Classification of rank conditions for multiple views of dynamical scenes

Yi Ma Kun Huang Yang Yang
Electrical & Computer Engineering Department
University of Illinois at Urbana Champaign

Abstract

In this paper we demonstrate the possibility of using *rank conditions* as a guiding principle to unify the study of multiple view geometry, for either static or dynamical scenes. A way of classifying rank conditions is proposed for various cases that may arise in the more general context where multiple views of a dynamical scene are considered. We demonstrate by concrete examples why a few doctrines in classic multiple view geometry should be updated in this new context and how. Not all results presented here are entirely new (although mostly recent), nor all belong to us. The purpose of this paper is to provide a brief overview all rank related issues (old or new) as well as their potential impact on multiple view geometry, in either theory or practice, for either static or dynamical scenes.

Keywords: multiple view geometry, rank conditions, dynamical scenes

1 Introduction: about rank conditions

Multi-focal tensor analysis has long been proposed as a basic (mathematical) tool for analyzing multiple images, with an especial emphasis on breaking image sequences or sets into relations among pairwise and triple-wise (and sometimes quadruple-wise) views (see the representative manuscripts [Hartley and Zisserman 2000; Faugeras et al. 2001]). Nevertheless, in most algorithms of practical importance that work for multiple images or image sequences, certain (global) rank condition (based factorization) is typically utilized for purposes such as feature tracking or even 3-D reconstruction. These rank based methods have been demonstrated mostly effective in cases of approximated camera models such as orthographic [Tomasi and Kanade 1992], affine [Quan and Kanade 1996], para-perspective [Poelman and Kanade 1997], weak-perspective (even for non-rigid motion) [Irani 1999; Torresani et al. 2001], just to name a few. The main reason why such rank based methods are favored is largely due to the numerical stability of the rank conditions and that such methods in principle can apply to all available images simultaneously without breaking a sequence or set into pairwise or triple-wise views.[1]

Surprisingly, (global) rank conditions for the exact perspective case have been investigated, understood, or utilized only to a limited extent, although there has been some primitive rank based factorization methods (e.g., [Triggs 1996]). Recently, it has been shown that such factorization methods can be very much improved by a more systematic study of rank conditions associated to an exact perspective camera model [Ma et al. 2001]. A universal rank condition indeed exists for all kinds of features such as points, lines, and planes (and even curves), and for all types of projections too. In addition, the so-derived rank conditions could easily replace the roll of multi-focal tensors in analyzing multiple images and dramatically simplify the study of multiple view geometry, both in theory and in algorithm.[2] In this paper, we intend to go one step further and make a provocative stand that

> **intrinsic constraints encountered in multiple view geometry are determinantal.**[3]

That is, constraints among multiple images can always be derived *directly* from certain rank conditions on a matrix, regardless of their nature. The word "directly" is important since it requires that the size of the matrix must be minimum so that its rank can no longer be further reduced. Although it is difficult to believe already that the above statement would be ultimately true, we in this paper will use it as a guiding principle and demonstrate how to unify most known or some unknown relationships among multiple images of a static or dynamical scene.[4] In any case, a few advantages and surprises from such a seemingly pre-mature belief will soon become evident:

1. *It allows us to easily categorize and study* all *intrinsic constraints (to be defined later) at least for point feature in spaces of arbitrary dimensions which will therefore allow us to extend existing multiple view geometry to cases when multiple moving objects, deforming surfaces, or non-rigid motions are present (the so-called dynamical scenes). In particular it extends and simplifies some previous understanding on this subject (e.g., [Wolf and Shashua 2001]).*

2. *It allows us to study certain non-conventional incidence relations among multiple images of point and line features. For instance, we will show how to derive using rank technique non-trivial intrinsic constraints for the case of linear trajectory triangulation, previously studied by [Avidan and Shashua 1999]. The results give a simple algebraic explanation why up to 9 images should be involved in such constraints.*

3. *It allows us to demonstrate that the richness of geometry that more complicated features such as curves may give rise to. In particular, we will show constructively that depending on the class of features considered, non-trivial irreducible constraints among multiple images may involve arbitrary number of views (in a rather classic sense!).*

All these results will be derived by using a single technique:

> **reducing the relations among multiple images to certain rank condition on a matrix.**

This paper intends to provide a conceptual understanding of rank related geometry and algebra that may arise in the context of multiple views of a static or dynamical scene. In current draft, we do not yet perform any sort of simulation or experiment. The statements themselves speak for the importance of such an understanding. Not all results presented here are entirely new (although mostly recent), nor all belong to us, but we believe that putting them together in

[1] It is also true that such methods can handle occlusions and outliers to some extent.

[2] For instance, the estimation and use of trifocal tensor is no longer necessary even for purposes such as feature matching, image transfer, and motion and structure reconstruction, see [Ma et al. 2001].

[3] This unavoidably relates multiple view geometry with a very important branch of study in modern Algebraic Geometry: determinantal varieties, see [Harris 1992; Room 1938].

