Generalized Rank Conditions in Multiple View Geometry with Applications to Dynamical Scenes*

Kun Huang[1], Robert Fossum[2], and Yi Ma[1]

[1] Electrical & Computer Engineering Dept., and Coordinated Science Lab.
[2] Mathematics Department, and Beckman Institute
University of Illinois at Urbana-Champaign, Urbana IL 61801, USA
{kunhuang,r-fossum,yima}@uiuc.edu

Abstract. In this paper, the geometry of a general class of projections from $\mathbb{R}^n$ to $\mathbb{R}^k$ $(k < n)$ is examined, as a generalization of classic multiple view geometry in computer vision. It is shown that geometric constraints that govern multiple images of hyperplanes in $\mathbb{R}^n$, as well as any incidence conditions among these hyperplanes (such as inclusion, intersection, and restriction), can be systematically captured through certain rank conditions on the so-called multiple view matrix. All constraints known or unknown in computer vision for the projection from $\mathbb{R}^3$ to $\mathbb{R}^2$ are simply instances of this result. It certainly simplifies current efforts to extending classic multiple view geometry to dynamical scenes. It also reveals that since most new constraints in spaces of higher dimension are *nonlinear*, the rank conditions are a natural replacement for the traditional multilinear analysis. We also demonstrate that the rank conditions encode extremely rich information about dynamical scenes and they give rise to fundamental criteria for purposes such as stereopsis in n-dimensional space, segmentation of dynamical features, detection of spatial and temporal formations, and rejection of occluding T-junctions.

Key words: multiple view geometry, rank condition, multiple view matrix, dynamical scenes, segmentation, formation detection, occlusion, structure from motion.

1 Introduction

Conventional multiple view geometry typically applies to the case that the scene is static and only the camera is allowed to move. Nonetheless, it is easy to show that, if a scene contains independently moving objects – referred to as a *dynamical scene*, we usually can embed (by a certain formal process) the problem into a space of higher dimension, with a point in the high-dimensional space now representing such as the location and velocity of a moving point in the physical three-dimensional world [13]. However, results and understanding regarding multiple view geometry in such high

* This material is based upon work partially supported by the U.S. Army Research Office under Contract DAAD19-00-1-0466 and UIUC ECE department startup fund. Any opinions, findings, and conclusions are those of the authors and do not necessarily reflect the views of the above agencies.

dimensional spaces are rather sporadic or incomplete at best. This motivates us to seek a systematic generalization of multiple view geometry to higher dimensional spaces.

In the computer vision literature, geometric relationships between image features and camera displacements are traditionally described by the so-called *multilinear* matching constraints [8, 7, 11]. Most of the work since has focused on the algebraic aspects of these multilinear constraints, along with the algorithms which followed from the same formulation. In this framework, different geometric primitives, i.e. points, lines, and planes are treated separately, and analysis of multiple views must rely on a reduction to pairwise [8], triple-wise [11, 4, 1, 6] or quadruple-wise views [12].[1] This line of work culminated in the publication of two monographs on this topic [5, 2]. However, it has recently been discovered that a unifying and yet simplifying tool for characterizing geometric relationships among multiple views is a so-called *rank condition* [10]. It gives rise to a *global* constraint for multiple images of multiple features and incidence relations among them. Consequently, certain *nonlinear* relationships among multiple (up to four) images are revealed. As we will also see in this paper, in fact the majority of constraints among multiple images in high dimensional spaces are going to be *nonlinear*. Therefore, the rank condition currently seems to be a reasonable tool left which allows us to systematically generalize multiple view geometry to higher dimensional spaces.

In this paper, our focus is on a *complete* characterization of intrinsic algebraic and geometric constraints that govern multiple k-dimensional images of hyperplanes in an n-dimensional space. We show that these constraints can be uniformly expressed in terms of certain rank conditions, which also simultaneously capture geometric relationships among the hyperplanes themselves, such as inclusion, intersection, and restriction. The importance of this study is at least two-fold: 1. In many applications, objects involved are indeed multi-faceted (polygonal) and their shape can be well modeled (or approximated) as a combination of hyperplanes;[2] 2. In some cases, there is not enough information or it is not necessary to locate the exact location of points in a high-dimensional space and instead, we may still be interested in identifying them up to some hyperplane (e.g., in the case of segmentation). As we will point out later, for the special case $n = 3$ and $k = 2$, our results naturally reduce to what is known in computer vision for points, lines, and planes. For the cases $n > 3$ and $k = 2$, our results provide a simpler explanation to extant study on dynamical scenes (e.g., [13]) based on tensor algebra. Since reconstruction is not the main focus of this paper, the reader is referred to [10] for how to use such constraints to develop algorithms for various reconstruction purposes. Nonetheless, since the rank conditions encode extremely rich information about dynamical scenes, we will show how to use it as a basic tool to conduct multiple view analysis in high dimensional spaces, including *stereopsis* in n-dimensional spaces, *segmentation* of independently moving feature points, detection of spatial and temporal *formations*, and rejection of *occluding T-junctions*.

Outline of this paper. Section 2 provides a general formulation of multiple view geometry from $\mathbb{R}^n$ to $\mathbb{R}^k$, including the concepts of camera, camera motion, image, coimage, and preimage. Section 3 fully generalizes classical rank conditions to hyperplanes of

[1] Although we now know quadruple wise constraints are completely redundant in the point case.

[2] In case the object consists of smooth curves and surfaces, it is not hard to show that the rank conditions can be easily generalized (see [9]).

arbitrary dimension in $\mathbb{R}^n$. Rank conditions for various incidence conditions (i.e. inclusion, intersection, and restriction) among hyperplanes are presented. In Section 4, we discuss numerous potential applications of the rank conditions through a few concrete examples and simulation results.

2 Problem formulation

2.1 Euclidean embedding of dynamical scenes

In classic multiple view geometry, we typically consider projecting a three-dimensional *static* scene onto a two-dimensional image plane with respect to multiple camera frames. The standard mathematical model for such a projection is

$$\lambda(t)\boldsymbol{x}(t) = \Pi(t)\boldsymbol{X}, \tag{1}$$

where $\boldsymbol{X} \in \mathbb{R}^4$ is the homogeneous representation of the 3-D coordinates of a feature point $\boldsymbol{p}$ (relative to a world coordinate frame), $\boldsymbol{x}(t) \in \mathbb{R}^3$ is its image at time t (also in homogeneous coordinates), $\Pi(t) \in \mathbb{R}^{3\times 4}$ is the projection matrix, and $\lambda(t)$ is the depth scale of the point $\boldsymbol{p}$ with respect to the current camera frame. Typically $\Pi(t) = [R(t), T(t)]$ where $R \in \mathbb{R}^{3\times 3}, T \in \mathbb{R}^3$ respectively are the rotation and translation of the camera frame relative to a pre-fixed world frame.[3].

