classification accuracy will suffer unless feature models are stored for a large number of viewpoints[3]. Other model-based recognition algorithms use a 2D or 3D geometric model for the object and use this model to predict the appearance of the object in a new image. The prediction requires that the pose(translation and orientation) of the stored model be determined with respect to the camera reference frame of the new image.

The classes are usually defined in terms of human concepts. For example, a classifier may be constructed to identify types of aircraft from aerial views. In the case of model-based vision, a 3D geometric model of each class is derived so that the salient features for each type are emphasized. In this approach, a direct link is established between geometry and concept. More recent work is aimed at establishing a link between *function* and class[14]. Again in this approach a geometric description of the object provides a basis for determining function.

2.2 Navigation

The goal of navigation differs somewhat from recognition in that the main function is to provide guidance to an autonomous vehicle. The vehicle is to maintain accurate following along a defined path. In the case of a road, it is desired to maintain a smooth path with the vehicle staying safely within the defined lanes. In the case of off-road travel, the vehicle must maintain a given route and the navigation is carried out with respect to landmarks.

[3] Recent algorithms that exploit projective invariants have defined viewpoint-invariant measures for planar shapes[10].

A secondary goal of navigation is obstacle avoidance. Here the vehicle must avoid 3D structures, usually of unknown class. The objective is to produce an accurate description of the 3D environment around the vehicle. In current navigation projects, this 3D structure is recovered by various techniques:

- laser range sensing,
- sonar range sensing,
- stereo,
- structure from motion.

The most relevant to our current discussion are the last two items. The main problem in stereo is to automatically identify correspondences between two images collected from a stereo camera configuration. A secondary problem is to carry out the calibration of these two cameras so that the image epipolar geometry and the resulting 3D coordinates can be computed.

In structure from motion, a time sequence of images is acquired. A set of correspondences between features in each element of the sequence is determined and maintained from frame to frame. The correspondences define both the camera position in space as well as the 3D structure of the points defined defined by the correspondences. For example, Longuet-higgins showed that 8 correspondences between two views are sufficient to determine both camera pose and 3D point coordinates[8]. It is generally not possible to determine the overall scale of the 3D coordinate space, even with a calibrated camera[4]since the distance and size of an object can be mutually adjusted without changing its image appearance.

2.3 Object modeling

A third goal of computer vision is somewhat related to the 3D structure recovery of the previous section. Here the central issue is to recover a complete and reasonably accurate 3D model of an object. The model is then used for a number of applications:

- to support object recognition, as described earlier,
- for image simulation, where image intensity data, is projected onto the surface of the object and provides realistic image of the object from any desired viewpoint.

Image simulation is used extensively for military training and other applications such as virtual reality for entertainment purposes. In the simulation of military sites, it is often necessary to provide a 3D terrain model, in addition to buildings or other cultural features. A model is constructed by positioning and adjusting geometric primitives over several views of the object simultaneously. Some systems use a complete set of CAD primitives, but the most structures are represented as simple polyhedral box shapes.

The goal of computer vision is to automated the model construction process so that a minimum of human intervention is required. One major approach to the automation of 3D structure extraction is the use of automated stereo algorithms[6]. The stereo correspondences are determined by image feature matching. For example, contextual clues such as shadows can be used to reinforce the validity of proposed correspondences. Another approach is to extrude the occluding boundaries of an object along the direction of view for a given camera position. This extrusion forms a solid which has an outer boundary corresponding to the object shape along that viewing direction. Then more *view solids* are constructed from other views.

[4] Here the term *calibrated* means that the internal parameters of the camera are known, such as focal length, principal point and image coordinate aspect ratio. The relationship between pixel location and projection ray angle relative to the principal ray is known for a calibrated camera.

Figure 2: The general problem setting for aerial photogrammetry.

The Boolean intersection of all these views define a reasonable approximation to the outer boundary of the object[15]. An advantage of this approach is that correspondences across views are not required.

Finally, a generic or parametric model for an object class can be defined in terms of geometric constraints such as line symmetry, coplanarity and incidence[11]. Usually a continuous space of 3D model configurations is defined by the constraint system. The specific model is determined by minimizing the error of projecting the model onto one or more image views of the object, subject to the constraints. This approach provides a mechanism of including general information about an object class, which can be provided a priori. The computation of error requires the definition of correspondences between the model features and image features. Currently, these correspondences are manually defined, but the heuristics used to define stereo correspondences can be applied here as well.

