

The joint image handbook

Matthew Trager, Martial Hebert, Jean Ponce

▶ **To cite this version:**

Matthew Trager, Martial Hebert, Jean Ponce. The joint image handbook. ICCV 2015, Dec 2015, Santiago, Chile. hal-01249171

HAL Id: hal-01249171
https://hal.science/hal-01249171

Submitted on 30 Dec 2015

HAL is a multi-disciplinary open access archive for the deposit and dissemination of scientific research documents, whether they are published or not. The documents may come from teaching and research institutions in France or abroad, or from public or private research centers.

L'archive ouverte pluridisciplinaire **HAL**, est destinée au dépôt et à la diffusion de documents scientifiques de niveau recherche, publiés ou non, émanant des établissements d'enseignement et de recherche français ou étrangers, des laboratoires publics ou privés.

The joint image handbook

Matthew Trager* Inria
Martial Hebert Carnegie Mellon University
Jean Ponce* École Normale Supérieure / PSL Research University

Abstract

Given multiple perspective photographs, point correspondences form the "joint image", effectively a replica of three-dimensional space distributed across its two-dimensional projections. This set can be characterized by multilinear equations over image coordinates, such as epipolar and trifocal constraints. We revisit in this paper the geometric and algebraic properties of the joint image, and address fundamental questions such as how many and which multilinearities are necessary and/or sufficient to determine camera geometry and/or image correspondences. The new theoretical results in this paper answer these questions in a very general setting and, in turn, are intended to serve as a "handbook" reference about multilinearities for practitioners.

1. Introduction

Point correspondences in multiple images can be characterized using conditions that are "multilinear" in homogeneous image coordinates (*e.g.*, epipolar and trifocal constraints [5, 13, 16, 17, 21]). These constraints are at the core of any structure-from-motion (SfM) system, where they are mainly used in two tasks: selecting matching points in different pictures, and estimating the camera parameters from these correspondences. Yet, after 35 years of study, dating back to Longuet-Higgins' seminal work on the essential matrix [13] (and at least to the sixties in photogrammetry [19]), practicioners and specialists alike would still be hard pressed today to answer many simple questions such as how many multilinear relations (and which ones) are necessary and/or sufficient to characterize correspondences, or to determine the corresponding camera parameters.

Partial results are available, but scattered in the literature, and they sometimes contradict each other [4, 6, 14]. The aim of this presentation is to give new and definite answers to this type of elementary but fundamental questions in a very general setting: our hope is to provide a practical "handbook" reference for useful results and facts on multilinearities. Our results are obtained by using elementary tools from algebraic geometry to characterize the *joint image*, introduced by Triggs in [20]: this set is formed by the n-tuples of matching points, and is in fact in a formal sense an (almost) exact replica of 3D-space "distributed" across multiple images. An advantage of this geometric viewpoint is that it does not require the analytic tools that have been exploited in the past for deriving multi-view constraints (such as Grassman-Cayley algebras [4], tensor calculus [20], Plücker coordinates and line geometry [14]). Moreover, the joint image simultaneously describes point correspondences and camera geometry independently of the choice of coordinates in space, and this natural setting may be useful for revisiting many existing algorithms (cf. the discussion in Section 4). All of our results apply to ordinary (affine or Euclidean) pinhole cameras with known intrinsic parameters as well as projective, uncalibrated ones. We thus believe that they are highly relevant in practice.

Closely related to our work is that of Heyden and Åström in [11], who also study the algebraic properties of multiview constraints. For example, these authors show that for cameras in general position, the epipolar (bilinear) constraints are sufficient for characterizing correspondences for $n \geq 4$ views; they also observe that for $n \geq 5$, "some" conditions can be dropped. We extend these results, characterizing the extent to which trilinear relations are required in the case of degenerate camera configurations (Proposition 5), and giving useful bounds on the number of necessary conditions for generic configurations (Proposition 6). Heyden and Åström also discuss in [11] an interesting property of the epipolar constraints for three cameras in general position: these conditions uniquely determine (up to projective ambiguity) the associated camera matrices; *however*, the trilinear conditions do not follow algebraically from the bilinear ones and, in fact, bilinear constraints are not sufficient for characterizing point correspondences in general. We discuss in Section 3 this somewhat paradoxical behavior in a more general setting, defining the notions of *weak* and *strong* characterizations of the joint image, that respectively determine camera parameters and point correspondences. For example, we will show that (perhaps suprisingly) the nine trilinearities encoded in the trifocal tensor are *not* sufficient to completely ensure correspondence among three views (Proposition 9), although they can be used to recover the corresponding projection matrices. On the other hand, for n views in general position, camera matrices can always be determined using $2n - 3$ epipolar relations, assuming these relate appropriate pairs of cameras (Proposition 7).

*Willow project team. DI/ENS, ENS/CNRS/Inria UMR 8548.

Main contributions:

• We discuss in full generality the difference between the constraints that determine camera geometry, and those that characterize correspondences. The distinction is related to the geometry (*i.e.*, the decomposition in irreducible components) of the set of n-tuplets that satisfy the different constraints (Section 3).

• We give a series of results that provide explicit conditions for characterizing correspondences as well as camera geometry. In particular, we clarify in full generality the relationship between bilinear, trilinear and quadrilinear constraints (Proposition 5), and discuss the problem of finding minimal sets of necessary and sufficient conditions (Propositions 6, 7), improving on results from [11].

• We focus more closely on the case of three views (Propositions 8, 9, 10), clarifying several properties of the space of trilinear constraints, such as the fact that it is always a vector space of dimension 10. More generally, we give the number d_n of linearly independent multilinear constraints for any n cameras (Proposition 3), a result that is consistent with a more technical theorem given in [1].

• We present a general discussion of the basic geometric properties of the joint image (such as its singular locus, cf. Proposition 4), and argue that its (weak or strong) characterizations can be useful in practical tasks (Section 4).

Mathematical background. Our analysis makes use of some elementary aspects of algebraic geometry. For the convenience of the reader, we have included a brief introduction to these topics in the supplementary material; for more details we refer for example to [2]. Technical proofs are deferred to the supplementary material (we provide intuitive proof sketches whenever possible), however the statements of our main results (Propositions 5, 6, 7, 9, 10) do not require any technical prerequisites.