Now suppose the scene contains independently moving (relative to the world frame) feature points. Then the above equation must be modified to

$$\lambda(t)\boldsymbol{x}(t) = \Pi(t)\boldsymbol{X}(t). \tag{2}$$

Since now $\boldsymbol{X}(t)$ is time-dependent, methods from classic multiple view geometry no longer apply. However, suppose we can find a *time-base* for $\boldsymbol{X}(t)$, i.e. we can express the 3-D coordinates $\boldsymbol{X}(t)$ of $\boldsymbol{p}$ in terms of a linear combination of some time-varying functions $b_i(t) \in \mathbb{R}^4$ of time t

$$\boldsymbol{X}(t) = [b_1(t), b_2(t), \ldots, b_k(t)]\bar{\boldsymbol{X}} \quad \in \mathbb{R}^4, \tag{3}$$

where $\bar{\boldsymbol{X}} \in \mathbb{R}^{n+1}$ is a time-independent vector of coefficients. As an important example, $\boldsymbol{X}(t)$ is described by linear dynamics. Substitute this back to the above equation, to get

$$\lambda(t)\boldsymbol{x}(t) = \bar{\Pi}(t)\bar{\boldsymbol{X}}, \tag{4}$$

where $\bar{\Pi}(t) \doteq \Pi(t)[b_1(t), b_2(t), \ldots, b_k(t)] \in \mathbb{R}^{3\times(n+1)}$. This equation is of the same form as (1) except that it represents a (perspective) projection from the space $\mathbb{R}^n$ to $\mathbb{R}^2$. It is then natural to expect that results in classic multiple view geometry should generalize to this class of projections as well.

If a dynamical scene allows such a time-base, we say that the scene allows a *Euclidean embedding*. To avoid redundancy, the time-base functions $b_i(\cdot)$ should be chosen to be independent functions. The two examples shown by Figure 1 are scenes which

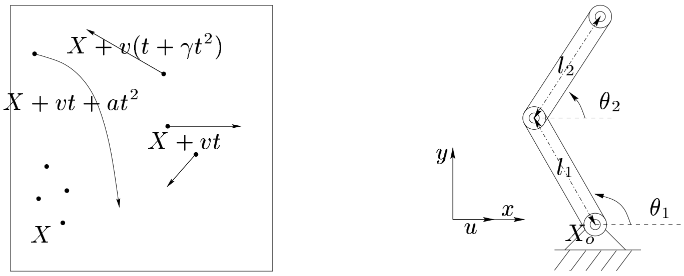


Fig. 1. a) Independent moving features; b) Two resolute joints.

do allow such an embedding. In the first case, since the coordinates of all points in the scene can be uniformly described by $X(t) = X + vt + at^2$ for some $X, v, a \in \mathbb{R}^3$, we can simply choose the embedded coordinates of a point to be $\bar{\boldsymbol{X}} = [X^T, v^T, a^T, 1]^T \in \mathbb{R}^{10}$. Then we have the following projection equation

$$\lambda(t)\boldsymbol{x}(t) = [R(t), R(t)t, R(t)t^2, T(t)]\bar{\boldsymbol{X}}. \tag{5}$$

Such an embedding would allow us to consider points with trajectories such as parabolic curves. In general, using similar techniques, one may also embed points on any rigid body as shown in Figure 1 b) of multiple links and joints (maybe of other types) into a high dimensional Euclidean space [3].

The above examples have shown the need for generalizing classic multiple view geometry to higher dimensional spaces. Although they both fall into the category of projection from $\mathbb{R}^n$ to $\mathbb{R}^2$ for some n, we will try to bring multiple view geometry into its full potential. That is, we will study the most general case by considering projections from $\mathbb{R}^n$ to $\mathbb{R}^k$ with arbitrary $k < n$.

2.2 Generalized multiple view geometry formulation

Homogeneous coordinates. In this paper we will use homogeneous coordinates for both points in $\mathbb{R}^n$ and its image in $\mathbb{R}^k$. Hence $\boldsymbol{X} = [X_1, X_2, \ldots, X_n, 1] \in \mathbb{R}^{n+1}$ is the coordinate for a point $\boldsymbol{p} \in \mathbb{R}^n$ and $\boldsymbol{x} = [x_1, x_2, \ldots, x_k, 1] \in \mathbb{R}^{k+1}$ is its image. However by abuse of language, we usually use $\boldsymbol{x}$ to denote as well the entire ray (1-dimensional subspace) spanned by $\boldsymbol{x}$, since any vector on this ray gives an equivalent (homogeneous) representation for the image of $\boldsymbol{p}$.

Image formation in high dimensional space. As a natural generalization of the perspective projection from $\mathbb{R}^3$ to $\mathbb{R}^2$, a perspective projection from $\mathbb{R}^n$ to $\mathbb{R}^k$ (with $k < n$) is described by the equation

$$\lambda(t)\boldsymbol{x}(t) = \Pi(t)\boldsymbol{X}, \tag{6}$$

[3] In case the camera is not calibrated, simply pre-multiply R and T by a calibration matrix $A \in \mathbb{R}^{3\times 3}$

where $\boldsymbol{x}(t) \in \mathbb{R}^{k+1}$ is the (homogeneous) image at time t of the point $\boldsymbol{X}$, $\lambda(t) \in \mathbb{R}$ is the missing depth scale and $\Pi(t) \in \mathbb{R}^{(k+1)\times(n+1)}$ is the *projection* matrix of full rank $k+1$. [4] For a dynamical scene, $\Pi(t)$ may depend on both the (relative) motion of the camera and the scene dynamics. Suppose multiple views of the same point (now in the high-dimensional space) are captured at time $t_1, \ldots, t_m$. The images of $\boldsymbol{X}$ then satisfy the equations

$$\lambda_i \boldsymbol{x}_i = \Pi_i \boldsymbol{X}, \quad i = 1, \ldots, m, \tag{7}$$

where $\lambda_i \doteq \lambda(t_i)$, $\boldsymbol{x}_i \doteq \boldsymbol{x}(t_i)$, and $\Pi_i \doteq \Pi(t_i) \in \mathbb{R}^{(k+1)\times(n+1)}$.[5] For the rest of the paper, we typically split $\Pi_i \doteq \left[\bar{R}_i \; \bar{T}_i\right]$ with $\bar{R}_i \in \mathbb{R}^{(k+1)\times(k+1)}$ and $\bar{T}_i \in \mathbb{R}^{(k+1)\times(n-k)}$. Note that here $\bar{R}_i, \bar{T}_i$ are not necessarily the motion (rotation and translation) of the moving camera, although they do depend on the motion.