3. THE GOALS OF PHOTOGRAMMETRY

The central theme of photogrammetry is **accuracy**. Photogrammetry developed in the last century, starting almost at the same time as the discovery of photography itself [5]. Initial applications were motivated by military considerations, but photogrammetry is now applied across a diverse set of commercial applications as well. A related field, *remote sensing* exploits many of the same techniques but perhaps with more emphasis on the radiometric aspects of image data while the main issue in photogrammetry is geometric accuracy. The most common camera model is *central projection*, which is a good approximation to image formation in a conventional camera. Accurate photogrammetric models also exist for other sensing geometries such as a moving linear array(SPOT) and the panoramic camera.

3.1 Mapping

The most important application area for photogrammetry is in the production of topographic maps. The images to support mapping are carefully collected and the internal parameters of the camera are known to great accuracy. The required imagery is usually collected from aircraft or from space from approximately a nadir(overhead) view. The general problem context is illustrated in Figure 2. The main technical issue is to compute the location of features on the ground as accurately as possible from corresponding sets of image features. The relationship between the camera positions and the earth coordinate frame is computed from a set of ground control points obtained from an accurate ground survey. Additional map features are located through triangulation among the set of aerial views.

There are many effects limiting the ultimate accuracy that can be achieved in photogrammetrically determining the 3D coordinates of a point on the ground. Some of the sources of error are:

- error in image feature position,
- error in ground control point position
- error in the camera projection model, e.g. radial distortion,
- numerical error in solving the projection equations.

When these effects are modeled, the resulting projection equations are non-linear. The goal is to find the set of camera parameters, image feature positions and ground control point positions which minimize the mean square error in projected image feature position and mean square 3D ground control point position

[5] The term **photogrammetry** was coined by the German geographer Kersten in 1855.

error. A set of normal equations are developed by differentiating the error cost function with respect to all of the projection variables[13]. The resulting solution provides the optimum values for all of the variables as well as error ellipsoids, which are derived from the overconstrained minimization process.

In this solution method, it is possible to introduce known accuracies a priori in terms of variances for both the ground control points and the image point locations so that all of the data can be combined with appropriate weighting. The method is easily generalized to the case of an arbitrary number of photographs and an arbitrary number of views. The only restriction is that each image should have a reasonable amount of overlap with some of the other images so that viewing constraints can be propagated.

3.2 Close range photogrammetry

A distinction is usually made for applications of photogrammetry that involve short viewing distances, compared to the thousands of feet involved in aerial photography. The main differences arise because the model of central projection may not be accurate enough or may not apply to the actual imaging conditions. Also close range photogrammetry implies a large range in spatial resolution for different applications. Because of this huge range, a single set of projection equations cannot be applied and solution techniques are problem dependent.

Typical applications for close-range photogrammetry are:

- architecture,
- anthropometrics(measurement of the human body),
- industrial metrology,
- archeological surveying.

The major application focus is the construction of detailed and accurate 3D models from a series of images. The images are usually taken from viewpoints that optimize the accuracy and completeness of the resulting model.

4. SHARED GOALS

The intersection of interest for the two fields centers on the theory and applications of the central projection camera.

4.1 Camera calibration

There has been a great deal of research in the computer vision community to solve what is the fundamental problem of photogrammetry. Camera calibration is defined as the determination of internal sensor parameters such as focal length, pixel skew and principal point. Once these parameters have been determined, the camera can be used in such applications as stereo to derive absolute positional measurements.

4.2 Pose determination

Pose determination is a technique central to model-based vision. This problem is known in the photogrammetry literature as *resectioning* where the position and orientation of a camera is determined from a set of known points in 3D space.

4.3 Model Projection

In model-based vision applications a 3D model is projected onto a set of features which are hypothesized to be a particular view of the object. When the projected model features are in close agreement with the observed image features, the hypothesis is confirmed. Also a recent advance called model supported exploitation(MSE) requires the interactive projection of a site model onto an aerial image. The projected model is then used to assist an image analyst or to guide localized computer vision algorithms. In both these cases, the problem of accurate projection of a set of 3D points is central to both computer vision and photogrammetry.

4.4 Model construction

As mentioned earlier, some applications of computer vision are aimed at the construction of a 3D model of the environment from a series of perspective images. Also, the models required by the MSE approach described in 4.3 are typically derived from a set of image views.

5. What are the differences?

We are now in a position to discuss the central theme of this paper, i.e. why isn't geometrically-oriented computer vision just a type of photogrammetry?