[4] We however do encourage people to provide an intuitive counter example if any!

a coherent framework means more than just supporting the above claims.[5] The purpose of this paper is to give an overview on rank related issues (old or new) as well as their potential impact on multiple view geometry, in either theory or algorithm, for either static or dynamical scenes.

2 Linear dynamical scenes

In classic multiple view geometry, people 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

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

where $\boldsymbol{Y} \in \mathbb{R}^4$ is the homogeneous representation of the 3-D coordinates of a feature point p (relative to a world coordinate frame),[6] $\boldsymbol{x}(t) \in \mathbb{R}^3$ 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 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.[7]

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

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

Since now $\boldsymbol{Y}(t)$ is too time-varying, all results from classic multiple view geometry no longer apply. However, suppose the temporal variation of $\boldsymbol{Y}(t)$ is not completely random. A simple but important case is when $\boldsymbol{Y}(t)$ can be described by an underlying linear system such that

$$\begin{aligned}\dot{\boldsymbol{X}}(t) &= A\boldsymbol{X}(t), \quad \boldsymbol{X} \in \mathbb{R}^{n+1},\ A \in \mathbb{R}^{(n+1)\times(n+1)}, & (3)\\ \boldsymbol{Y}(t) &= C\boldsymbol{X}(t), \quad \boldsymbol{Y}(t) \in \mathbb{R}^4,\ C \in \mathbb{R}^{4\times(n+1)}. & (4)\end{aligned}$$

Such linear dynamics can be used to model a rather large range of multiple moving objects or even non-rigid motions, e.g., water, smoke, multi-body links [Soatto et al. 2001; Huang et al. 2002]. Such modeling is a matter of system identification which we will discuss briefly later.

Comment 1 *We here use a homogeneous representation for the system state $\boldsymbol{X}$. For an n-dimensional linear system $\dot{X} = AX,\ X \in \mathbb{R}^n$, we can always augment the state as $\boldsymbol{X} = [X^T, 1]^T \in \mathbb{R}^{n+1}$ and obtain a new (homogeneous) linear system*

$$\dot{\boldsymbol{X}} = \begin{bmatrix} A & 0 \\ 0 & 0 \end{bmatrix} \boldsymbol{X}.$$

From the theory of ordinary differential equations (ODE),[8] we know that we can write down the solution of $\boldsymbol{Y}(t)$ in terms of the initial value of $\boldsymbol{X} = \boldsymbol{X}(0) \in \mathbb{R}^{n+1}$ as $\boldsymbol{Y}(t) = Ce^{At}\boldsymbol{X}$. Substitute this into the equation (2) we get $\lambda(t)\boldsymbol{x}(t) = \Pi(t)Ce^{At}\boldsymbol{X}$. If we call the matrix $\Pi(t)Ce^{At} \in \mathbb{R}^{3\times(n+1)}$ as the new projection matrix $\bar{\Pi}(t)$, i.e.

$$\bar{\Pi}(t) \doteq \Pi(t)Ce^{At}, \tag{5}$$

we obtain a perspective projection from $\mathbb{R}^n$ to $\mathbb{R}^2$

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

represented by homogeneous coordinates. Note that here two things potentially contribute to the time-variation of $\bar{\Pi}(t)$: the camera motion $\Pi = [R, T]$ and the dynamics of the point $\dot{\boldsymbol{X}} = A\boldsymbol{X}$. The above equation is of the same form as (1) except that it represents a (perspective) projection from the space $\mathbb{P}^n$ to $\mathbb{P}^2$. It is then natural to expect that results in classic multiple view geometry should generalize to this class of projections as well. We will refer to the above process as a (Euclidean) *embedding* of a dynamical scene into a space of higher dimension.

Comment 2 (Static scene as a special case) *If we let $n = 3, A = 0$ and $C = I_{4\times 4}$ in the above model, we simply get back to the classical perspective projection model studied in the computer vision literature. Hence the above model (6) can be viewed as a (dynamical) generalization of the classical model. Its advantage is yet to be revealed later.*

3 Perspective observability

More generally, we can consider a similar but generalized projection model from $\mathbb{P}^n$ to $\mathbb{P}^k$ as

$$\lambda(t)\boldsymbol{x}(t) = \bar{\Pi}(t)\boldsymbol{X} \tag{7}$$

where $\boldsymbol{x} \in \mathbb{R}^{k+1}$ and $\boldsymbol{X} \in \mathbb{R}^{n+1}$.

Definition 1 (Homogeneous equality) *We say two non-zero vectors $u, v \in \mathbb{R}^m$ are "homogeneously equal" if $u = \alpha v$ for some scalar $\alpha \in \mathbb{R}$, denoted as $u \simeq v$.*

Definition 2 (Perspective observability) *We say that the model (7) is* perspectively unobservable *if there exist two (initial) states $\boldsymbol{X}^1 \not\simeq \boldsymbol{X}^2$ in $\mathbb{R}^{n+1}$ which may give rise to the same sequence of image $\boldsymbol{x}^1(t) \simeq \boldsymbol{x}^2(t) \in \mathbb{R}^{k+1},\ \forall t \in [0, t_f]$. The model is called* observable *if no such states exist.*

Having unobservable states is certainly an undesirable property for any kind of modeling. We hence would like to know for the general model (7), under what conditions such a (bad) thing will or will not occur.