Notation. We assume a fixed coordinate frame for $\mathbb{P}^3$, and identify points with their homogeneous coordinate vectors. A camera in $\mathbb{P}^3$ will be described by a matrix $\mathcal{M} \in \mathbb{R}^{3\times 4}$ of full rank, defined up to scale: such a matrix describes a linear projection from $\mathbb{P}^3 \setminus \{\boldsymbol{c}\}$ to $\mathbb{P}^2$, where $\boldsymbol{c} \in \mathbb{P}^3$ is given by the nullspace of $\mathcal{M}$ and represents the optical center, or pinhole, of the camera [6]. Cameras and the associated projection matrices will be identified. Points in $\mathbb{P}^3$ or $\mathbb{P}^2$ will be represented by bold letters, while coordinates will be in normal font, with superscripts to indicate indices (*e.g.*, $\boldsymbol{p} = [p^1; p^2; p^3; p^4] \in \mathbb{P}^3$). The action of a camera $\mathcal{M}$ in $\mathbb{R}^{3\times 4}$ will be written as $\mathcal{M}\boldsymbol{p} \sim \boldsymbol{u}$, for $\boldsymbol{p}$ in $\mathbb{P}^3$ and $\boldsymbol{u}$ in $\mathbb{P}^2$, where $\sim$ expresses equality up to non-zero scalars between the coordinate vectors representing projective points.

2. The joint image

This section presents Triggs' joint image [20] (also known as the *multi-view variety* [1]), that will be the central object of our analysis throughout the paper. After giving some formal definitions, we derive the basic *multilinear* algebraic constraints that can be used to characterize correspondences. We then analyze the (closure of the) joint image as an algebraic variety, pointing out some of its geometric properties.

2.1. Defintions

Let $\mathcal{M}_1, \ldots, \mathcal{M}_n$ be n projective cameras with distinct centers $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$.

Definition 1. *An n-tuple of image points $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ is a* correspondence *if there exists $\boldsymbol{p}$ in $\mathbb{P}^3 \setminus \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\}$ such that $\mathcal{M}_i\boldsymbol{p} \sim \boldsymbol{u}_i$ for all $i = 1, \ldots, n$.*

The joint image $\mathcal{I}_n(\mathcal{M}_1, \ldots, \mathcal{M}_n)$ *[20], is the subset of $(\mathbb{P}^2)^n$ formed by image correspondences.*

Although the joint image depends on the camera matrices $\mathcal{M}_1, \ldots, \mathcal{M}_n$, we will often denote it simply with $\mathcal{I}_n$ when no confusion can arise.

It was noted in [11] that the joint image $\mathcal{I}_n$ (which Heyden and Åström refer to as the "natural descriptor") is *not* an algebraic set, in other words it cannot be described as the zero-set of a family of polynomial equations.

Definition 2. *The* joint image variety $\overline{\mathcal{I}}_n(\mathcal{M}_1, \ldots, \mathcal{M}_n)$ *is the Zariski closure of the joint image.*

In the Zariski topology, closed sets coincide with algebraic sets so, in practice, $\overline{\mathcal{I}}_n$ is simply the smallest set containing $\mathcal{I}_n$ which can be described by polynomial equations. As illustrated by the following example, the distinction between $\mathcal{I}_n$ and $\overline{\mathcal{I}}_n$ is well understood for simple cases.

Example 1. Given two cameras $\mathcal{M}_1$ and $\mathcal{M}_2$, any correspondence $(\boldsymbol{u}_1, \boldsymbol{u}_2)$ in $\mathcal{I}_2$ satisfies the algebraic relation $\boldsymbol{u}_1^T \boldsymbol{F} \boldsymbol{u}_2 = 0$ (the *epipolar constraint*), where $\boldsymbol{F} \in \mathbb{R}^{3\times 3}$ is the *fundamental matrix* associated with $\mathcal{M}_1$ and $\mathcal{M}_2$ [6]. However, the set $\overline{\mathcal{I}}_2$ of pairs of points satisfying this bilinear constraint is strictly larger than $\mathcal{I}_2$. Indeed, if $\boldsymbol{e}_1$ is the first epipole (given by the left null-space of $\boldsymbol{F}$), then $(\boldsymbol{e}_1, \boldsymbol{u}_2)$ will be in $\overline{\mathcal{I}}_2$ for all $\boldsymbol{u}_2 \in \mathbb{P}^2$. However, a pair $(\boldsymbol{e}_1, \boldsymbol{u}_2)$ is an actual correspondence only when $\boldsymbol{u}_2$ coincides with the second epipole $\boldsymbol{e}_2$. The joint image is in fact given by $\mathcal{I}_2 = \overline{\mathcal{I}}_2 \setminus \mathcal{C}_2$, where

$$\mathcal{C}_2 = (\{\boldsymbol{e}_1\} \times \mathbb{P}^2) \cup (\mathbb{P}^2 \times \{\boldsymbol{e}_2\}) \setminus \{(\boldsymbol{e}_1, \boldsymbol{e}_2)\}. \quad (1)$$

Note that $\mathcal{C}_2$ is a distinguished set that can be computed directly from the fundamental matrix $\boldsymbol{F}$.

More generally, there is always a non-empty set $\mathcal{C}_n = \overline{\mathcal{I}}_n \setminus \mathcal{I}_n$ containing n-tuples of points that are not actual

correspondences, but that will satisfy *any* set of algebraic equations that are also satisfied by correspondences. This set is described explicitly in the following proposition.

Proposition 1. *Given $n \geq 3$ cameras with non-collinear (distinct) pinholes, one has $\mathcal{I}_n = \overline{\mathcal{I}}_n \setminus \mathcal{C}_n$, where*

$$\mathcal{C}_n = \bigcup_{i=1}^{n} (\boldsymbol{e}_{i1} \times \ldots \times \mathbb{P}^2_{(i)} \times \ldots \times \boldsymbol{e}_{in}). \tag{2}$$

Here $\mathbb{P}^2_{(i)}$ indicates $\mathbb{P}^2$ at position i in the product, and $\boldsymbol{e}_{ij}$ denotes the epipole in image j relative to image i. If $n = 2$, or more generally if the cameras have collinear pinholes, then one must remove from $\mathcal{C}_n$ the n-tuple of epipoles $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$ (in this case there is only one epipole in each image).

