GPSfM: Global Projective SFM Using Algebraic Constraints on Multi-View Fundamental Matrices

Yoni Kasten* Amnon Geifman* Meirav Galun Ronen Basri
Weizmann Institute of Science
{yoni.kasten,amnon.geifman,meirav.galun,ronen.basri}@weizmann.ac.il

Abstract

This paper addresses the problem of recovering projective camera matrices from collections of fundamental matrices in multiview settings. We make two main contributions. First, given $\binom{n}{2}$ fundamental matrices computed for n images, we provide a complete algebraic characterization in the form of conditions that are both necessary and sufficient to enabling the recovery of camera matrices. These conditions are based on arranging the fundamental matrices as blocks in a single matrix, called the n-view fundamental matrix, and characterizing this matrix in terms of the signs of its eigenvalues and rank structures. Secondly, we propose a concrete algorithm for projective structure-from-motion that utilizes this characterization. Given a complete or partial collection of measured fundamental matrices, our method seeks camera matrices that minimize a global algebraic error for the measured fundamental matrices. In contrast to existing methods, our optimization, without any initialization, produces a consistent set of fundamental matrices that corresponds to a unique set of cameras (up to a choice of projective frame). Our experiments indicate that our method achieves state of the art performance in both accuracy and running time.

1. Introduction

This paper considers the problem of recovering projective camera matrices from collections of fundamental matrices. Many multiview structure from motion (SFM) pipelines begin, given n images, $I_1, ..., I_n$, by robustly estimating fundamental matrices between image pairs from collections of point matches, e.g., using RANSAC. However, in typical settings, only a subset of the $\binom{n}{2}$ pairwise fundamental matrices can be estimated, and the estimated matrices may be subject at times to significant errors. Moreover, fundamental matrices are defined through a homogeneous equation and can thus assume any scale factor, but a consistent setting of these scales is important in multi-view settings [20, 22] (in analogy to resolving the distance between cameras in a calibrated setting). Consequently, accurate recovery of camera matrices is interesting both from theoretical and practical standpoints. Our paper makes contributions to both of these aspects.

An important theoretical question is, given a collection of $\binom{n}{2}$ fundamental matrices, whether these fundamental matrices are *consistent*, in the sense that there exist n camera matrices that produce these fundamental matrices. Below we address this question by providing a set of algebraic constraints that form both necessary and sufficient conditions for the consistency of fundamental matrices. Our formulation, which extends the partial list of necessary conditions introduced in [22], considers the symmetric matrix F of size $3n \times 3n$, formed by stacking all $\binom{n}{2}$ fundamental matrices. It provides a complete characterization of F in terms of the signs of its eigenvalues and rank patterns.

An advantage of our algebraic characterization is that it can readily be used to construct optimization algorithms to recover camera matrices. In the second part of this paper we introduce an efficient algorithm to recover projective camera matrices directly from measured fundamental matrices. Our algorithm, which utilizes the consistency constraints presented in this paper, uses global optimization for camera recovery, overcoming noise and missing measurements. It further avoids one of the main difficulties in a previous approach [22] – the need to accurately recover a scale factor for each of the estimated fundamental matrix. This allows us to obtain state-of-the-art results without any initialization. We demonstrate the utility of our method by applying it to uncalibrated image collections of various sizes. Our experiments indicate that our method outperforms previous methods in both accuracy and runtime.

1.1. Previous work

The recovery of projective camera matrices was addressed in several lines of work.

Incremental algorithms [13, 15, 19] process the images sequentially, interleaving camera and depth recovery for ev-

*Equal contributors

ery new image. Such methods can be sensitive to the order of processing and may suffer from drift, due to accumulation of errors. The use of bundle adjustment for every new image reduces such drift, but is computationally demanding.

Factorization-based methods [5, 6, 12, 15, 16, 23, 25] factor a measurement matrix that includes all point matches across views into an (unknown) product of camera matrices, depth values, and 3D point locations. These methods typically yield very large optimization problems and are often approached by splitting the problem into smaller subproblems.