As an extension to classic linear systems theory, the perspective observability issue has been extensively studied by Ghosh et. al. [Ghosh and Loucks 1995] when the projection matrix $\bar{\Pi}(t)$ is induced from a linear system: $\bar{\Pi}(t) = Ce^{At}$. If so, a generalized Popov-Belevitch-Hautus rank criterion can be used to verify the observability:

Proposition 1 (Generalized PBH rank test) *Assume $\bar{\Pi}(t) = Ce^{At}$ in the projection model (7). Then it is* perspectively observable *if*

$$\boxed{\textit{rank}\begin{bmatrix}(A-\lambda_i I)(A-\lambda_j I)\\ C\end{bmatrix} = n+1}$$

for every pair of eigenvalues (λ_i, λ_j) (maybe the same) of A. This condition is also necessary if all the eigenvalues of A are real.

Please see [Ghosh and Loucks 1995] for a proof. The reader can easily verify for the special case when A is a diagonal matrix. Note that this test only applies to image sequence captured by a monocular camera. For more general scenarios we have:

[5] The use of rank based methods to characterize relations among multiple images has a long history, although it has never been formally established as a framework, despite some earlier effort e.g., [Sparr 1994].

[6] We here use $\boldsymbol{Y}$ to save $\boldsymbol{X}$ for a new use.

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

[8] Recall that the solution to the ordinary differential equation $\dot{\boldsymbol{X}}(t) = A\boldsymbol{X}(t)$ can be written as $\boldsymbol{X}(t) = e^{At}\boldsymbol{X}(0)$ where $e^{At} \doteq I + tA + \frac{t^2}{2}A + \cdots + \frac{t^n}{n!}A + \cdots$.

Proposition 2 (A necessary and sufficient condition) *Let $\bar{\pi}_i^T(t) \in \mathbb{R}^{n+1}$ be the i^{th} row of $\bar{\Pi}(t) \in \mathbb{R}^{(k+1)\times(n+1)}$. The projection model (7) is perspectively observable if and only if, over time $t \in [0, t_f]$ the skew-symmetric matrices below*

$$\bar{\pi}_i(t)\bar{\pi}_j^T(t) - \bar{\pi}_j(t)\bar{\pi}_i^T(t) \ \in \mathbb{R}^{(n+1)\times(n+1)}, \quad i, j = 1, \dots, k$$

span the entire space of $(n+1)\times(n+1)$ skew-symmetric matrices, i.e. they span the entire lie algebra $so(n+1)$ of the rotation group $SO(n+1)$.

Proof: See [Ma et al. 2002] ■

Notice that the above condition easily generalizes to the situation where more than multiple views (of the same moving points) are taken at the same time. All we need to do is to account for all the skew-symmetric matrices generated by all the projection matrices $\bar{\Pi}_1(t), \dots, \bar{\Pi}_m(t)$ associated to say m views.

Let us take a closer look at the perspective observability in a multiple view setting. For instance, multiple *fixed* cameras are filming the same moving scene. Now assume the projection matrix associated to the i^{th} camera takes the form $\bar{\Pi}_i(t) = \Pi_i C e^{At}$ where $\Pi_i \in \mathbb{R}^{3\times 4}, C \in \mathbb{R}^{4\times(n+1)}, A \in \mathbb{R}^{(n+1)\times(n+1)}$, what kind of conditions do we have for the perspective observability then? From our experience with 3-D vision, we know that if there are more than two views of a 3-D scene (dynamical or not) the projection model would always be observable as long as the pair (A, C) is. Hence observability is really not a serious issue in the multiple view setting as long as one is willing to pay the price for establishing exact (or sometimes can be relaxed) correspondence across multiple views and images. A much more important issue then is how to determine the system parameters (A, C) and the camera locations $\Pi = \{\Pi_1, \dots, \Pi_m\}$ from all the corresponding trajectories of points and lines etc from those images.

4 Multiple view intrinsic constraints

Essentially the problem of recovering (Π, A, C) is a systems identification problem. Despite a vast body of literature on systems identification, little is genuinely useful for our special case here. Hence we must resort to the help from multiple view geometry. A key step towards this goal is to identify all constraints that multiple views of a static or dynamical scene may have to offer. We are particularly interested in the so-called "intrinsic constraints", to be precise:

Definition 3 (Irreducible intrinsic constraint) *A constraint among multiple images is called* intrinsic *if it can be expressed in terms of an algebraic (i.e. polynomial) equation of the form $f(\Pi, \boldsymbol{x}) = 0$ where $\boldsymbol{x}$ denotes image quantities (maybe from multiple images) and Π the (multiple) camera locations. A constraint is* irreducible *if it cannot be expressed in terms of constraints that involve strictly less number of images.*[9]

The main reason why we are interested in this type of equations is because they directly relate information about the camera and scene motion with measurements on the images.