Since Π is full rank, there always exists a matrix $g \in \mathbb{R}^{(n+1)\times(n+1)}$ in the general linear group $GL(n+1, \mathbb{R})$ such that $\Pi_1 g$ is in the standard form

$$\Pi_1 g = \left[I_{(k+1)\times(k+1)} \; 0_{(k+1)\times(n-k)}\right], \tag{8}$$

hence for simplicity, we will always assume Π_1 is itself in the above form already.[6] The reader should be aware that algebraically we do not lose any generality in doing so.

The following two assumptions make the future study well conditioned:

1. Motion of the camera is generic, i.e. for any p-dimensional hyperplane in $\mathbb{R}^n$, the its image in $\mathbb{R}^k$ is a hyperplane whose dimension is $q = \min\{p, k\}$.
2. Any hyperplane in $\mathbb{R}^n$ whose image is to be studied has a dimension $p < k$. If $p \geq k$, then its image will occupy the whole image plane for a generic motion and hence does not provide any useful information.)

The two assumptions above guarantee that we always have $q = p$ in this paper.

Remark 1 (Degenerate motions). Note that motions which violate the first assumption correspond to degenerate configurations which comprises just a zero-measure set of the overall configuration space of the camera and object. In addition, they would only induce minor changes in the results of this paper. A detailed analysis for these degenerate cases can be found in [3] and is omitted here.

Image, coimage, and preimage. For a p-dimensional hyperplane $H^p \subseteq \mathbb{R}^n$ whose points satisfy the equation $\Lambda \boldsymbol{X} = 0$ (where $\Lambda \in \mathbb{R}^{(n-p)\times(n+1)}$ is of rank $n-p$), it corresponds to a $(p+1)$-dimensional subspace (i.e. a hyperplane passing through the origin) $G^{p+1} \subseteq \mathbb{R}^{n+1}$ w.r.t. the camera frame. The image of H^p is then a p-dimensional hyperplane S^p in the image space $\mathbb{R}^k$ (since $p < k$), it corresponds to

[4] if $\mathrm{rank}(\Pi) = k' < k+1$, then the problem simply becomes the projection from $\mathbb{R}^n$ to $\mathbb{R}^{k'-1}$.

[5] Usually for a static scene with a moving camera, we have $\Pi_i = \Pi_1 g_i = \Pi_1 \left[\begin{smallmatrix} R_i & T_i \\ 0 & 1 \end{smallmatrix}\right]$, where $R_i \in \mathbb{R}^{n\times n}$ and $T_i \in \mathbb{R}^n$ are usually the rotation and translation of the camera in $\mathbb{R}^n$.

[6] If g is an affine transformation, it simply corresponds to a different choice of the world reference frame. If g has to be a projective transformation, then it distinguishes the perspective and orthographic projections. For details, see [3].

a $(p+1)$-dimensional subspace $U^{p+1} \subset \mathbb{R}^{k+1}$. Hence the image can be described by the span of a matrix $\boldsymbol{s} = [u_1, u_2, \ldots, u_{p+1}] \in \mathbb{R}^{(k+1)\times(p+1)}$ or by its maximum complementary orthogonal space (with respect to $\mathbb{R}^{k+1}$), which is spanned by $\boldsymbol{s}^{\perp} = [v_1, v_2, \ldots, v_{k-p}] \in \mathbb{R}^{(k+1)\times(k-p)}$ such that $(\boldsymbol{s}^{\perp})^T \boldsymbol{s} = 0$. For clarity, we then call $\boldsymbol{s}$ the *image* of the hyperplane and call $\boldsymbol{s}^{\perp}$ its *coimage*. The reader must notice that they are equivalent ways of expressing the same geometric entity on the image plane. In the rest of this paper, we will use $H^p \subseteq \mathbb{R}^n$ and $G^{p+1} \subseteq \mathbb{R}^{n+1}$ interchangeably to

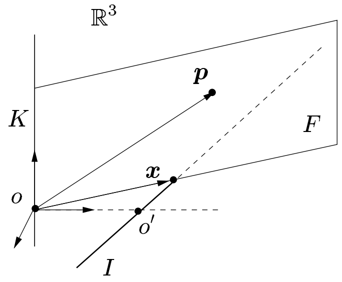


Fig. 2. An image $\boldsymbol{x}$ of a point $\boldsymbol{p} \in \mathbb{R}^3$ under a perspective projection from $\mathbb{R}^3$ to $\mathbb{R}^1 (\doteq I)$. K is the subspace orthogonal to the subspace spanned by o and the image plane I. The plane F corresponds to the preimage of $\boldsymbol{x}$ which is the subspace spanned by $\boldsymbol{p}$ and K.

refer to the same object in $\mathbb{R}^n$ and use $S^p \subseteq \mathbb{R}^k$ and $U^{p+1} \subseteq \mathbb{R}^{k+1}$ for the same image entity in $\mathbb{R}^k$.

One difference between the image formation in high dimensional space with the classical 3-D case is that the difference between the dimension of the ambient space and that of the image space might be larger than one. This leads to the notion of preimage. For any subspace $U^{p+1} \subseteq \mathbb{R}^{k+1}$ in the image space, if its equation is $(\boldsymbol{s}^{\perp})^T \boldsymbol{x} = 0$, then define its *preimage* to be the set $\boldsymbol{F} = \{\boldsymbol{X} \in \mathbb{R}^{n+1} : (\boldsymbol{s}^{\perp})^T \Pi \boldsymbol{X} = 0\}$, where Π is the corresponding projection matrix. Geometrically, the preimage $\boldsymbol{F}$ is the largest set in $\mathbb{R}^{n+1}$ that can give rise to the same image U^{p+1}. Its dimension is

$$\dim(\boldsymbol{F}) = \dim(\boldsymbol{F} \cap \mathbb{R}^{k+1}) + \dim(\boldsymbol{F} + \mathbb{R}^{k+1}) - \dim(\mathbb{R}^{k+1}) = (n-k) + p + 1.$$

It corresponds to a $(n-k+p)$-dimensional subspace F in $\mathbb{R}^n$. Figure 2 illustrates the notions of image and preimage for a special case when $n = 3, k = 1$ and $p = q = 0$. The dimension of F is $3 - 1 + 0 = 2$.