Global methods. A number of "global methods" were proposed recently demonstrating both accurate and efficient recovery of camera matrices from pairwise measurements (essential or fundamental matrices), mostly in a calibrated setting [11, 27, 18, 7, 2]. Sweeney el al. [24] attempts to improve the consistency of fundamental matrices by minimizing the discrepancy of reprojected points in three views (through an "epipolar point transfer"). They, however, cannot achieve projective recovery since their method does not guarantee the consistency of the improved fundamental matrices. Sengupta et al. [22] attempt to enforce rank constraints on the measured fundamental matrices. Complicated by the need to simultaneously recover suitable scale factors, their method is sensitive to errors and requires highly accurate initialization, which was achieved by applying the state-of-the-art, *calibrated* LUD algorithm [18], defeating the purpose of projective camera recovery without calibration.

Solvability of viewing graphs. A number of papers seek to design algorithms that can identify "solvable viewing graphs" [14, 18, 20, 24]. A *viewing graph* captures the pattern of missing fundamental matrices. Let $G = (V, E)$ be a graph such that a node $v_i \in V$ represents image I_i and an edge $e_{ij} \in E$ exists if the fundamental matrix F_{ij} relating I_i and I_j is available. G is called *solvable* if the corresponding n camera matrices can be determined uniquely (up to a 4×4 projective transformation) despite the missing fundamental matrices. Identifying solvable graphs is equivalent to asking, given a partial set of fundamental matrices, if F can be completed uniquely to satisfy our algebraic constraints.

Finally, both our paper and [22] explore algebraic properties of multiview fundamental matrices (MVFs) and propose optimization schemes for utilizing these properties toward camera pose recovery. However, these two papers differ in several significant respects. First, [22] provides a set of necessary algebraic constraints, for the consistency of MVF F, while we provide a complete set of necessary and sufficient algebraic conditions. Moreover, our conditions are specified directly in terms of the MVFs, in contrast to [22] which rely on the construction of an auxiliary, unknown matrix. Our direct formulation further leads to a significantly simpler optimization algorithm, including a new algorithm for recovering the projective camera matrices, from a consistent MVF F, which is lacking in [22]

2. Algebraic constraints of n-view fundamental matrices

Let $I_1, \ldots, I_n$ denote a collection of n images of a static scene captured respectively by projective cameras $P_1, ..., P_n$. Each camera P_i is represented by a 3×4 matrix $P_i = K_i R_i^T [I, -\mathbf{t}_i]$, where K_i is a 3×3 calibration matrix, $\mathbf{t}_i \in \mathbb{R}^3$ and $R_i \in SO(3)$ respectively denote the location and orientation of P_i in some global coordinate system, and I denotes the 3×3 identity matrix. Below we further denote $V_i = K_i^{-T} R_i^T$, so $P_i = V_i^{-T}[I, -\mathbf{t}_i]$. Consequently, let $\mathbf{X} = (X, Y, Z)^T$ be a scene point in the global coordinate system. Its projection onto I_i is given by $\mathbf{x}_i = \mathbf{X}_i / Z_i$, where $\mathbf{X}_i = (X_i, Y_i, Z_i)^T = K_i R_i^T (\mathbf{X} - \mathbf{t}_i)$.

We next denote the fundamental matrix between images I_i and I_j by F_{ij}. In [2, 22] it was shown that F_{ij} can be written as

$$F_{ij} = K_i^{-T} R_i^T (T_i - T_j) R_j K_j^{-1} = V_i (T_i - T_j) V_j^T, \quad (1)$$

where $T_i = [\mathbf{t}_i]_\times$. It can be readily verified that this definition of F_{ij} is consistent with the standard properties of fundamental matrices [8], including (a) $P_i^T F_{ij} P_j$ is skew symmetric, and (b) $\mathbf{e}_{ik}^T F_{ij} \mathbf{e}_{jk} = 0$, where $\mathbf{e}_{ik}$ denotes the projection of the center of camera k onto camera i, i.e., $\mathbf{e}_{ik} = P_i \mathbf{t}_k$.