5.1 Grouping and combinatorics

A major driving force for a difference in treatment of the calculations surrounding image projection arises from the combinatorics associated with grouping fragmented image features. Most model-based object recognition algorithms depend on groups of line segments and vertices that have been extracted from image data through the detection of step discontinuities in image intensity.

It is necessary to group these fragments into a set of a certain minimum complexity in order to continue with the next level of processing such as matching to the features of an object model. For example, a minimum of six points are required to linearly determine the projection matrix. That is, given a set of six 3D points and the corresponding 2D image locations, determine the 3x4 projection matrix that maps the 3D points onto the image points[12]. In principle, once these six points have been determined, the full model can be projected on to the image and verified as the correct class. However, such an algorithm would be hopeless because of the combinatorics. For example, if the image contains even a hundred point features, the resulting number of possible six-point model combinations exceeds one billion!

There are several possible ways to deal with this complexity:

- use an approximation to image projection requiring fewer features,
- use higher level features such as a vertex and the edges incident on the vertex,
- decompose the image projection transformation and interleave grouping with projection,
- use contextual information to limit the combinations that have to be tested.

Thus, it is common in computer vision research to employ camera models that are only approximations to central projection in order to reduce the complexity of recognition. For example, many computer vision recognition algorithms assume a more limited form of camera geometry – affine projection. Affine projection is also known as weak perspective. Affine projection assumes that the camera viewing distance is large compared with the depth change of the object along the principal ray direction. For affine projection, only three control points are required to derive the the model projection parameters. So for 100 points, the number of combinations is reduced to less than 200,000. Similarly, two vertices, along with their associated incident edge directions, is sufficient to determine the parameters of affine projection[5] and the number of combinations is reduced to a mere 5000.

There is considerable advantage to be gained by decomposing the projection into a series of separable effects, so that grouping can be interleaved with projection. For example, the 3x3 central projection transformation between two planes can be uniquely decomposed into three matrices, as follows.

$$T = \begin{bmatrix} 1 & 0 & 0 \\ 0 & 1 & 0 \\ a & b & 1 \end{bmatrix} \begin{bmatrix} c & 0 & 0 \\ d & f & 0 \\ 0 & 0 & 1 \end{bmatrix} \begin{bmatrix} \cos\theta & \sin\theta & t_1 \\ -\sin\theta & \cos\theta & t_2 \\ 0 & 0 & 1 \end{bmatrix}$$

The decomposition isolates the effects of perspective, internal camera calibration(an affine transformation) a Euclidean transformation of the world plane.

The terms in the perspective matrix can be determined by identifying two or more vanishing points in the image. The determination of vanishing points involves grouping lines pairwise. This strategy of grouping around vanishing points is often effective since many man-made features are aligned along consistent directions. The effects of perspective can then be removed and additional simple grouping strategies applied to determine the affine and Euclidean transformation parameters[1].

Context can be used in many ways to reduce the combinatorial problem. For example, if the object is assumed to be planar, only four point correspondences are required to determine the associated 3x3 projection matrix linearly. Also, introducing assumptions about the viewpoint can also reduce complexity. Feature combinations producing camera locations that fall into a predetermined forbidden range do not have to be verified.

All of these strategies lead to a difference in emphasis from the classical approaches in photogrammetry. In the case of computer vision *accuracy* is given lower priority in order to derive a small number of object hypotheses. In photogrammetry, the emphasis is on deriving a *globally* consistent geometric description. For the purpose of recognition, it is not essential that the projection used to map a single object onto the image be consistent with that derived from other objects in the scene. The requirement for computer vision is that the average error of projection for a single model be small compared to the error resulting from projecting an incorrect model.

5.2 Invariants and uncalibrated cameras

Recent developments in computer vision research have demonstrated that it is not essential to have any knowledge about camera position or camera calibration in order to carry out recognition and model projection tasks.

Measurements can be derived from small sets of geometric features that are invariant to both camera viewpoint and to camera calibration. These measurements are called geometric invariants. Geometric invariants thus provide a direct index into a model library without computing a camera projection or assuming any camera calibration. An additional advantage of eliminating the direct dependence of indexing from camera calibration is that the derivation of camera parameters is often numerically ill-conditioned. The resulting parameters will have large uncertainties unless the number of control points is large. On the other hand, invariant functions can be constructed, which vary in a smoothly and continuously with respect to small variations in feature geometry.