The proof of this result follows easily from the characterization of $\overline{\mathcal{I}}_n$ that we will discuss in Section 2.2, see the supplemental material for details. Note that the set $\mathcal{C}_n$ is always a distinguishable set, *i.e.*, it contains special n-tuples of points that can easily be detected as spurious solutions (as for the case for $n = 2$ discussed in Example 1). This means that, in practice, we do not lose any actual information by replacing $\mathcal{I}_n$ by its closure $\overline{\mathcal{I}}_n$. In our study, we will talk about equations that "characterize correspondences", referring to polynomial constraints that actually describe the joint image variety $\overline{\mathcal{I}}_n$.

2.2. Algebraic properties of the joint image

Multilinear conditions. Following [1, 7, 10, 11], given an n-tuple of image points $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$, we can define the $3n \times (n + 4)$ matrix

$$\mathcal{U}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) = \begin{pmatrix} \mathcal{M}_1 & \boldsymbol{u}_1 & \mathbf{0} & \ldots & \mathbf{0} \\ \mathcal{M}_2 & \mathbf{0} & \boldsymbol{u}_2 & \ldots & \mathbf{0} \\ \vdots & \vdots & \vdots & \ddots & \vdots \\ \mathcal{M}_n & \mathbf{0} & \mathbf{0} & \ldots & \boldsymbol{u}_n \end{pmatrix}. \tag{3}$$

A *necessary* condition for $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ to form a correspondence is clearly that $\mathcal{U}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ be rank deficient (since there would not exist a nonzero vector $[\boldsymbol{p}; \lambda_1; \ldots; \lambda_n]$ in the nullspace of the matrix otherwise). Hence, the maximal minors of (3) are polynomial conditions in image coordinates that all correspondences *must* satisfy. In fact, one can show that they are also sufficient to define $\overline{\mathcal{I}}_n$ [1]:

$$\overline{\mathcal{I}}_n = \{(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \in (\mathbb{P}^2)^n \mid \mathcal{U}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n) \text{ is not full rank}\}.$$

The constraints given by the maximal minors of (3) are easily seen to be *multilinear*, in other words, they are polynomials in $\mathbb{R}[x_1, y_1, z_1, \ldots, x_n, y_n, z_n]$ that are linear in each triplet of variables x_i, y_i, z_i $(i = 1, \ldots, n)$ associated with an image.[1]

[1] It will be useful to define more generally a *k-linear* polynomial in $\mathbb{R}[x_1, y_1, z_1, \ldots, x_n, y_n, z_n]$ (for $k \leq n$) as a polynomial which involves only k triplets of variables and is linear in each triplet that appears (so a multilinear polynomial is the same as an n-linear polynomial). In particular, k-linear polynomials for $k = 2, 3, 4$ will also be described as "bilinear", "trilinear", and "quadrilinear", respectively.

Of course, many other polynomial constraints can be obtained by considering the minors of (3) based on $k \leq n$ of the original camera matrices. This yields families of k-linear relations for $2 \leq k \leq n$, which we will refer to as the *k-linearities*. In practice, it is easy to see that only k-linearities with $k \leq 4$ need to be considered: this is closely related to the fact the multi-view tensors do not exist for more than four views [10].

Proposition 2. *Every n-linearity is of the form mP where m is a monomial factor and P is a k-linearity with $k \leq 4$. This implies that bilinearities, trilinearities and quadrilinearities are sufficient to characterize $\overline{\mathcal{I}}_n$.*

Proof. The result follows from the fact that a non-vanishing minor of $\mathcal{U}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ requires choosing $n+4$ rows, with at least one row associated with each camera: this distinguishes k cameras with $2 \leq k \leq 4$ for which more than one row is chosen. The monomial factors can be removed from the constraints since each k-linearity is multiplied by sets of monomials that cannot vanish simultaneously. □

One can show that the k-linearities for $2 \leq k \leq 4$ are sufficient to generate the largest *ideal* associated to $\overline{\mathcal{I}}_n$ [1]: in practice this means that *all* multi-view constraints can always be deduced algebraically from these basic relations (even if derived using other approaches *e.g.*, the trilinearities in [14], obtained using line geometry). However, the complete description of the joint image based on all the bilinear, trilinear and quadrilinear constraints is generally very redundant. We will see in Section 3 that the quadrilinear constraints are always completely unnecessary (a well known fact [4]), and, more importantly, much fewer bilinear and trilinear conditions can actually be used.

Vector spaces of multilinearities. The multilinear relations that vanish on $\overline{\mathcal{I}}_n$ (*i.e.*, the k-linearities for $k = n$) form a *vector space*, the dimension of which is given by the following proposition. The result can also be deduced from Theorem 3.6 in [1], although expressed in more technical terms (the authors provide the "multigraded Hilbert function" for the ideal associated to $\overline{\mathcal{I}}_n$).

Proposition 3. *Given n cameras $\mathcal{M}_1, \ldots, \mathcal{M}_n$, the multilinear polynomials that vanish on $\overline{\mathcal{I}}_n$ form a vector space of dimension $d_n = 3^n - \binom{n+3}{3} + n$.*

Proof sketch. It is sufficient to compute the dimension of the vector space generated by the initial terms associated to the multilinear constraints. Since the maximal minors of $\mathcal{U}(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ defined in (3) form a Gröbner basis, the result follows from a counting argument involving the associated initial monomials. □

For example, let us point out that $d_2 = 1$ (the epipolar constraint is the only bilinear relation for two views),

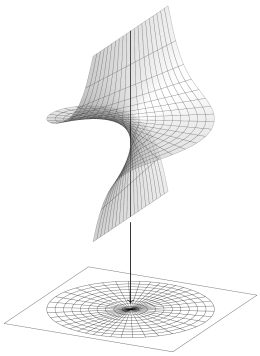


Figure 1. Example of a birational equivalence: the projective plane $\mathbb{P}^2$ and its so-called *blow-up* at a point are identical, apart from a single point of $\mathbb{P}^2$, which is "expanded" into a line [9]. The relationship between $\mathbb{P}^3$ and $\overline{\mathcal{I}}_n$ is analogous: in fact, we show in the supplemental material that the joint image variety (assuming non-collinear pinholes) is given by this same construction (the blow-up) applied to $\mathbb{P}^3$ at all of the camera pinholes.