3 Generalized rank conditions on multiple view matrix

A typical problem in multiple view geometry is to systematically express all intrinsic constraints among multiple images of an object (in $\mathbb{R}^n$). By intrinsic we mean that such constraints should not explicitly depend on the location (or structure) of the object in

$\mathbb{R}^n$. In this section, we will give a complete description of such constraints including those for various incidence relations among different objects. We will only present the theorems without giving proof. For the complete proof please refer to [10, 3].

With the notation introduced in the preceding section, we may formally define a so-called *multiple view matrix* as following:

Definition 1 (Formal multiple view matrix). *We define a* multiple view matrix M *as*

$$M \doteq \begin{bmatrix} (D_2^{\perp})^T \bar{R}_2 D_1 & (D_2^{\perp})^T \bar{T}_2 \\ (D_3^{\perp})^T \bar{R}_3 D_1 & (D_3^{\perp})^T \bar{T}_3 \\ \vdots & \vdots \\ (D_m^{\perp})^T \bar{R}_m D_1 & (D_m^{\perp})^T \bar{T}_m \end{bmatrix} \tag{9}$$

where the D_i's and $D_i^{\perp}$'s stand for images and coimages of some hyperplanes respectively. The actual values of D_i's and $D_i^{\perp}$'s are to be determined in context.

Theorem 1 (Rank condition for multiple images of one hyperplane). *Given m images $\boldsymbol{s}_1, \ldots, \boldsymbol{s}_m$ and coimages $\boldsymbol{s}_1^{\perp}, \ldots, \boldsymbol{s}_m^{\perp}$ of a p-dimensional hyperplane H^p in $\mathbb{R}^n$, choose in the above multiple view matrix $D_1 = \boldsymbol{s}_1$ and $D_i{}^{\perp} = \boldsymbol{s}_i^{\perp}, i = 2, \ldots, m$, then the resulting matrix M satisfies*

$$\boxed{0 \leq \mathit{rank}(M) \leq (n-k).} \tag{10}$$

If the hyperplane happens to be a point (i.e. $p = 0$), the theorem easily implies the following result:

Corollary 1 (Multilinear constraints for $\mathbb{R}^n \to \mathbb{R}^k$). *For multiple ($k$-D) images of a point in $\mathbb{R}^n$, non-trivial algebraic constraints involve up to $(n-k+2)$-wise views. These constraints* happen to be multilinear *and the tensor associated to the $(n-k+2)$-view relationship in fact induces all the other types of tensors associated to smaller numbers of views.*

In the classic case $n = 3, k = 2$, this corollary reduces to the well-known fact in computer vision that irreducible constraints exist up to triple-wise views, and furthermore the associated (tri-focal) tensor induces all (bi-focal) tensors (i.e. the essential matrix) associated to pairwise views. Of course, besides the examples given in Section 2.1, extra knowledge on the motion of features sometime allows us to embed the problem in a lower dimensional space. These special motions have been studied in [13], but only incomplete lists of constraints among multiple images were given. Our results here clearly *complete* such efforts and imply a much richer set of constraints, not just for point features but also for hyperplanes of any dimension. One must notice that for hyperplanes with dimension higher than 0, most algebraic constraints (as result of the above theorem) will be however *nonlinear*, especially when $n > 3$. Hence traditional multilinear analysis will no longer apply. Our approach can also be applied to much more general scenarios and capture all kinds of incidence relations among objects in a high dimensional spaces.

If there are two hyperplanes with one including the other, they give rise to the following theorem which further generalizes Theorem 1:

Theorem 2 (Rank condition with inclusion). *Consider a p_2-dimensional hyperplane H^{p_2} belonging to a p_1-dimensional hyperplane H^{p_1} in $\mathbb{R}^n$. m images $\boldsymbol{x}_i \in \mathbb{R}^{(k+1)\times(p_2+1)}$ of the H^{p_2} and m images $\boldsymbol{s}_i \in \mathbb{R}^{(k+1)\times(p_1+1)}$ of the H^{p_1} relative to the i^{th} camera frame are given ($i = 1, \dots, m$). Let the D_i's and $D_i^{\perp}$'s in the multiple view matrix M have the following values*

$$\begin{cases} D_i^{\perp} \doteq \boldsymbol{x}_i^{\perp} \in \mathbb{R}^{(k+1)\times(k-p_2)} & \text{or } \ \boldsymbol{s}_i^{\perp} \in \mathbb{R}^{(k+1)\times(k-p_1)}, \\ D_i \ \doteq \boldsymbol{x}_i \in \mathbb{R}^{(k+1)\times(p_2+1)} & \text{or } \ \boldsymbol{s}_i \in \mathbb{R}^{(k+1)\times(p_1+1)}. \end{cases} \tag{11}$$

Then for all possible instances of the matrix M, we have the two cases:

1. case one: *If $D_1 = \boldsymbol{s}_1$ and $D_i^{\perp} = \boldsymbol{x}_i^{\perp}$ for some $i \geq 2$, then*

$$\boxed{\mathit{rank}(M) \leq (n-k) + (p_1 - p_2),}$$

2. case two: *Otherwise,*

$$\boxed{0 \leq \mathit{rank}(M) \leq n-k.}$$

Since $\text{rank}(AB) \geq (\text{rank}(A) + \text{rank}(B) - n)$ for all $A \in \mathbb{R}^{m\times n}, B \in \mathbb{R}^{n\times k}$, we have $\text{rank}[(D_i^{\perp})^T \bar{R}_i D_1] \geq (p_1 - p_2)$ if the matrix $\bar{R}_i \in \mathbb{R}^{(k+1)\times(k+1)}$ is full rank for some $i \geq 2$. So the rank condition for the case one can be improved with a tight lower bound

$$\boxed{(p_1 - p_2) \leq \text{rank}(M) \leq (n-k) + (p_1 - p_2).}$$

This theorem can be easily generalized to any set of cascading hyperplanes

$$H^{p_l} \subseteq H^{p_{l-1}} \subseteq \cdots \subseteq H^{p_1},$$

for some $l \in \mathbb{Z}_+$. We omit the details for simplicity.

For two hyperplanes intersecting at a third, we have

Theorem 3 (Rank condition with intersection). *Consider hyperplanes H^{p_1}, H^{p_2}, and H^{p_3} with $H^{p_3} \subseteq H^{p_1} \cap H^{p_2}$. Given their m images relative to m camera frames as above, let the D_i's and $D_i^{\perp}$'s in the multiple view matrix M have the following values: $D_1 \doteq \boldsymbol{x}_1$, and $D_i^{\perp} \doteq \boldsymbol{x}_i^{\perp}, \boldsymbol{r}_i^{\perp}, \boldsymbol{s}_i^{\perp}$ being the coimages of $H^{p_3}, H^{p_1}, H^{p_2}$ respectively. Then we have*

$$\boxed{0 \leq \mathit{rank}(M) \leq (n-k).}$$

This theorem can be easily generalized to a family of intersecting hyperplanes

$$H^{p} \subseteq H^{p_l} \cap H^{p_{l-1}} \cap \ldots \cap H^{p_1}, \tag{12}$$

for some $l \in \mathbb{Z}_+$. We here omit the details for simplicity.