As we have known, all intrinsic constraints in classic multiple view geometry are "determinantal", i.e. they come from certain rank conditions on various types of matrices, namely

$$\text{rank}(M(\Pi, \boldsymbol{x})) \leq k, \quad M(\Pi, \boldsymbol{x}) \in \mathbb{R}^{m\times n}$$

[9]The class of intrinsic constraints (with respect to which the irreducibility is considered) depend on our prior knowledge imposed or assumed upon the class of image features considered.

for some $k < \min\{m, n\}$ (see [Ma et al. 2001] for a proof). Determinants of minors of the matrix M then *directly* give rise to all types of intrinsic constraints, such as epipolar constraints, and trilinear constraints etc. In a dynamical setting, the situation becomes much more complicated. New types of determinantal constraints arise together with constraints which are sometimes hard to describe in terms of traditional matrix rank. Nevertheless, we have identified three types of intrinsic constraints below and intend to support the claim:

Thesis 1 *Intrinsic constraints encountered in multiple view geometry are determinantal.*

These three types of constraints are classified according to how strongly correspondence needs to be established among multiple views or image sequences; and consequently, how much information they may provide about the camera and scene motion.

4.1 Type one: corresponding dynamical views

First suppose we have multiple views $\boldsymbol{x}_1(t), \dots, \boldsymbol{x}_m(t)$ of a point p at time t. Notation wise, we will often put a "wide hat" ($\widehat{\ }$) on a vector $u \in \mathbb{R}^3$ to indicate a skew symmetric matrix $\widehat{u} \in \mathbb{R}^{3\times 3}$ such that $\widehat{u}v = u \times v$ for all $v \in \mathbb{R}^3$.[10] Then the matrix below

$$W_p(t) \doteq \begin{bmatrix} \widehat{\boldsymbol{x}_1(t)}\Pi_1 \\ \vdots \\ \widehat{\boldsymbol{x}_m(t)}\Pi_m \end{bmatrix} \quad \in \mathbb{R}^{3m\times 4} \tag{8}$$

satisfies the rank condition

$$\boxed{\text{rank}(W_p(t)) \leq 3,} \tag{9}$$

for all t since the kernel would be exactly $\boldsymbol{Y}(t) = Ce^{At}\boldsymbol{X}$ if enough number of views are taken so that rank$(W_p(t)) = 3$.[11] This rank condition was initially recognized in [Faugeras and Mourrain 1995]. In our context, we may regard this case as if the scene is static since a "synchronized" correspondence among multiple views has been established at each time t. According to [Ma et al. 2001], one may further reduce the above rank condition and obtain a more compact multiple view matrix,[12] from which we immediately have:

Proposition 3 (Classic intrinsic constraints) *The rank condition rank$(W_p) \leq 3$ gives rise to non-trivial intrinsic constraints involving $2 \leq m \leq 3$ images (of point features). Non-trivial constraints exist up to 4 images if mixed point and line features are considered.*

This is somewhat a good news since if we have the correspondence, 1. the recovery of Π can be decoupled from (A, C); 2. corresponding images at different time can be treated as images of different points at the same time! Hence solving the reconstruction problem from multiple views for a scene with linear dynamics is really a multiple view geometry problem for a static scene in the traditional sense [Hartley and Zisserman 2000; Faugeras et al. 2001], followed by a standard linear system identification problem.

[10]The other notation $[u]_\times$ is also often used in the computer vision literature.

[11]The only difference between the matrix W_p and the so-called multiple view matrix M_p in [Ma et al. 2001] is M_p is W_p further reduced with respect to a reference (usually the first) image. Hence M_p is more compact and allows direct factorization and computation of relative camera configurations. But algebraically the rank conditions on W_p and M_p are equivalent.

[12]W_p is not yet factorizable with respect to the camera motion Π hence further reduction is needed. The idea to reduce W_p is by fixing any camera frame to be the reference [Ma et al. 2001].

4.2 Type two: one dynamical sequence

Now suppose we take multiple images (from a sequence) $\boldsymbol{x}(t_1), \dots, \boldsymbol{x}(t_m)$ of a point p from a single view at different times $t_1, t_2, \dots, t_m$. In this case we have $\bar{\Pi}(t) = \Pi C e^{At}$.[13] Then it is direct to show that the matrix below

$$\bar{W}_p \doteq \begin{bmatrix} \widehat{\boldsymbol{x}(t_1)}\bar{\Pi}(t_1) \\ \vdots \\ \widehat{\boldsymbol{x}(t_m)}\bar{\Pi}(t_m) \end{bmatrix} \in \mathbb{R}^{3m\times(n+1)} \tag{10}$$

satisfies the rank condition

$$\boxed{\operatorname{rank}(\bar{W}_p) \leq n,} \tag{11}$$

and the kernel would be exactly $\boldsymbol{X}$ if enough number of images are used here so that $\operatorname{rank}(\bar{W}_p) = n$. This is the type of rank conditions that have been studied previously in [Huang et al. 2002]. Notice that it is truly of the same nature as the type one except that it is for a projection in a higher dimensional space. It is this type of rank conditions that give rise to constraints used in [Wolf and Shashua 2001] for a few interesting dynamical cases. But more generally we have:

Proposition 4 (Intrinsic constraints from dynamics) *Irreducible intrinsic constraints arise from image sequence of (a point of) an n-dimensional dynamical system (i.e. the rank condition rank$(\bar{W}_p) \leq n$) involves up to n images.*

This simple rank condition nevertheless easily explains and further completes constraints given in [Wolf and Shashua 2001] and it can even be extended to study subspaces of more than zero dimension (i.e. lines, planes etc). If images of lines or subspaces of higher dimension in $\mathbb{R}^n$ are considered, irreducible constraints may involve more than n images [Huang et al. 2002]. Note that in order to utilize this type of constraints, tracking of each individual point within an image sequence is required.

4.3 Type three: multiple views of a trajectory

There is also the case when we know the motion of the (feature) point up to certain class of trajectories but do not have a precisely synchronized or timed dynamical model as above two cases. For example, we may know that a point is moving on a straight line but we do not know if it is moving with constant speed or with acceleration. Nevertheless, such knowledge may still imposes non-trivial constraints on the images and camera configurations. This type of constraints are the most intriguing, not particularly easy to utilize, but the least studied in the literature. For simplicity, in this paper, we will however restrict the trajectories to those of linear dynamical systems.[14] We do not pretend we are able to give a complete study of all such trajectories at this point. We only demonstrate here a few interesting examples and show how intrinsic constraints among such images can be derived from rank techniques.

Remark 1 (Beyond four images) *It is widely perceived in computer vision community that irreducible intrinsic constraints among multiple images exist only up to four views. This is however only partially true. It is mostly true if we only consider geometric primitives like points, lines, and planes. It is far from true if a more general class of primitives are considered. As we will demonstrate below with a class of curves, irreducible intrinsic constraints may exist up to any number of views. The rank condition will still be an effective tool to characterize such intrinsic constraints.*

4.3.1 Rank condition for linear trajectory triangulation

Consider the scenario where a point is moving on a straight line L and multiple images of this point are taken at different vantage points.[15] See Figure 1. This is a case studied as "trajectory triangulation" previously by [Avidan and Shashua 1999]. Here let us revisit this problem from the rank based approach and see if we can get better insight to it.

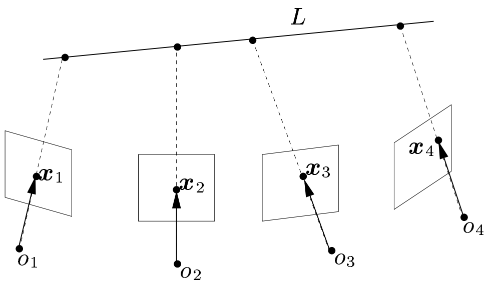

Figure 1: Images of points on a straight line L are taken at different vantage points. In fact there is always a line across any given four lines. Hence a fifth image is needed to impose nontrivial constraints.

Suppose the line L is described by a rank 2 matrix $\pi \in \mathbb{R}^{2\times 4}$ such that points $\boldsymbol{Y}$ on the line satisfy $\pi\boldsymbol{Y} = 0$. It is easy to see that the matrix $\begin{bmatrix} W_p \\ \pi \end{bmatrix} \in \mathbb{R}^{(3m+2)\times 4}$ in general has full rank 4, which gives null algebraic constraints (on entries of the matrix). Nevertheless, after a closer look at the internal structure of the above matrix, we realize that its sub-matrices have the property

$$\operatorname{rank}\begin{bmatrix} \widehat{\boldsymbol{x}_i}\Pi_i \\ \pi \end{bmatrix} \leq 3, \quad i = 1, \dots, m \tag{12}$$

since the ray from each image point and the line L in 3-D intersect.[16] This condition in fact imposes some not so trivial constraints among the multiple images $\boldsymbol{x}_1, \dots, \boldsymbol{x}_m$ and the camera configuration $\Pi_1, \dots, \Pi_m$. After the extrinsic parameters (i.e. π) of the line are eliminated, as we will show below, one in fact will get an effective rank condition on a matrix of a proper dimension. It will then be easy to see from such condition that non-trivial constraints will be among 5 – 9 views instead of 2 – 4.

Choose a base point $\boldsymbol{Y}_o = [X_o, Y_o, Z_o, 1]^T \in \mathbb{R}^4$ on the line L and let $\boldsymbol{V} = [v^T, 0]^T, v \in \mathbb{R}^3$ be the direction of the line, it is direct to see that the condition (12) is equivalent to $\widehat{\boldsymbol{x}_i}\Pi_i(\boldsymbol{Y}_o + \lambda_i \boldsymbol{V}) = 0, \lambda_i \in \mathbb{R}$ for $i = 1, \dots, m$.[17] Call $\boldsymbol{y}_i(v) \doteq -\widehat{\boldsymbol{x}_i}\Pi_i\boldsymbol{V}$, then we get the perspective projection like equation $\lambda_i \boldsymbol{y}_i(v) = \widehat{\boldsymbol{x}_i}\Pi_i\boldsymbol{Y}_o$ which can be used to eliminate extrinsic parameters $\boldsymbol{Y}_o$ and λ_i. The result is, again, a rank deficient matrix