$d_3 = 10$ (there are always 10 linearly independent trilinearities, cf. Section 3.4), and $d_4 = 50$; these facts can also be verified computationally using Gröbner bases.

2.3. Geometric properties of the joint image

By definition, $\overline{\mathcal{I}}_n$ is the closure of the image of the "joint-projection" map

$$\begin{array}{ccc} \mathbb{P}^3 \setminus \{\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n\} & \longrightarrow & \mathbb{P}^2 \times \ldots \times \mathbb{P}^2 \\ \boldsymbol{p} & \longmapsto & \mathcal{M}_1\boldsymbol{p} \times \ldots \times \mathcal{M}_n\boldsymbol{p}. \end{array} \tag{4}$$

This map is usually injective, the only exception being when all of the cameras have collinear pinholes (in particular, when there are only two views), in which case two points lying on the baseline will have the same images. The inverse function, where it is well-defined, is *(exact) triangulation*, that is, the operation of recovering spatial coordinates from corresponding image points. Note that this is a rational map, *i.e.*, it can be described using polynomial expressions (because it amounts to computing the intersection of visual rays). The existence of rational maps that are the inverse of each other for "generic" points is expressed in the language of algebraic geometry by saying that $\overline{\mathcal{I}}_n$ and $\mathbb{P}^3$ are *birationally equivalent* (Figure 1). Intuitively, this says $\overline{\mathcal{I}}_n$ is a model of $\mathbb{P}^3$ embedded in $\mathbb{P}^2 \times \ldots \times \mathbb{P}^2$, which immediately implies that the joint image is irreducible (it is not the union of proper subvarieties) and has dimension 3.

Since $\mathbb{P}^2 \times \ldots \times \mathbb{P}^2$ has dimension $2n$, one could hope to be able to describe $\overline{\mathcal{I}}_n$ using $2n - 3$ constraints. However, typically one cannot represent an algebraic set of codimension m as the intersection of m hypersurfaces (when this is possible, the set is called a "complete intersection"). It is true, however, that *at least* this many conditions are necessary and one can always use this minimum number of constraints for *local* characterizations of the joint image (at least away from singularities, see below).

Example 2. Consider three cameras $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ with non-collinear pinholes. Let B_{ij} be the epipolar constraint between views i and j, and T be any trilinear constraint that does not vanish on the product of the "trifocal lines" (*i.e.*, the projections of the plane containing the pinholes). Consider the following sets of constraints

$$S_1 : \{B_{12}, B_{23}, B_{13}\},\ S_2 : \{B_{12}, B_{13}, T\}. \tag{5}$$

Both S_1 and S_2 give *minimal* and *local* descriptions of the joint image variety $\overline{\mathcal{I}}_3$: for example, if $(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3)$ does not lie on the trifocal lines, then S_1 is sufficient for establishing whether they form a correspondence [14]. The same holds for S_2 if one excludes a more complicated set of spurious solutions (see the supplemental material). However, in order to obtain a *global* description of $\overline{\mathcal{I}}_3$, one has to consider all four equations in $S_1 \cup S_2$, even though this characterization will be locally redundant.

The following proposition deals with the singularities of $\overline{\mathcal{I}}_n$. The proof is technical, and deferred to the supplemental material.

Proposition 4 (Singularities of the joint image variety)**.** *When the camera pinholes are not collinear, $\overline{\mathcal{I}}_n$ is smooth. When they are collinear (in particular, for $n = 2$ views), then $\overline{\mathcal{I}}_n$ has a unique singular point given by the n-tuple of epipoles $(\boldsymbol{e}_1, \ldots, \boldsymbol{e}_n)$.*

The joint image and camera matrices. It is clear that the association between camera matrices and $\overline{\mathcal{I}}_n$ does not depend on the reference frame in $\mathbb{P}^3$, in other words $\overline{\mathcal{I}}_n(\mathcal{M}_1, \ldots, \mathcal{M}_n) = \overline{\mathcal{I}}_n(\mathcal{M}_1 T, \ldots, \mathcal{M}_n T)$ for all T in $GL_4(\mathbb{R})$. Note also that for any $S_1, \ldots, S_n$ in $GL_3(\mathbb{R})$, $\overline{\mathcal{I}}_n(\mathcal{M}_1, \ldots, \mathcal{M}_n)$ and $\overline{\mathcal{I}}_n(S_1\mathcal{M}_1, \ldots, S_n\mathcal{M}_n)$ are completely equivalent, since they are identical up to linear changes of variables. Conversely, it is important to emphasize that *the joint image completely characterizes the set of cameras*, up to changes of coordinates in $\mathbb{P}^3$: indeed, all SFM methods are based on the property that camera parameters can be recovered given a sufficient number of correspondences across multiple views (at least 7 correspondences for $n = 2$ and at least 6 for $n \geq 3$), and the joint image describes *all* matches between the views.

3. Main results

We now resume our study of the different sets of multilinear constraints that can be used to describe the joint image. First, however, we make the important observation that it may be possible to recover the joint image (and camera parameters) from sets of constraints that do not actually guarantee correspondence globally. This leads us to study, in Section 3.2, the relationship between the bilinear, trilinear, and quadrilinear constraints. We analyze in Section 3.3 some practical sets of epipolar constraints that can be used for generic configurations. Finally, we discuss in Section 3.4 the important case of three views.

3.1. Characterizations of the joint image

Let us assume that we are given multilinear polynomials $P_1, \ldots, P_s$ that are annihilated by all elements of $\overline{\mathcal{I}}_n$, and denote with $\mathcal{W} \supseteq \overline{\mathcal{I}}_n$ the algebraic set defined by these polynomials. Interestingly, $\mathcal{W}$ may completely determine camera geometry even when it is strictly larger than $\overline{\mathcal{I}}_n$ and thus does not characterize correspondences. The following example illustrates this behavior (see also [11]):

Example 3. Consider three cameras $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ with non-collinear pinholes. We have already observed that the three epipolar constraints do not yield a global description of $\overline{\mathcal{I}}_3$, and in fact:

$$\mathcal{W} = \{B_{12} = B_{13} = B_{23} = 0\} = \overline{\mathcal{I}}_3 \cup \mathcal{V}_t, \qquad (6)$$

where $\mathcal{V}_t$ is the product of the trifocal lines [11]. However, it has also been observed that the epipolar constraints (or the fundamental matrices) are sufficient to recover the camera matrices [4, 6]. From (6), we see that $\overline{\mathcal{I}}_3$ appears as an irreducible component of the larger set $\mathcal{W}$ (see the example in Figure 2). In practice, polynomial equations for irreducible components can be computed by means of primary decomposition (see [2] or Section A of the supplementary material): this means that *all* constraints defining $\overline{\mathcal{I}}_3$ (*e.g.*, trilinearities) can be recovered indirectly, even if they are not algebraic combinations of the epipolar conditions. This gives an algebraic justification for how the epipolar constraints determine camera geometry.