When 3D feature geometry can be reconstructed up to a projective transformation of space from two or more uncalibrated camera views[4]. The images can be acquired by completely different and unknown central projections. Further, the projection of 3D features is determined completely from the projective epipolar structure of a collection of images. Thus it is possible to determine the projection of a 3D model in a new unknown image without knowing the either 3D coordinates of the model or the parameters of the image projection. With a minimum of 8 feature correspondences among three views, any number of additional features can be projected from two given views to a third view[2].

Both these ideas lead to an approach to computer vision which is *view-centered*. The emphasis is on representing objects in terms of small feature sets and associated invariant functions instead of compiling this information into a conventional 3D model. The advantage is that models can be acquired directly from image observations with a minimum of human interaction. Due to unavoidable image segmentation errors, a topologically complete 3D model can not usually be constructed without manual intervention. By contrast, the emphasis in photogrammetry is *world-centered*, i.e., the goal is to derive an accurate model of the world.

6. An approach to collaboration

In view of the shared goals and differences in emphasis, what should be the focus of collaboration? First it is essential that existing photogrammetry algorithms be understood and adopted by the computer vision community. To this end, a review article or even a monograph should be jointly authored by a representatives of both communities. In this way, differences in terminology and problem statement can be resolved and mutually agreeable notation established. Obvious topics to be reviewed are:

- camera calibration,
- stereo,
- accurate model construction,
- and navigation.

Second, research effort can be shared on unresolved problems associated with geometry-based computer vision. Joint research projects should be encouraged by specific funding for such collaborations. An ideal topic for such joint activities is the relationship between geometric configurations and the robustness of the resulting image measurements. Some specific issues are:

- error characteristics of invariants,
- error characteristics of projective structures,
- identification of critical configurations.

The new view-centered approaches often make use of projective homogeneous coordinates. Work is only just beginning in the computer vision community to develop an understanding of the minimization of error in these coordinate systems[7]. Similarly, there is little currently known about the error behavior of invariant feature measurements. It is clear that variations in feature configuration and camera distortions will affect the value of invariants, but usable theories to simply characterize this behavior are not yet available.

Finally, it is important to develop an understanding of critical configurations in the context of these new approaches. It is has long been recognized that the resectioning problem is degenerate for certain configurations of points in space and the center of projection. Early German photogrammetrists even constructed physical models of the degenerate quadric surfaces where resectioning fails in order to visualize the critical spatial relationships between camera center and 3D control points. View-based methods can fail in a similar manner. For example, model transfer is degenerate when the plane formed by the three centers of projection intersects the field of view in the scene[6].

It is hoped that this paper has helped to clarify some of the issues that have impeded progress in the past and presented some useful suggestions to improve collaboration between photogrammetry and computer vision in the future.

More Stuff

[6] An observation made by R. Hartley.

A further difference between computer vision and photogrammetry lies in the emphasis placed in computer vision applications on fast, preferably linear techniques. Such techniques sometimes sacrifice some precision for speed. By contrast, photogrammetry places emphasis on the tried least-squares minimization techniques, which though slow usually produce optimal results. The need for speed in computer vision applications arises from the need to test large numbers of hypotheses often in real time. In the navigation of a robot, for instance, the position of the robot must be continually computed from its view of the surrounding scene. Extreme precision is not usually required. In addition, for computer vision applications it is often impossible to assume *a priori* knowledge of the camera parameters, particularly the external orientation of the camera. For this reason, computer vision has emphasized the development of algorithms that require no knowledge of camera placement.

An example of this is the use of the Direct Linear Transformation (DLT) method of resectioning [16]. In this method, one solves for the entries of the camera matrix M directly from a set of at least six world to image correspondences. The algorithm is linear and hence very fast. In addition it requires no initial guess of the camera parameters. For many purposes it gives adequate (though not optimum) results in the presence of noise.