In practice, there are situations when all hyperplanes being observed belong to a p-dimensional ambient hyperplane in $\mathbb{R}^n$,[7] and the location of this ambient hyperplane

[7] Here we no longer require that p is less than k since the image of this ambient hyperplane is not of interest.

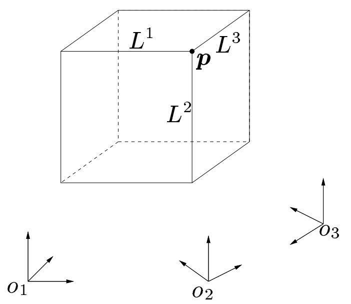
relative to the world reference frame is fixed. In general, a p-dimensional hyperplane H^p in $\mathbb{R}^n$ can be described by a full-rank $(n-p) \times (n+1)$ matrix $\Lambda \in \mathbb{R}^{(n-p)\times(n+1)}$ such that any point $\boldsymbol{p} \in H^p$ satisfies: $\Lambda \boldsymbol{X} = 0$, where $\boldsymbol{X} \in \mathbb{R}^{n+1}$ is the homogeneous coordinate of the point $\boldsymbol{p}$. We call the matrix Λ the *homogeneous representation* for H^p. Of course, such a representation is not unique – any two matrices Λ_1 and Λ_2 with the same kernel give rise to the same hyperplane. For convenience, we usually divide the $(n-k) \times (n+1)$ matrix Λ into two parts

$$\Lambda = [\Lambda^1, \Lambda^2], \quad \text{with} \quad \Lambda^1 \in \mathbb{R}^{(n-p)\times(k+1)}, \quad \Lambda^2 \in \mathbb{R}^{(n-p)\times(n-k)}. \tag{13}$$

Then we have

Theorem 4 (Rank condition with restriction). *If all the feature hyperplanes belong to an ambient hyperplane homogeneously described by the matrix $\Lambda \in \mathbb{R}^{(n-p)\times(n+1)}$, then by appending a block of rows $\left[\Lambda^1 D_1 \ \Lambda^2\right]$ to the multiple view matrix M, all the rank conditions in the above three theorems remain the same for the new matrix.*

Example 1 (Multiple views of a cube). The above theorems allow us to express all incidental constraints associated to a particular object. For instance, as shown in Figure 3, there are three edges L^1, L^2 and L^3 intersecting at each vertex $\boldsymbol{p}$ of a cube. Then from three images of the cube from three vantage points, we get the following multiple view matrix M (see Figure 3) associated to the vertex $\boldsymbol{p}$, where $\boldsymbol{x}_i \in \mathbb{R}^3$ is the image of the vertex $\boldsymbol{p}$ in the i^{th} view, and $\boldsymbol{l}_i^j \in \mathbb{R}^{3\times 2}$ is the image of the j^{th} edge L^j in the i^{th} view.[8]



$$M = \begin{bmatrix} (\boldsymbol{x}_2^{\perp})^T R_2 \boldsymbol{x}_1 & (\boldsymbol{x}_2^{\perp})^T T_2 \\ (\boldsymbol{l}_2^{1\perp})^T R_2 \boldsymbol{x}_1 & (\boldsymbol{l}_2^{1\perp})^T T_2 \\ (\boldsymbol{l}_2^{2\perp})^T R_2 \boldsymbol{x}_1 & (\boldsymbol{l}_2^{2\perp})^T T_2 \\ (\boldsymbol{l}_2^{3\perp})^T R_2 \boldsymbol{x}_1 & (\boldsymbol{l}_2^{3\perp})^T T_2 \\ (\boldsymbol{x}_3^{\perp})^T R_3 \boldsymbol{x}_1 & (\boldsymbol{x}_3^{\perp})^T T_3 \\ (\boldsymbol{l}_3^{1\perp})^T R_3 \boldsymbol{x}_1 & (\boldsymbol{l}_3^{1\perp})^T T_3 \\ (\boldsymbol{l}_3^{2\perp})^T R_3 \boldsymbol{x}_1 & (\boldsymbol{l}_3^{2\perp})^T T_3 \\ (\boldsymbol{l}_3^{3\perp})^T R_3 \boldsymbol{x}_1 & (\boldsymbol{l}_3^{3\perp})^T T_3 \end{bmatrix} \in \mathbb{R}^{10\times 2}$$

Fig. 3. Images of a standard cube from three vantage points.

The condition $\text{rank}(M) \leq 1$ then expresses the incidence condition among the vertex and three edges in terms of their three images, without explicitly referring to their 3-D location. However still more can be said. The above matrix M only captures the fact that the edges all pass the vertex in all views, it does not captures the incidence condition for the edges themselves and also not capture the fact that the four vertices are on the same plane. These constraints can be described by other (types of) multiple view matrices which are also just instances of above theorems.

[8] Traditionally the symbol $\boldsymbol{l}$ is used to describe the *coimage* of a line. In our case, that becomes $\boldsymbol{l}^{\perp}$, which is indeed a three dimensional vector.

Remark 2 (Rank values). In the above rank conditions, the highest rank value typically corresponds to a generic configuration, every value in between the highest and lowest corresponds to a different class (or type) of degenerate configurations of the object w.r.t. the camera positions (see [3]).

Remark 3 (Fixed camera). Note that in many practical situations, as we will see from examples in the following section, one typically has a *fixed* camera recording a dynamical scene. In this case, the projection matrix Π might have a column with all zeros for all views/time. Then in above theorems, the rank of the multiple view matrix should drop 1.

4 Applications and examples

In classic multiple view geometry, a primary purpose of deriving the above rank conditions or the constraints among multiple images is for a full reconstruction of camera motion as well as the location of the objects being observed. Technical conditions and algorithms for similar purposes in higher dimensional spaces are however still largely unknown. Nonetheless, in this section, we demonstrate through a few examples how information about the camera motion and scene structure is extensively encoded by the multiple view matrix and its associated rank conditions.