Generalizing the previous example, we give the following definition:

Definition 3. *A set of multilinear constraints $P_1, \ldots, P_s$ is referred to as a* weak *characterization of the joint image (or of correspondences) when it* uniquely *determines camera geometry. A* strong *characterization is a set of conditions that describe the joint image variety in the usual sense,* i.e.*, $\overline{\mathcal{I}}_n = \{P_1 = \ldots = P_s = 0\}$, so they directly give conditions for correspondence.*

As the terminology suggests, a "strong" characterization of the joint image is also "weak", since we know that the joint image uniquely determines camera geometry (cf. Section 2.2).

In practice, multilinearities provide a weak characterization of the joint image whenever the associated variety $\mathcal{W}$ contains $\overline{\mathcal{I}}_n$ as an irreducible component, as in Example 3. Note that once camera projections are recovered, correspondences can subsequently be correctly characterized (in other words, it is always possible to verify whether a candidate correspondence is actually an extraneous solution). Weak characterizations have the advantage of usually being simpler and, in many cases, they can be used in place of strong ones, since they provide sufficient conditions for correspondence away from spurious components (*i.e.*, they are local characterizations, see Example 2). As we will see in our discussion of the "trifocal" trilinearities in Section 3.4, it is actually likely that sets of weakly sufficient conditions are sometimes used unknowingly, because correspondences can be assumed general enough not to lie on a spurious component (*e.g.*, generic image points are not epipoles, *etc.*).

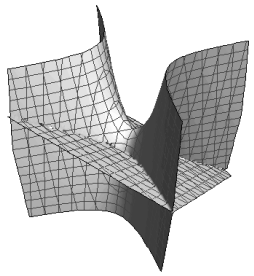


Figure 2. An algebraic surface $p(x, y, z) = 0$ with two irreducible components, that can be recovered by factoring $p(x, y, z)$. In general, when considering more than a single constraint, irreducible components can still be recovered, but factorization has to be replaced with primary decomposition of ideals [2].

3.2. Dependencies among multi-view constraints

We have observed in Section 2.2 that it is possible to describe the joint image using constraints that involve at most four views, that is, using k-linearities for $k \leq 4$ (Proposition 2). However, Heyden and Åström [11] point out that for $n \geq 4$, assuming all the pinholes to be in general position, the bilinear constraints are already sufficient to generate the trilinear and quadrilinear ones. We now extend this result, clarifying the role of the different families of constraints for all possible camera configurations. The proof of the following proposition amounts to reducing it to the case $n = 4$ and verifying all the relations computationally: we refer the reader to the supplemental material for details.

Proposition 5. *Assume n cameras are given.*

1. Bilinearities and trilinearities always *strongly characterize $\overline{\mathcal{I}}_n$, independently of the camera configurations.*

2. Bilinear constraints alone *strongly characterize $\overline{\mathcal{I}}_n$, if and only if the pinholes are* not all coplanar.

3. Bilinear constraints alone *weakly characterize $\overline{\mathcal{I}}_n$ if and only if the pinholes are* not all collinear.

We should observe that the first point of Proposition 5 is well known (see for example [4]), and indeed reconstruction methods are generally only based on epipolar and trifocal constraints. The second point can be deduced from the analysis for four points in general position in [11], although the authors do not point out this general fact. The last subtle point is, to the best of our knowledge, new, at least

	Non coplanar	Coplanar	Collinear
Bil.	Strong	Weak	Not sufficient
Bil.+Tril.	Strong	Strong	Strong

Table 1. Summary of the results of Proposition 5.

with such a general formulation: it clearly shows that trilinear relations are essential *only* if *all* of the cameras have collinear pinholes, since otherwise the epipolar relations are sufficient to completely capture the geometry among all the views. See Table 1.

3.3. Subsets of epipolar constraints

We have considered so far *complete* families of bilinear and trilinear constraints. These sets of conditions are redundant in general: for example, it has been observed in [11] that for five cameras in general position, correspondences can be (strongly) characterized using 9 bilinear constraints instead of the complete set of 10. Although it is difficult to make general statements on minimal sets of necessary constraints, we can make some useful remarks.

Proposition 6. *Assume that we are given $n \geq 4$ cameras with pinholes $\boldsymbol{c}_1, \ldots, \boldsymbol{c}_n$ (the case $n = 3$ is treated in more detail in Section 3.2). See Figure 3.*

(A) If the pinholes (say) $\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3$ are not coplanar with any other $\boldsymbol{c}_i$ for $i \geq 4$, then $\{B_{12}, B_{13}, B_{23}\} \cup \{B_{1i}, B_{2i}, B_{3i}\}_{i=4,\ldots,n}$ are $3n-6$ bilinearities that strongly characterize the joint image.

(B) If the pinholes (say) $\boldsymbol{c}_1, \boldsymbol{c}_2$ are not collinear with any other $\boldsymbol{c}_i$ for $i \geq 3$, then $B_{12} \cup \{B_{1i}, B_{2i}\}_{i=3,\ldots,n}$ are $2n-3$ bilinearities that weakly characterize the joint image.

With slightly modified hypotheses, similar results hold for the sets $\{B_{i,i+1}, B_{i,i+2}, B_{i,i+3}\}_{i=1,\ldots,n-3}$ and $\{B_{i,i+1}, B_{i,i+2}\}_{i=1,\ldots,n-2}$, that are still respectively strong and weak characterizations of the joint image based on $3n-6$ and $2n-3$ constraints.