$$U_p \doteq \begin{bmatrix} \widehat{\boldsymbol{y}_1(v)}\widehat{\boldsymbol{x}_1}\Pi_1 \\ \vdots \\ \widehat{\boldsymbol{y}_m(v)}\widehat{\boldsymbol{x}_m}\Pi_m \end{bmatrix} \in \mathbb{R}^{3m\times 4} \tag{13}$$

which has $\boldsymbol{Y}_o$ in its kernel. Notice that the matrix U_p is still a function of extrinsic parameters v. We hence denote it as $U_p(v)$.

[13] Since here the camera is fixed, the matrix $\Pi(t)$ is constant Π.

[14] For example, straight lines, circles, polynomial curves that can be realized as trajectories of linear dynamical systems.

[15] If the point is moving at a constant speed or acceleration, it is essentially reducible to a case for the second type of rank conditions.

[16] The kernel of the sub-matrix is exactly the point where they intersect.

[17] Note that $\boldsymbol{Y}_o + \lambda_i V$ are homogeneous coordinates of points on the line L.

It is easy to see that if $v = [v_1, v_2, v_3]^T$ and $v_3 \neq 0$, the matrix $U_p(v)$ can be written as

$$U_p(v) \doteq \tilde{W}_p \begin{bmatrix} v & 0 & -\frac{v_1}{v_3}v & 0 \\ 0 & v & -\frac{v_2}{v_3}v & 0 \\ 0 & 0 & 0 & v \end{bmatrix}$$

for a matrix $\tilde{W}_p \in \mathbb{R}^{3m\times 9}$ which depends only on the images $\boldsymbol{x}_i$'s and camera positions Π_i's.[18] Since $\mathrm{rank}(U_p) \leq 3$, it yields that the matrix $\tilde{W}_p$ is of

$$\boxed{\mathrm{rank}(\tilde{W}_p) \leq 8.} \tag{14}$$

Therefore determinants of (9×9 minors of) $\tilde{W}_p$ give constraints among 5 to 9 images. Note that the three rows associated to each image $\boldsymbol{x}_i$ in $\tilde{W}_p$ are trivially linearly dependent simply because the matrix $\widehat{\boldsymbol{y}}$ is always of rank 2. Hence in order to pick 9 non-trivial rows, we must need at least 5 images.

The geometry of these constraints can be described in terms of the so-call "ruled surface" determined by three lines in general position: the family of lines each of which intersects with three given lines form a 2-D surface (see Figure 2). Since such a surface can be parameterized by a family of lines, it is called a *ruled surface*. Another 3-D line in general intersects with such a surface (typically

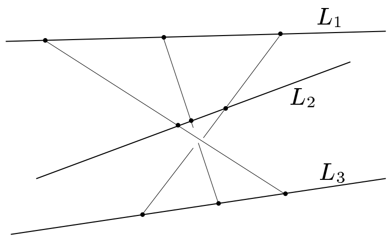

Figure 2: A ruled surfaced formed by lines intersecting at three given lines.

at two points). Hence there is at least one line which intersects with any 4 image rays and constraint among them is null. So a 5^{th} image is needed as the rank condition predicts. This necessary condition was first observed by [Avidan and Shashua 1999]. Nevertheless, the rank condition also reveals *other* non-trivial relations: including those that could have involved up to 9 images. The reason why 9 images should be involved is because the line passing through image rays 1-5 is not necessary the same one passing through rays 5-9. The role of such constraints among more than 5 images is just like that of the trilinear constraints in the classic multiple view geometry: Here they guarantee the two lines (each determined by 5 image rays) coincide while the trilinear constraints force the two points (each determined by two rays that satisfy epipolar constraints) to coincide. It is then easy to verify (by tracing back the above derivation) that the upper rank bound $\mathrm{rank}(\tilde{W}_p) = 8$ is indeed both necessary *and sufficient* for the existence of a unique line passing through any number of image rays. We hence have proved:

Proposition 5 (Intrinsic constraints for linear trajectories) m *images in different views correspond to points on a straight line in 3-D iff rank*$(\tilde{W}_p) \leq 8$*, which implies non-trivial constraints among 5 to 9 views.*

4.3.2 Irreducible algebraic relations among arbitrary number of views

Exercising the same philosophy, in principle we can construct irreducible intrinsic constraints among *any number of views*. Consider a class of (generalized spiral) curves specified in terms of 3-D coordinates

$$\gamma_\alpha : \; s \in \mathbb{R}_+ \;\mapsto\; [x, y, z] \doteq \left[r\sin(s), r\cos(s), \sum_{i=0}^{n} \alpha_i s^i \right].$$