4.1 Multiple view stereopsis in n-dimensional space

For a p-dimensional hyperplane H^p in $\mathbb{R}^n$, if we can obtain its coimages $\boldsymbol{s}_i^{\perp}$'s, with known projection matrices $\Pi_i = \left[\bar{R}_i, \bar{T}_i\right]$, the question is "What is the least number of images we need in order to determine H^p?" In the classical 3-D space, this is known as *stereopsis*. To this end, we need to introduce another matrix

$$C = \begin{bmatrix} \left(\boldsymbol{s}_1^{\perp}\right)^T \bar{R}_1 & \left(\boldsymbol{s}_1^{\perp}\right)^T \bar{T}_1 \\ \left(\boldsymbol{s}_2^{\perp}\right)^T \bar{R}_2 & \left(\boldsymbol{s}_2^{\perp}\right)^T \bar{T}_2 \\ \vdots & \vdots \\ \left(\boldsymbol{s}_m^{\perp}\right)^T \bar{R}_m & \left(\boldsymbol{s}_m^{\perp}\right)^T \bar{T}_m \end{bmatrix} \in \mathbb{R}^{[m(k-p)]\times(n+1)},$$

which is related to M by

$$\text{rank}(C) = \text{rank}(M) + (k-p). \tag{14}$$

If G^{p+1} is the corresponding subspace in $\mathbb{R}^{n+1}$ for H^p, then $G^{p+1} \subseteq ker(C)$. Hence when the rank of M reaches its upper bound $n-k$, the rank of C reaches its upper bound $n-p$, which means that the kernel of C has dimension $p+1$. This further implies that $G^{p+1} = ker(C)$ and we can reconstruct H^p uniquely by calculating the kernel of C. On the other hand, if $\text{rank}(C) = l < n-p$, then we can only recover the hyperplane up to an $(n-l)$-dimensional hyperplane in $\mathbb{R}^n$.

Note that C is a stack of m $(k-p)\times(n+1)$ matrices $\left[(\boldsymbol{s}_i^{\perp})^T\bar{R}_i \ (\boldsymbol{s}_i^{\perp})^T\bar{T}_i\right]$, $i = 1,\ldots,m$. The kernel of the i^{th} block is the preimage $\boldsymbol{F}_i$ of $\boldsymbol{s}_i$. Hence, each view actually contributes to a reduction in the dimension of the kernel of C and $\ker(C) =$

$\cap_{i=1}^{m} \boldsymbol{F}_i$. In order to reconstruct the original subspace G^{p+1}, the dimension of the kernel should be reduced from $\dim(\boldsymbol{F}_1) = n-k+p+1$ to $p+1$. The reduction in dimension is then $n-k$. If K is the kernel of matrix C composed of $i-1$ views, then the dimension reduction of the kernel contributed by the i^{th} view is

$$\begin{aligned} \dim(K) - \dim(K \cap \boldsymbol{F}_i) &= \dim(K + \boldsymbol{F}_i) - \dim(\boldsymbol{F}_i) \\ &\leq (n+1) - (n-k+p+1) = k-p. \end{aligned} \tag{15}$$

Thus, in order to uniquely determine H^p, we need at least $m = \lceil \frac{n-k}{k-p} \rceil + 1$ views under general configuration. However, this is the "optimal" case such that each view can contribute maximum dimension reduction of the kernel of C. The maximum number of general views required is $m = (n-k+1)$ in which case each view only contributes to a one-dimensional reduction of the kernel of C. For example, in the special case for point features, we have $p = 0$. Hence the minimum number of independent views required is then $\lceil \frac{n-k}{k} \rceil + 1$. When $n = 3, k = 2$, this number reduces to 2 which is well-known for 3-D stereopsis. For the dynamical scene problem with $n = 6, k = 2$ studied in [13], in general we need 5 views to reconstruct the point unless we have an optimal configuration for which 3 views suffice.

4.2 Segmentation and formation detection

Segmentation by the rank. Now we consider the dynamical scene example we introduced in the beginning of this paper. For any point moving with up to constant acceleration, it can be described by a homogeneous coordinate in the $\mathbb{R}^{10}$: $\boldsymbol{X} = \left[X^T, v^T, a^T, 1\right]^T \in \mathbb{R}^{10}$, where $X \in \mathbb{R}^3$, $v \in \mathbb{R}^3$, and $a \in \mathbb{R}^3$ are the point's initial location, velocity, and acceleration, respectively. With respect to a fixed camera, its image $\boldsymbol{x}_i$ at time t_i then satisfies: $\lambda_i \boldsymbol{x}_i = \Pi_i \boldsymbol{X}$, where $\lambda_i \in \mathbb{R}$ and

$$\Pi_i = \left[\bar{R}_i\ \bar{T}_i\right] \in \mathbb{R}^{3\times 10}, \quad \bar{R}_i = I \in \mathbb{R}^{3\times 3}, \quad \bar{T}_i = \left[It_i\ \ It_i^2/2\ \ 0\right] \in \mathbb{R}^{3\times 7}. \tag{16}$$

Hence the associated multiple view matrix is

$$M = \begin{bmatrix} \left(\boldsymbol{x}_2^{\perp}\right)^T \boldsymbol{x}_1 & t_2 \left(\boldsymbol{x}_2^{\perp}\right)^T & \frac{t_2^2}{2} \left(\boldsymbol{x}_2^{\perp}\right)^T & 0 \\ \left(\boldsymbol{x}_3^{\perp}\right)^T \boldsymbol{x}_1 & t_3 \left(\boldsymbol{x}_3^{\perp}\right)^T & \frac{t_3^2}{2} \left(\boldsymbol{x}_3^{\perp}\right)^T & 0 \\ \vdots & \vdots & \vdots & \vdots \\ \left(\boldsymbol{x}_m^{\perp}\right)^T \boldsymbol{x}_1 & t_m \left(\boldsymbol{x}_m^{\perp}\right)^T & \frac{t_m^2}{2} \left(\boldsymbol{x}_m^{\perp}\right)^T & 0 \end{bmatrix} \in \mathbb{R}^{[2(m-1)]\times 8}.$$

From the rank condition and remark 2 we know that $\mathrm{rank}(M) < 9 - 2 = 7$.

By randomly choosing values for X, v, and a for numerous points, simulation results showed that we always have:

1. $\mathrm{rank}(M) = 4$, if the point is static, i.e. $a = 0, v = 0$.
2. $\mathrm{rank}(M) = 5$, if the point is moving with constant velocity, i.e. $a = 0$.
3. $\mathrm{rank}(M) = 6$, if the point is moving with constant acceleration, i.e. $a \neq 0$.