Proof. If $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ is an n-tuple of image points that satisfy all of the constraints given in (A), then Proposition 5 guarantees that the visual rays associated to $(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3, \boldsymbol{u}_i)$ converge for all $i = 4, \ldots, n$. Since $\boldsymbol{c}_1, \boldsymbol{c}_2, \boldsymbol{c}_3$ are necessarily not collinear, this implies that *all* the visual rays intersect, so that $(\boldsymbol{u}_1, \ldots, \boldsymbol{u}_n)$ is in fact a correspondence.

Similarly, the set of constraints given in (B) allow one to determine a consistent set of camera parameters: it is enough to note that B_{12}, B_{1i}, B_{2i} are weakly sufficient for views $(1, 2, i)$, so that after having fixed $\mathcal{M}_1, \mathcal{M}_2$ compatible with B_{12}, one can uniquely recover all of the remaining cameras. □

Remark. The $2n-3$ weakly sufficient bilinearities given in Proposition 6 (B) define an algebraic set $\mathcal{W}$ of dimension 3: this can be shown by induction, observing that every new view contributes two more independent constraints. In particular, $\mathcal{W}$ must contain $\mathcal{I}_n$ as a component of maximal dimension. This relates to our discussion in the beginning of Section 3, and confirms Conjecture 6.2 in [11].

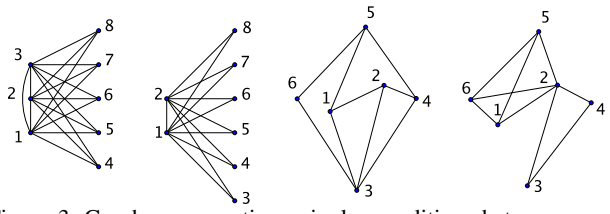


Figure 3. Graphs representing epipolar conditions between cameras in general position: the first two graphs consider $n = 8$ cameras and describe, respectively, a strong characterization based on $3n - 6 = 18$ constraints, and a weak one based on $2n - 3 = 13$ constraints, both according to Proposition 6. The third graph is a minimal configuration for $n = 6$ nodes that satisfies the property (P) given in Proposition 7, and thus describes a weak characterizations using $2n - 3 = 9$ conditions. The fourth graph is a configuration of $2n - 3 = 9$ conditions that does not satisfy (P), and in fact, one easily shows that the corresponding constraints are not weakly sufficient.

When the pinholes of n cameras are in general position (*i.e.*, no four of them are coplanar), $2n - 3$ bilinearities can be used to recover camera geometry, however not all choices of this many constraints will work. The following practical result gives conditions for $2n-3$ epipolar relations to be sufficient for characterizing camera projections.

Proposition 7. *Consider $n \geq 3$ cameras with pinholes in general position, and let G be a graph with n nodes corresponding to the cameras, and edges representing epipolar relations between them. Assume G has the following property:*

(P) G can be constructed from a 3-cycle by adding vertices of degree two, one at the time.

Then the epipolar conditions associated with the edges of G weakly characterize the joint image, and thus they uniquely determine camera geometry. Note that the minimum number of edges for a graph satisfying (P) is $2n - 3$. See Figure 3.

Proof. The proof is by induction on n. For $n = 3$, the graph G is a cycle, and we know that the epipolar constraints between three views are weakly sufficient. Property (P) clearly allows the use of the inductive hypothesis. Having recovered a consistent set of cameras $\mathcal{M}_1, \ldots, \mathcal{M}_{n-1}$ (that will be unique up to homographies in $\mathbb{P}^3$), we can then use two epipolar constraints involving the n-th view to uniquely recover $\mathcal{M}_n$. □

Interestingly, graphs considered by Proposition 7 form a subset of the family of *Laman graphs* [12], which characterize minimally rigid systems of rods and joints in the plane.

3.4. The case of three views

The study of three-view geometry is traditionally based on *trifocal tensors* [5]. It is well known that such tensors also encode 9 trilinear conditions for point correspondences: assuming that $\mathcal{T}$ distinguishes the first view, these

can be expressed as

$$u_1^i u_2^j u_3^k (\epsilon_{jpr}\epsilon_{kqs}\mathcal{T}_i^{pq}) = 0_{rs}, \tag{7}$$

where we use Einstein summation notation (all indices running from 1 to 3), and denote by ϵ_{ijk} the Levi-Civita permutation symbol. These trilinear constraints correspond to maximal minors of the matrix $\mathcal{U}(\boldsymbol{u}_1, \boldsymbol{u}_2, \boldsymbol{u}_3)$ defined in (3), more precisely to the nine minors arising by considering all rows associated to one fixed camera, and two rows from each of the other two.

Let us now clarify some issues related to the necessity or sufficiency of the conditions (7) (all of the results in this section are shown by direct computation using Gröbner bases, after fixing triplets of cameras for both non-collinear as well as the collinear case; see Section B5 of the supplemental material for details).

The first point to address is linear independence. It is frequently reported that only four of the nine conditions are independent [5, 6]: this fact is true if one interprets the equations in (7) as conditions for *recovering coefficients of the trifocal tensor* $\mathcal{T}$, using a known triplet of points. This is *not* the same as the independence of the constraints as polynomials in the point coordinates, and we believe the two notions may have been confused (although some references clearly say "independent in the tensor components" [10]). Regarding the independence as trilinear conditions, we have the following new result. We recall that *all* of the trilinearities form a vector space of dimension 10 (Proposition 3).

Proposition 8. *The nine trilinearities encoded in a trifocal tensor (Eq. (7)) span a vector space of dimension* 8.

Interestingly, assuming non-collinear pinholes, the trilinearities defined by Eq.(7) are *not* (strongly) sufficient for guaranteeing point correspondence, *i.e.*, they describe a set that is strictly larger than the joint image variety $\overline{\mathcal{I}}_3$. We believe that this fact has not been pointed out in previous literature (although it is closely related to the known degeneracies of transfer based on the trifocal tensor [6, Section 15.3.2], see our discussion in the supplemental material).