One such spiral is shown in Figure 3. Notice that a straight line can be viewed as the degenerate case as $r = 0$. For simplicity, here we assume that $r \in \mathbb{R}_+, n \in \mathbb{Z}_+$ are given and fixed but the coefficients $\alpha = \{\alpha_i \in \mathbb{R}\}$ could vary among the class of spirals. Suppose multiple images $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ are given whose rays are tangent to the supporting cylinder $[r\sin(s), r\cos(s), z]$ and intersect with the spiral from different views, as shown in Figure 3. Such image points are easy to identify in images of the spiral. For example, we can take image points that both lie on the (image) spiral and the "contour" of the (image) spiral which typically is two "straight lines" in the image plane, see Figure 4. We can further assume that the images are ordered in a way that their indices grows monotonically in s and the increase in s between consecutive ones is less than 2π.[19]

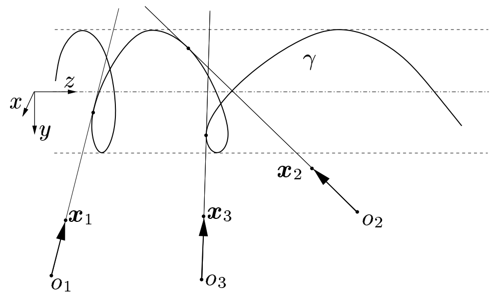

Figure 3: Images of points on a spiral curve γ are taken at different vantage points. Also suppose the image rays tangent to the curve.

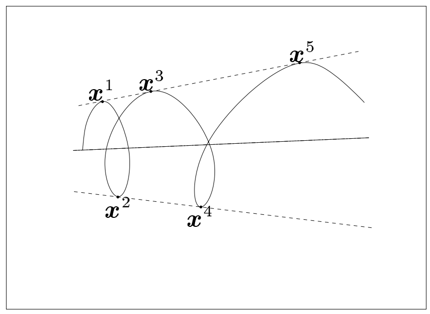

Figure 4: Image points $\{\boldsymbol{x}^i\}_{i=1}^5$ on the supporting cylinder (the "contour" of the spiral). The ray connecting camera center o and $\boldsymbol{x}^i$ will be both tangent to the cylinder and intersecting with the spiral in 3-D.

The i^{th} image ray L_i can be expressed as $\widehat{\boldsymbol{x}_i}\Pi_i \in \mathbb{R}^{3\times 4}$ since this matrix is of rank two and its kernel contains exactly points on the image ray. The line L_i then tangent to the cylinder at a

[18] Here we assume $v_3 \neq 0$. The reader should notice that by doing this we do not lose any generality since we can always change the reference coordinates to make $v_3 \neq 0$, or if v_1 or v_2 is non-zero, we can simply switch their roles by exchanging the first three columns of $U_p(v)$.

[19] A ray of such intersects with the cylinder at a unique point, the value of t for this ray hence can be determined by (x and y coordinates of) this point.

unique point p_i. Suppose the (x, y, z)-coordinates of this point are $(r\sin(s_i), r\cos(s_i), z_i)$. Then s_i is uniquely determined by the ray $L_i : \widehat{\boldsymbol{x}_i}\Pi_i$ (with the known supporting cylinder). Therefore, s_i is a function of both $\boldsymbol{x}_i$ and Π_i, denoted as $s_i(\boldsymbol{x}_i, \Pi_i)$. The z-coordinate of the point p_i is also determined by the ray and the cylinder, denoted as $z_i(\boldsymbol{x}_i, \Pi_i)$. For n images we have

$$\underbrace{\begin{bmatrix} 1 & s_1 & s_1^2 & \cdots & s_1^n & z_1 \\ 1 & s_2 & s_2^2 & \cdots & s_2^n & z_2 \\ \vdots & \vdots & \vdots & \ddots & \vdots & \vdots \\ 1 & s_n & s_m^2 & \cdots & s_m^n & z_m \end{bmatrix}}_{\breve{W}_p \in \mathbb{R}^{m \times (n+2)}} \begin{bmatrix} \alpha_0 \\ \alpha_1 \\ \vdots \\ \alpha_n \\ -1 \end{bmatrix} = 0 \qquad (15)$$

satisfies the rank condition

$$\boxed{\text{rank}(\breve{W}_p) \leq n + 1.} \qquad (16)$$

Note that the matrix $\breve{W}_p$ is a (matrix) function of $\boldsymbol{x}$ and Π only. Hence, for $m \geq n + 2$, this rank condition imposes non-trivial determinantal constraints among its $(n + 2) \times (n + 2)$ sub-matrices. More specifically one can easily verify that: 1. Any $m \leq n$ images (rays) tangent to the cylinder $[r\sin(s), r\cos(s), z]$ belong to some, in fact infinitely many, spirals; 2. It takes $m = n + 1$ such images (rays) in general position to determine all the coefficients α_i's uniquely (and there always exists one);[20] 3. Therefore, non-trivial irreducible constraints must involve at least $m = n+2$ such images (rays).