These results can be explained by studying the kernel of M. Note that the kernel of M always contains a trivial vector $l = \left[0\ 0_{3\times1}\ 0_{3\times1}\ 1\right]^T$. For the case that $v = 0$ and $a = 0$, we have $\boldsymbol{x}_i$=$\boldsymbol{x}_1$. So the first column of M is always 0. A basis of the kernel of M is then l, $\left[1\ 0_{3\times1}\ 0_{3\times1}\ 0\right]^T$, $\left[0\ \boldsymbol{x}_1^T\ 0_{3\times1}\ 0\right]^T$ and $\left[0\ 0_{3\times1}\ \boldsymbol{x}_1^T\ 0\right]^T$.[9] If we only have $a = 0$, then a basis for the kernel of M is l, $\left[0\ \lambda_1\boldsymbol{x}_1^T\ 2v^T\ 0\right]^T$, and $\left[\lambda_1\ v^T\ 0_{3\times1}\ 0\right]^T$. Finally for the most general case, the basis for the kernel is l and $\left[\lambda_1\ v^T\ a^T\ 0\right]^T$. The upper bound 7 for the rank is never achieved because the camera location is fixed. There will always be a one-parameter family of ambiguity for the location of a point in $\mathbb{R}^{10}$.

Hence rank conditions provide a simple method for doing segmentation or grouping of feature points in a dynamical scene. Given the correspondences of the images points at different time, we can easily differentiate background (static) points, points moving on a straight line with constant velocity and points moving with parabolic trajectory. Since we approximate the motion up to the second order, most practical scenarios studied in the literature so far [13] fall into this case.

Formation detection by the rank. Given the image correspondences for a set of points, sometimes we want to detect whether they move (relative to each other) in a similar fashion. If so, we say they move in a *formation*. As we will see, this is a more general notion than *grouping* of static points. In practice, moving together in “formation” often implies that features considered are more likely from the same object. Here we demonstrate a special case: images of four points with constant 3-D velocities, and show how the formation information can be encoded in the rank conditions. Denote the four points as $\boldsymbol{p}^1$, $\boldsymbol{p}^2$, $\boldsymbol{p}^3$ and $\boldsymbol{p}^4$, and their velocities as v^1,v^2,v^3 and v^4, respectively. The corresponding images in the i^{th} frame are $\boldsymbol{x}_i^1$, $\boldsymbol{x}_i^2$, $\boldsymbol{x}_i^3$ and $\boldsymbol{x}_i^4$. In general, the four image points can be linked to generate six virtual lines. Intersections of these lines give three virtual points on the image plane, as shown in Figure 4, which are labeled by $\boldsymbol{x}_i^5$, $\boldsymbol{x}_i^6$ and $\boldsymbol{x}_i^7$. Then we can associate multiple view matrices M^5, M^6 and M^7 to these (virtual) images of points and lines. However, since these image points may not necessarily correspond to physical points (moving or not) in 3-D space, the rank of the corresponding matrix may vary according to the relative motion among the four points. Table 1 summarizes results from simulation performed in various cases. In the table, it is clear that when the four points are coplanar and moving in that plane with the same velocity, then the additional intersection points correspond to physical points in 3-D space which are just intersections of the lines connecting the four points in 3-D. They should move with the same velocity, which is why we always obtain rank 5 for the three matrices. For the other cases, virtual images typically do *not* correspond to any physical points. Still the rank of associated matrices can tell us information about the formation of the feature points under different scenarios. These results give us simply a glimpse of

[9] The kernel can be calculated in the following way, let the vector be $V = \left[u_0, u_1^T, u_2^T, 0\right]^T \in \mathbb{R}^8$ where $u_0 \in \mathbb{R}$ and $u_1, u_2 \in \mathbb{R}^3$. If $V \in \ker(M)$, then $\left[(\boldsymbol{x}_i^\perp)^T \boldsymbol{x}_1\ (t_i\boldsymbol{x}_i^\perp)^T\ (t_i^2\boldsymbol{x}_i^\perp/2)^T\ 0\right] V = 0$ for all $i = 2, \ldots, m$. Combined with the fact that $\lambda_i\boldsymbol{x}_i = \lambda_1\boldsymbol{x}_1 + t_i v + \frac{t_i^2}{2}a$, we can write this into a “polynomial” of t_i. By setting the coefficients to zeros we can solve for u_0, u_1 and u_2.

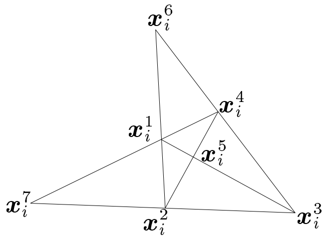


Fig. 4. Three apparent "intersections" ($\boldsymbol{x}_i^5$, $\boldsymbol{x}_i^6$ and $\boldsymbol{x}_i^7$) generated by the four image points.

Motion formations	Spatial formations	rank(M^5)	rank(M^6)	rank(M^7)
$v^1 = v^2 = v^3 = v^4$	Coplanar	5	5	5
$v^1 = v^3, v^2 = v^4 = kv^1$ $k \in \mathbb{R}$ and $k \neq 1$	Coplanar	5	6	6
$v^1 = v^2 = v^3$	Coplanar	6	6	6
$v^1 = v^2 = v^3 = v^4$	Not coplanar	6	6	6

Table 1. Relationship between the rank of the multiple view matrices associated with the virtual image points and the spatial & motion formations of the four points in 3-D space.

how rich 3-D information we may gain by merely playing with the rank of the multiple view matrix. Although our current results do not provide an analytical explanation to the relationship between the formation and the rank conditions, at least they give some necessary conditions which can be used at early stage of establishing correspondence, grouping, or segmentation. We believe a thorough study will lead to fruitful theoretical results.

4.3 Distinguish corners and occluding T-junctions by the rank

A fundamental problem which troubles structure from motion is that feature points, as "corners" or "T-junctions", extracted from the image do not necessarily correspond to physical points in the 3-D world. For example, in Figure 5, three blocks are on a table. Blocks 1 and 2 are adjacent to each other hence the corner $\boldsymbol{p}$ can be treated as a real 3-D point. However, the "point" $\boldsymbol{q}$ appears in the image as the occluding T-junction of the two edges L^1 and L^2. It does not correspond to any physical point in 3-D space. The question is: "Can we extract some visual information from a moving scene so that these two types of feature points can be distinguished?" Suppose that the table is rotating around its normal, denoted as the X-axis, with angular velocity ω. The axis intersects the table at a point $[0, Y_0, Z_0]^T$ (where Y_0 and Z_0 are unknown). Then for any point with initial coordinate $[X, Y, Z]^T$, its image $\boldsymbol{x}(t)$ at any time t satisfies

$$\lambda(t)\boldsymbol{x}(t) = \Pi(t)\bar{\boldsymbol{X}} = \begin{bmatrix} 1 & 0 & 0 & 0 & 0 & 0 \\ 0 & 1 & 0 & \cos(\omega t) & -\sin(\omega t) & 0 \\ 0 & 0 & 1 & \sin(\omega t) & \cos(\omega t) & 0 \end{bmatrix} \bar{\boldsymbol{X}}, \tag{17}$$