Proposition 9. *If the pinholes $\mathcal{M}_1, \mathcal{M}_2, \mathcal{M}_3$ are not collinear, then the constraints (7) (assuming that $\mathcal{T}$ distinguishes the first view) describe a set $\mathcal{W} = \overline{\mathcal{I}}_3 \cup \mathcal{S}_{12} \cup \mathcal{S}_{13}$, where*

$$\begin{aligned} \mathcal{S}_{12} &= \{\boldsymbol{e}_{12} \times \boldsymbol{e}_{21} \times \boldsymbol{u}_3 \in (\mathbb{P}^2)^3 \,|\, \boldsymbol{u}_3 \in \mathbb{P}^2\}, \\ \mathcal{S}_{13} &= \{\boldsymbol{e}_{13} \times \boldsymbol{u}_2 \times \boldsymbol{e}_{31} \in (\mathbb{P}^2)^3 \,|\, \boldsymbol{u}_2 \in \mathbb{P}^2\}, \end{aligned} \tag{8}$$

and $\boldsymbol{e}_{ij}$ is the epipole in image i relative to the camera j.

A geometric justification for this result is given by Figure 4. In practice, the set of spurious correspondences (8) described by all nine trilinearities is very limited, since two of the three image points are always constrained to be epipoles. However, it is interesting to observe that the trilinearities expressed by the trifocal tensor are in some sense

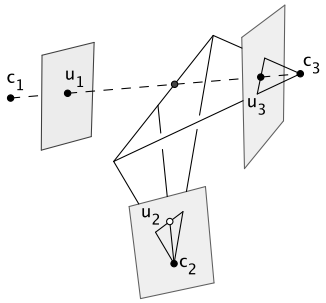


Figure 4. Geometrical explanation for Proposition 9: the points $\boldsymbol{u}_1$ and $\boldsymbol{u}_3$ are epipoles for $\mathcal{M}_1$ and $\mathcal{M}_3$; for *any* choice of $\boldsymbol{u}_2 \in \mathbb{P}^2$ in the second image (shown in white), all lines through $\boldsymbol{u}_2$ and $\boldsymbol{u}_3$ will give rise to a point-line-line correspondence with $\boldsymbol{u}_1$.

not "complete": indeed, according to Proposition 8, they always span a space of dimension 8, strictly included in the vector space of dimension 10 spanned by all trilinearities.[2] Based on our discussion in Section 3.1, the trilinearities (7) are only "weakly sufficient", *i.e.*, they define a larger set than $\overline{\mathcal{I}}_3$ but still uniquely characterize camera matrices (indeed, the trifocal tensor encodes camera geometry).

By considering any subset of the nine trilinearities (7), the set of spurious correspondences will obviously be larger than the one described by Proposition 9.

Example 4. Consider the camera matrices

$$\begin{pmatrix} 1 & 0 & 0 & 1 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 1 \\ 0 & 0 & 1 & 0 \end{pmatrix}, \begin{pmatrix} 1 & 0 & 0 & 0 \\ 0 & 1 & 0 & 0 \\ 0 & 0 & 1 & 1 \end{pmatrix}, \tag{9}$$

and the trilinearities $T_{47}, T_{58}, T_{69}, T_{48}$, where T_{ij} denotes the trilinear condition associated with the minor of (3) obtained by "excluding" rows i, j [11]. One can verify that these constraints describe a set $\mathcal{W}$ containing $\overline{\mathcal{I}}_3$ together with eight other (spurious) irreducible components. See Section B7 of supplemental material for details.

We conclude this section with the following proposition, that extends some results given in [14]. The proof is computational, and given in the supplementary material.

Proposition 10. *Consider three cameras.*

- *If the pinholes are non-collinear:*

1. *For* any *trilinearity T that does not vanish on the product of the trifocal lines, $\{B_{12}, B_{13}, B_{23}, T\}$ gives a strong characterization of the joint image.*
2. *The epipolar constraints $\{B_{12}, B_{13}, B_{23}\}$ uniquely determine camera geometry,* i.e.*, they give a weak characterization of the joint image.*

[2] The fact that a set of trilinear constraints does not linearly generate the whole space of trilinearities is not by itself sufficient to conclude that the conditions do not describe $\overline{\mathcal{I}}_3$: in fact, in the case of collinear pinholes, the trifocal trilinearities do characterize $\overline{\mathcal{I}}_3$. See the supplemental material for a discussion on this subtle point.

• *If the cameras have collinear pinholes:*

1. *A strong characterization of the joint image is given by* $\{B_{12}, B_{13}, B_{23}, T_1, T_2\}$ *where* T_1 *and* T_2 *are (sufficiently general) trilinear constraints.*
2. *Two epipolar constraints together with one (sufficiently general) trilinearity* $\{B_{12}, B_{13}, T\}$ *uniquely determine camera geometry,* i.e.*, they give a weak characterization of the joint image.*

4. The joint image in practice

We argue in in this section that that our theoretical description of the joint image and the associated multilinearities may be quite useful in practical settings.

4.1. Projective vs. euclidean cameras

One could object to the practicality of our analysis the fact that physical cameras are always euclidean, often with known internal parameters, thanks to Exif tags in JPEG images. However, the projective framework used in our presentation is simply more general and can easily be adapted to more practical and constrained settings. For example, in order to deal with actual pictures, one can introduce the *affine joint image* $\mathcal{J}_n(\mathcal{M}_1, \ldots, \mathcal{M}_n)$ for perspective cameras as the subset of $(\mathbb{R}^2)^n$ formed by n-tuples of *affine* correspondences $(\tilde{\boldsymbol{u}}_1, \ldots, \tilde{\boldsymbol{u}}_n)$. Since algebraic characterizations of (the closure of) $\mathcal{J}_n$ are effectively the same as for $\overline{\mathcal{I}}_n$, up to *dehomogeneization* (that is, setting $z_i = 1$ for all i), one realizes that all the results discussed in Section 3 also remain valid in the affine case. The only point to note is that, in special cases, weak descriptions of $\overline{\mathcal{I}}_n$ may specialize to strong ones for the affine joint image (namely when the spurious components do not appear in the affine charts). We also observe that the affine constraints are in general *not* multilinear but instead "multiaffine". Finally, using known intrinsic parameters basically amounts to considering only cameras the form $\mathcal{M}_i = (R, t)$, where R is a 3×3 rotation matrix (assuming normalized image coordinates), while all of our results clearly hold for any choice of 3×4 matrices of full rank. Restricting ourselves to calibrated cameras, the multilinear relations that characterize correspondences will simply *automatically* yield more constrained expressions (*e.g.*, a fundamental matrix will also satisfy the conditions for being an *essential matrix* [6]).