The Relative Placement Problem More interesting is case in which no ground truth (knowledge of world points) is assumed, only a set of correspondences between a pair of images. The traditional approach in both the computer vision and photogrammetry communities has been to assume that the internal camera parameters of the cameras are known[7], and the task is to solve for 5 parameters of relative camera placement. Subsequently, the 3D scene may be reconstructed up to scaled Euclidean transformation. Longuet-Higgins ([**?**]) gave a very simple linear algorithm for solving this problem. In his approach, a matrix Q is defined, which has subsequently come to be known as the *essential matrix*. The essential matrix is defined by the equation

$$\mathbf{u}_i'^{\top} Q \mathbf{u}_i = 0 \tag{3}$$

where $\mathbf{u}_i$ and $\mathbf{u}_i'$ are homogeneous vectors representing a pair of matched points in two images. The matrix Q is defined only up to a constant factor. Given enough correspondences (at least 8) it is possible to solve for the entries of Q using a linear least-squares method. The basic result concerning the essential matrix is that Q may be factored as a product $Q = RS$, where R is a rotation matrix representing the orientation of the second camera with respect to the first, and S is a skew-symmetric matrix of the form

$$\begin{pmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{pmatrix}$$

where $(t_x, t_y, t_z)^{\top}$ is a vector representing the placement of the second camera in the coordinate frame of the first camera. Longuet-Higgins gave a method of factoring the Q, thereby solving the relative placement problem. A simpler method for factorizing Q is given in [**?**]. It has recently been shown ([**?**]) that the essential matrix may be used to solve the relative placement problem in the case where the magnification factors (or focal lengths) of the two cameras are unknown. The other calibration parameters, including the principal points of the two images must be assumed known, however. Thus, one may solve for the two focal lengths and the relative placement using only image correspondences. This is the maximum amount of information that may be computed using image correspondences. The algorithm of [**?**] is non-iterative and very fast.

[7] In this case, the problem may be reduced to one in which the calibration matrix K is the identity.

Interesting work has been done on describing critical configurations for the relative placement problem. These are situations in which relative placement is not unique.

The minimum situation in which only five matched points are given is of particular theoretical interest. In this case, the solution to the relative placement problem is not unique. It has been shown that in this case a maximum of 10 solutions (or 20 counting "conjugate solutions") exist to a relative placement problem. A difficult proof of this result is found in [**?**]. A much simpler (and conceptually enlightening) proof is found in [**?**].

The method of Longuet-Higgins has been shown to suffer from instability in the presence of noise. For optimum results, an iterative method is needed. Horn ([**?**, **?**]) gives two different versions of iterative algorithms for solving the relative placement problem based on representation of rotations using quaternions. These algorithms are among the best available. The methods are closely related to methods of least-squares optimization using the normal equations. The linear method of Longuet-Higgins may be used as an initial guess for iteration.

Methods for Uncalibrated Cameras The need to calibrate cameras has always been a thorn in the side of the computer vision worker. The photogrammetrist typically works with very high quality, highly expensive cameras for which the calibration may be computed to high accuracy. The photographs are usually taken under highly controlled conditions for which in addition accurate knowledge of the external calibration is also known. This applies especially to satellite images for which accurate ephemeris information is often available. In computer vision applications on the other hand the source of images is not always so well known. In many applications (such as intelligence applications) the calibration of the camera may be entirely unknown. In robotics applications, a roving robot may be moving over rough terrain while zooming and unzooming its camera. Neither the internal or external camera parameters will be known. Many papers have dealt with methods of calibrating cameras, for the most part relying on complicated, accurately measured calibration jigs (for instance see [**?**, **?**, **?**]). Because of the problematic nature of camera calibration, a recent line of work has dealt with photogrammetric problems for uncalibrated cameras.

The direct linear transformation method described above is an example of the sort of algorithm that works for uncalibrated cameras, determining the camera calibration and the exterior orientation simultaneously. Apparently the first method given for extracting physically meaningful camera parameters from the camera matrix was given by Ganapathy ([3]). However, his method is unnecessarily complex. In fact, all that is needed is to compute the factorization (1) using the QR decomposition, after which the external orientation may be read from the matrix E and the interior calibration from the entries of K.

The relative placement problem for uncalibrated cameras is again of interest. In the absence of ground-control, the relative camera placement can not be determined uniquely from two (or any number of) views, and the scene can not be reconstructed uniquely. Recently, however, it was shown by Faugeras ([**?**]) and Hartley et. al. ([**?**]) that given several image correspondences, sufficient in number to allow the essential matrix to be computed, the scene may then be reconstructed up to a projective transformation of three-dimensional space[8]. In addition, the camera transformations of the cameras may be determined up to simultaneous multiplication by a 4×4 matrix H. In other words, one can find a set of camera matrices M_i for each of the cameras. These may not be the "correct" camera matrices, but there will exist a 4×4 matrix H such that matrices M_iH are simultaneously correct. This result seems to be basic to an analysis of multiple images using uncalibrated cameras.