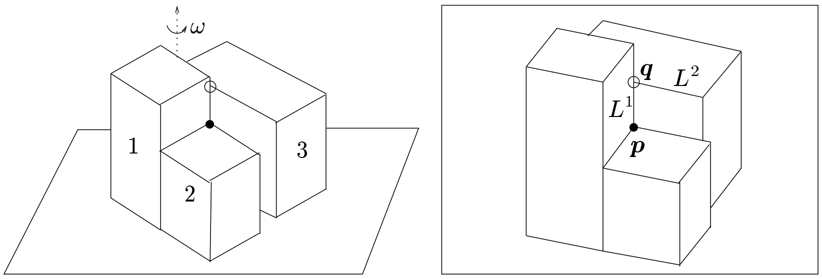


Fig. 5. Three rectangular blocks on a table with blocks 1 and 2 adjacent to each other. Point $\boldsymbol{p}$ is the conner formed by blocks 1 and 2. Point $\boldsymbol{q}$ appears on image as the occluding T-junction of lines L^1 and L^2. In 3-D space these two lines do *not* intersect.

where $\bar{\boldsymbol{X}} = \left[X, Y_0, Z_0, \tilde{Y}, \tilde{Z}, 1\right]^T \in \mathbb{R}^6$ with $\tilde{Y} = Y - Y_0$, and $\tilde{Z} = Z - Z_0$. Under this setup, the image can be viewed as a projection from $\mathbb{R}^5$ to $\mathbb{R}^2$. Since the projection matrix always has the last column being 0, the rank for the multiple view matrix associated to any point (in $\mathbb{R}^5$) should be strictly less than 3 (again, because the camera is not moving). Simulation results demonstrate that for a sufficient number of images we have:[10]

1. $\text{rank}(M) = 2 < 3$ for images of the point $\boldsymbol{p}$,
2. $\text{rank}(M) = 3$ for virtual images of the T-junction $\boldsymbol{q}$.

This example shows that at least in some cases the rank condition can serve as a criterion for determining whether or not a "feature point" in the image actually corresponds to some physical point in 3-D (from their motion). It may provide a means to reject T-junctions which are the result of occlusion, like the point $\boldsymbol{q}$ in Figure 5.

5 Summary

The main result in this paper is the presentation of generalized rank conditions associated to the multiple view matrix for perspective projection from $\mathbb{R}^n$ to $\mathbb{R}^k$. These conditions provide a complete set of intrinsic constraints that govern multiple images of objects in high dimensional spaces. The theory is general enough to enable geometric analysis for many dynamical scenes (after embedded into a higher dimensional Euclidean space), as an extension to classic multiple view geometry. In addition to its potential for purposes such as recovering camera motion and scene structure and dynamics, many new problems and phenomena arise in the setting of dynamical scenes from broad applications of the multiple view matrix and its rank conditions. They include (but not limited to): stereopsis, segmentation and grouping, formation detection,

[10] Note that at time $t = 0$, the projection matrix is not in the standard form $[I, 0]$, so we need to multiply another matrix to each $\Pi(t)$ to obtain this form, and the multiple view matrix is also modified accordingly.

and occlusion detection. A full geometric and algebraic characterization for these problems and phenomena remains largely open.

In this paper we did not address at all how to use such rank conditions to facilitate the recovery of camera motion and scene dynamics. But it provides a systematic way to eliminate redundant parameters and reduce constraints among image sequences to its (almost) minimum. Further estimation of unknown parameters using either tensorial techniques or direct minimization is a matter of algorithm design. In addition to such theoretical endeavor, we are currently conducting experiments on videos of multiple moving objects (mobile robots) as well as for the purpose of tracking and estimating human body movement.

References

1. S. Avidan and A. Shashua. Novel view synthesis by cascading trilinear tensors. *IEEE Transactions on Visualization and Computer Graphics (TVCG)*, 4(4), pp.293-306, 1998.
2. O. Faugeras, Q.-T. Luong, and T. Papadopoulo. *Geometry of Multiple Images*. The MIT Press, 2001.
3. R. Fossum, K. Huang, Y. Ma General rank conditions in multiple view geometry. *UIUC, CSL Technical Report, UILU-ENG 01-2222 (DC-203)*, October 8, 2001.
4. R. Hartley. Lines and points in three views - a unified approach. In *Proceedings of 1994 Image Understanding Workshop*, pp. 1006–1016, Monterey, CA USA, 1994. OMNIPRESS.
5. R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, 2000.
6. A. Heyden and K. Åström. Algebraic properties of multilinear constraints. *Mathematical Methods in Applied Sciences*, 20(13), pp.1135-1162, 1997.
7. Y. Liu and T.S. Huang. Estimation of rigid body motion using straight line correspondences *IEEE Workshop on Motion: Representation and Analysis*, Kiawah Island, SC, May 1986.
8. H. C. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293, pp.133-135, 1981.
9. Y. Ma, K. Huang, and J. Košecká. New rank deficiency condition for multiple view geometry of line features. *UIUC, CSL Technical Report, UILU-ENG 01-2209 (DC-201)*, May 8, 2001.
10. Y. Ma, K. Huang, R. Vidal, J. Košecká, and S. Sastry. Rank conditions of multiple view matrix in multiple view geometry. *UIUC, CSL Technical Report, UILU-ENG 01-2214 (DC-220)*, submitted to IJCV, June 18, 2001.
11. M. Spetsakis and Y. Aloimonos. Structure from motion using line correspondences. *International Journal of Computer Vision*, 4(3):171–184, 1990.
12. B. Triggs. Matching constraints and the joint image. In *Proceedings of Fifth International Conference on Computer Vision*, pp.338-343, Cambridge, MA, USA, 1995. IEEE Comput. Soc. Press.
13. L. Wolf and A. Shashua. On projection matrices $\mathcal{P}^k \rightarrow \mathcal{P}^2, k = 3, \dots, 6$, and their applications in computer vision. In *Proceedings of the Eighth International Conference on Computer Vision*, pp.412-419, Vancouver, Canada, 2001.
14. G. Sparr. A common framework for kinetic depth, reconstruction and motion for deformable objects. *Proceedings of the Fourth European Conference on Computer Vision*, pp.471-482, Cambridge, England, 1994.