Increasing the degree of the polynomial $\sum_{i=0}^{n} \alpha_i s^i$, such constraints may involve any number of views. We hence may conclude:

Proposition 6 (Irreducible constraints among m views) *From images of (non-corresponding) points on certain class of 3-D curves, non-trivial irreducible (intrinsic) constraints may involve up to m views for any large integer $m \in \mathbb{Z}_+$. Above generalized spiral is an example.*

Note that to utilize this type of constraints, one do not need to synchronize among image sequences nor to precisely time the tracked features; all one needs to know is the class of trajectories which the motion of the feature obeys, such as straight lines, conics, or (generalized) spirals etc.

5 Conclusions

In this paper, based constructive examples, we have argued that the extant effort in computer vision community to reduce multiple view analysis to pairwise and triple-wise images is

1. *not really necessary:* the rank conditions are more universal and directly instantiates all possible intrinsic constraints among points, lines, and planes; they make the computation of relative camera configuration much easier than the use of multi-focal tensors.

2. *very much limited:* there are intrinsic constraints among more than 4 images even only incidence relations among points and lines are considered; there are intrinsic constraints among arbitrary number of views if dynamical scenes or more general classes of geometric primitives (e.g., curves) are considered.

In general, rank conditions serve as a guiding principle and unifying tool for the study of (or search for) a wide range of features and incidence relations, particularly powerful for analyzing a large number of images. However, rank based algorithms for the exact perspective case have only been investigated to a very limited extent (see [Ma et al. 2001; Huang et al. 2002] for some detail). The full potential of rank based methods in applications such as feature tracking or matching, image sequence synchronization, model identification, and 3-D motion and structure reconstruction in both static and dynamical cases is yet to be explored. A detailed and systematic study about relations among rank conditions associated to different types of camera models (approximated or not) and their practical implications is also largely due.

[20] This is because the first n columns of $\breve{W}_p$ form the so-called *Van De Monde* matrix which has rank n if s_i are all different.

References

AVIDAN, S., AND SHASHUA, A. 1999. Trajectory triangulation of lines: Reconstruction of a 3d point moving along a line from a monocular image sequence. In *Proceedings of CVPR*.

FAUGERAS, O., AND MOURRAIN, B. 1995. On the geometry and algebra of the point and line correspondences between N images. In *Proceedings of Fifth International Conference on Computer Vision*, IEEE Comput. Soc. Press, Cambridge, MA, USA, 951–6.

FAUGERAS, O., LUONG, Q.-T., AND PAPADOPOULO, T. 2001. *Geometry of Multiple Images*. The MIT Press.

GHOSH, B. K., AND LOUCKS, E. P. 1995. A perspective theory for motion and shape estimation in machine vision. *SIAM Journal on Control and Optimization 33*, 5 (Sept.), 1530–1559.

HARRIS, J. 1992. *Algebraic Geometry: A first course*. Springer.

HARTLEY, R., AND ZISSERMAN, A. 2000. *Multiple View Geometry in Computer Vision*. Cambridge.

HUANG, K., MA, Y., AND FOSSUM, R. 2002. Generalized rank conditions in multiple view geometry and its applications to dynamical scenes. In *Proceedings of ECCV*.

IRANI, M. 1999. Multi-frame optical flow estimation using subspace constraints. In *Proceedings of ICCV*.

MA, Y., HUANG, K., VIDAL, R., KOSECKA, J., AND SASTRY, S. 2001. Rank conditions on the multiple view matrix. *submitted to International Journal of Computer Vision* (November).

MA, Y., HUANG, K., AND YANG, Y. 2002. Classification of rank conditions for multiple views of dynamical scenes. *UIUC CSL Technical Report* (May).

POELMAN, C. J., AND KANADE, T. 1997. A paraperspective factorization method for shape and motion recovery. *IEEE Transactions on Pattern Analysis and Machine Intelligence 19*, 3, 206–18.

QUAN, L., AND KANADE, T. 1996. A factorization method for affine structure from line correspondences. In *Proceedings of the CVPR*, 803–808.

ROOM, T. G. 1938. *The Geometry of Determinantal Loci*. Cambridge at the University Press.

SOATTO, S., DORETTO, G., AND WU, Y. 2001. Dynamic textures. In *Proceedings of ICCV*, 439–46.

SPARR, G. 1994. A common framework for kinetic depth, reconstruction and motion for deformable objects. In *Proceedings of ECCV*.

TOMASI, C., AND KANADE, T. 1992. Shape and motion from image streams under orthography. *Intl. Journal of Computer Vision 9*, 2, 137–154.

TORRESANI, L., YANG, D., ALEXANDER, E., AND BREGLER, C. 2001. Tracking and modeling non-rigid objects with rank constraints. In *Proceedings of CVPR*.

TRIGGS, B. 1996. Factorization methods for projective structure and motion. In *Proceedings of 1996 Computer Society Conference on Computer Vision and Pattern Recognition*, IEEE Comput. Soc. Press, San Francisco, CA, USA, 845–51.

WOLF, L., AND SHASHUA, A. 2001. On projection matrices $\mathcal{P}^k \rightarrow \mathcal{P}^2, k = 3, \ldots, 6$, and their applications in computer vision. In *Proceedings of the international conference on computer vision*, 412–419.