4.2. The distance to the joint image

Let us now illustrate a (potential) practical use of our analysis with one example. If we assume that we have measured a set of (noisy) image points matched across multiple images, and that we know an estimate of the camera parameters, then the *reprojection error* measures the mean-squared distance between the detected points and optimally reprojected points using the given cameras [6]. It is easy to realize that the computation of the reprojection error is equivalent to measuring the *distance to the joint image* in $(\mathbb{R}^2)^n$ of the given noisy correspondences [3]: this can be expressed as a constrained minimization problem in $(\mathbb{R}^2)^n$, where the contribution of a single n-tuple of measured affine points $\boldsymbol{c} \in (\mathbb{R}^2)^n$ is given by

$$\min_{\bar{\boldsymbol{c}} \in \mathcal{J}_n} ||\boldsymbol{c} - \bar{\boldsymbol{c}}||^2. \tag{10}$$

An exact optimization of (10) is expensive, and generally requires parameterizing the joint image using auxiliary variables associated with points in $\mathbb{R}^3$, then applying gradient descent-type methods.[3] For this reason, approximations of the reprojection error have been considered, for example the so-called *Sampson error* [15]. Essentially, this is a measure of the distance to a local linear approximation of the variety $\overline{\mathcal{J}}_n$. We refer to [6, Chapter 4.2.6] for details. However, a limitation of the Sampson error for $n \geq 3$ views is that it critically depends on the choice of equations for describing $\overline{\mathcal{J}}_n$, and using too many conditions will result in a higher computational cost. According to our discussion in Section 3, weak characterizations of the joint image may prove useful in this setting, since they provide good local approximations of the joint image but involve much fewer equations: it would be interesting to verify experimentally whether simple "weak" versions of the Sampson error (perhaps based on $2n - 3$ bilinear constraints) can actually be used to efficiently recover camera parameters for generic configurations.

5. Conclusions

The goal of this paper was to provide a clear and general overview on the geometry of the joint image and the different sets of algebraic conditions that can be used to characterize it. In summary, we have shown that for $n \geq 4$ generic views, only epipolar conditions are required: $3n-6$ constraints are sufficient for a complete description of correspondences (*i.e.*, a *strong* characterization), while $2n - 3$ are enough to recover camera geometry (a *weak* characterization). In the case of $n = 3$ views, bilinearities must be used with at least one trilinearity for strong characterizations, while all of the nine relations encoded in the trifocal tensor (or any subset of them) will generally yield some extraneous solutions.

In our opinion, a pleasant aspect of the joint image is that it allows revisiting most practical tasks in multi-view geometry (if not all!) in a natural setting that completely avoids the introduction of a three-dimensional coordinate frame (that would necessarily suffer projective ambiguity). We expect its role in computer vision algorithms to become increasingly important in the future.

Acknowledgments. This work was supported in part by the ERC grant VideoWorld, the Institut Universitaire de France, and ONR MURI N000141010934.

[3] Direct approaches that algebraically solve conditions for stationarity have also been proposed, however these are actually feasible only for two or three views, see [3, 8, 18].

References

[1] C. Aholt, B. Sturmfels, and R. Thomas. A Hilbert scheme in computer vision. *Canadian Journal of Mathematics*, 65:961–988, 2013. 2, 3

[2] D. A. Cox, J. Little, and D. O'Shea. *Ideals, varieties, and algorithms: an introduction to computational algebraic geometry and commutative algebra*. Springer, 2007. 2, 5

[3] J. Draisma, E. Horobeţ, G. Ottaviani, B. Sturmfels, and R. R. Thomas. The euclidean distance degree of an algebraic variety. *Foundations of Computational Mathematics*, pages 1–51, 2015. 8

[4] O. Faugeras and B. Mourrain. On the geometry and algebra of the point and line correspondences between n images. Technical Report 2665, INRIA, 1995. 1, 3, 5

[5] R. Hartley. Lines and points in three views and the trifocal tensor. *IJCV*, 22(2):125–140, 1997. 1, 6, 7

[6] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge University Press, 2000. 1, 2, 5, 7, 8

[7] R. I. Hartley and F. Schaffalitzky. Reconstruction from projections using grassmann tensors. In *Computer Vision-ECCV 2004*, pages 363–375. Springer, 2004. 3

[8] R. I. Hartley and P. Sturm. Triangulation. *Computer vision and image understanding*, 68(2):146–157, 1997. 8

[9] R. Hartshorne. *Algebraic Geometry*. Encyclopaedia of mathematical sciences. Springer, 1977. 4

[10] A. Heyden. Tensorial properties of multiple view constraints. *Mathematical Methods in the Applied Sciences*, 23:169–202, 2000. 3, 7

[11] A. Heyden and K. Astrom. Algebraic properties of multilinear constraints. *Mathematical Methods in the Applied Sciences*, 20(13):1135–1162, 1997. 1, 2, 3, 5, 6, 7

[12] G. Laman. On graphs and rigidity of plane skeletal structures. *Journal of Engineering mathematics*, 4(4):331–340, 1970. 6

[13] H. Longuet-Higgins. A computer algorithm for reconstructing a scene from two projections. *Nature*, 293:133–135, 1981. 1

[14] J. Ponce and M. Hebert. Trinocular geometry revisited. In *CVPR*, 2014. 1, 3, 4, 7

[15] P. D. Sampson. Fitting conic sections to very scattered data: An iterative refinement of the bookstein algorithm. *Computer graphics and image processing*, 18(1):97–108, 1982. 8

[16] A. Shashua. Algebraic functions for recognition. *PAMI*, 17(8):779–789, 1995. 1

[17] M. Spetsakis and Y. Aloimonos. Structure from motion using line correspondences. *IJCV*, 4(3):171–183, 1990. 1

[18] H. Stewenius, F. Schaffalitzky, and D. Nister. How hard is 3-view triangulation really? In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 686–693. IEEE, 2005. 8

[19] M. Thompson, R. Eller, W. Radlinski, and J. Speert, editors. *Manual of Photogrammetry*. ASPRS, 1966. Third Edition. 1

[20] B. Triggs. Matching constraints and the joint image. In *ICCV*, 1995. 1, 2

[21] J. Weng, T. Huang, and N. Ahuja. Motion and structure from line correspondences: closed-form solution, uniqueness, and optimization. *PAMI*, 14(3):318–336, 1992. 1