For many applications reconstruction up to projective transformation was sufficient. One example of this is "model transfer" as defined by Barrett ([**?**]). If a model of an object is constructed from its image in two uncalibrated views, then its aspect in a third image may be computed exactly once the third image

[8] A projective transformation of projective n-space P^n is a mapping represented by an invertible linear transformation on homogeneous coordinates

is registered to the reconstructed scene. Specifically, once six points in the third image are matched with points in the first two, then the camera matrix of the third camera may be computed (relative to the first two views) by the DLT method (or any other suitable method of resection) and this camera matrix may be used to project the model into the third image.

If ground control points are also available, it is possible to compute a Euclidean reconstruction of the image. In [**?**] a stereo terrain extraction method is described whereby the terrain is constructed up to a projective transformation using image-to-image correspondences and then the correct projective transformation of space is computed to transform the terrain to the correct Euclidean frame, using ground-control points. This is a linear method that accomplishes camera calibration and scene reconstruction simultaneously given image-to-image correspondences and ground-control points. Another method of imposing Euclidean constraints to translate a reconstructed scene to the correct Euclidean frame has been described in [9].

Autocalibration Additional information is available in analyzing multi-image sets if it is assumed that all the images are taken with the same camera (with the same unknown calibration). Indeed, if at least three views of a scene are given, along with image correspondences, then it has been shown that the calibration of the camera may be computed without the need for ground truth. This procedure is termed autocalibration, since it may be used to calibrate a moving camera without the need for special calibration rigs. This result has perhaps been implicitly known to photogrammetrists, but only recently has it been investigated thoroughly by Maybank and Faugeras ([**?**]). They use techniques of Algebraic Geometry to analyze systems of equations due to Kruppa (**??**) and prove the feasibility of autocalibration. Unfortunately, their method has not been turned into a practicable algorithm.

References

[1] J. R. Beveridge and E.M. Riseman. Can too much perspective spoil the view? In *Proc. DARPA Image Understanding Workshop*, pages 655–663, 1992.

[2] E. Barrett et al. Linear resection, intersection, and perspective independent model matching in photogrammetry:theory. In *Proc. SPIE Conf. on Applications of Digital Image Processing XIV, Vol 1567*, pages 142–170, 1991.

[3] S. Ganapathy. Decomposition of transformation matrices for robot vision. *Pattern Recognition Letters*, 2:410–412, 1989.

[4] R. Hartley, R. Gupta, and T. Chang. Stereo from uncalibrated cameras. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 761–764, 1992.

[5] A. Heller and J.L. Mundy. The evolution and testing of a model-based object recognition system. In *Computer Vision and Applications, R. Kasturi and R. Jain, eds, IEEE Computer Society Press.*, 1991.

[6] R. Irvin and D. McKeown. Methods for exploiting the relationship between buildings and their shadows in aerial imagery. *IEEE Transactions on Systems Man and Cybernetics*, 19:1564–1575, 1989.

[7] K. Kanatani. *Geometric Computation for Machine Vision.* Oxford University Press, Oxford, UK, 1993.

[8] H.C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, Sept 1981.

[9] R. Mohr, F. Veillon, and L. Quan. Relative 3D reconstruction using multiple uncalibrated images. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 543 – 548, 1993.

[10] J.L. Mundy and A. Zisserman. *Geometric Invariance in Computer Vision.* MIT Press, Boston, MA, 1992.

[11] V.-D. Nguyen, J. L. Mundy, and D. Kapur. Modeling generic polyhedral objects by constraints. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, 1991.

[12] L. G. Roberts. Machine perception of three-dimensional solids. In J. T. Tippett *et al.*, editor, *Optical and Electro-Optical Information Processing*, pages 159–197. MIT Press, 1965.

[13] C. C. Slama, editor. *Manual of Photogrammetry.* American Society of Photogrammetry, Falls Church, VA, fourth edition, 1980.

[14] L. Stark and K. Bowyer. Indexing function-based categories for generic recognition. In *Proc. IEEE Conf. on Computer Vision and Pattern Recognition*, pages 795–797, 1992.

[15] J.R. Stenstrom and C.I. Connolly. Constructing object models from multiple images. *International Journal of Computer Vision*, 9:185–212, 1992.

[16] I.E. Sutherland. Sketchpad: A man-machine graphical communications system. Technical Report 296, MIT Lincoln Laboratories, 1963. Also published by Garland Publishing Inc, New York, 1980.