Note, however, that (1) attributes a scale to each F_{ij} that relates it to some global coordinate system. In practice, given two images, a fundamental matrix is determined only up to a multiplicative factor. We will denote an estimated fundamental matrix by $\hat{F}_{ij}$ and assume, in case it is estimated accurately, that $\hat{F}_{ij} = \lambda_{ij} F_{ij}$ with unknown $\lambda_{ij} \neq 0$.

We next construct a matrix from all $\binom{n}{2}$ fundamental matrices.

Definition 1. *A matrix $F \in \mathbb{S}^{3n}$, whose 3×3 blocks are denoted by F_{ij}, is called an n-view fundamental matrix if $rank(F_{ij}) = 2$ for all $i \neq j$ and $F_{ii} = 0$.*

We use $\mathbb{S}^{3n}$ to denote the space of all $3n \times 3n$ symmetric matrices. The symmetry of F implies that $F_{ij} = F_{ji}^T$.

Definition 2. *An n-view fundamental matrix F is called consistent if there exist camera matrices $P_1, ..., P_n$ of the form $P_i = V_i^{-T}[I, \mathbf{t}_i]$ such that $F_{ij} = V_i([\mathbf{t}_i]_\times - [\mathbf{t}_j]_\times) V_j^T$.*

A consistent F, therefore, takes the form

$$F = \begin{pmatrix} 0 & F_{12} & \ldots & F_{1n} \\ F_{21} & 0 & \ldots & F_{2n} \\ \vdots & & & \vdots \\ F_{n1} & F_{n2} & \ldots & 0 \end{pmatrix}$$

and each F_{ij} is scaled properly in accordance with the global coordinate system. We further refer to the $3 \times 3n$ i^{th} block row of F by F_i.

Our main theoretical result is summarized below in Theorem 1, which specifies a set of necessary and sufficient algebraic conditions for the consistency of F in terms of its eigenvalue sign and rank patterns. These, in turn, will be used in later sections to construct a new optimization algorithm for global recovery of projective camera matrices from noisy fundamental matrices.

Theorem 1. *An n-view fundamental matrix F is consistent with a set of n cameras whose centers are not all collinear if, and only if, the following conditions hold:*

1. *$Rank(F) = 6$ and F has exactly 3 positive and 3 negative eigenvalues.*
2. *$Rank(F_i) = 3$ for all $i = 1, ..., n$.*

Below we provide a proof sketch. The full proof is deferred to the supplementary material. To prove the theorem we first state that for a symmetric rank 6 matrix F the following three conditions are equivalent:

(i) F has exactly 3 positive and 3 negative eigenvalues.

(ii) $F = XX^T - YY^T$ with $X, Y \in \mathbb{R}^{3n\times 3}$ and $rank(X) = rank(Y) = 3$.

(iii) $F = UV^T + VU^T$ with $U, V \in \mathbb{R}^{3n\times 3}$ and $rank(U) = rank(V) = 3$.

In particular, using the eigen-decomposition of F, let $F\boldsymbol{x}_i = \alpha_i \boldsymbol{x}_i$ and $F\boldsymbol{y}_i = -\beta \boldsymbol{y}_i$, $\alpha_i, \beta_i > 0$, the columns of X and Y respectively may include $\sqrt{\alpha_i}\boldsymbol{x}_i$ and $\sqrt{\beta_i}\boldsymbol{y}_i$, and U, V are related to X, Y through

$$U = (X - Y)/\sqrt{2}, \quad V = (X + Y)/\sqrt{2}. \qquad (2)$$

Next, to show the necessary condition, let F be a consistent, n-view fundamental matrix, then clearly (1) can be written in matrix form as $F = UV^T + VU^T$, where $U, V \in \mathbb{R}^{3n\times 3}$ whose 3×3 blocks respectively are $U_i = V_i T_i$ and V_i, implying condition 1. Condition 2 holds because not all cameras are collinear, since if conversely $rank(F_i) < 3$ for some i then there exists a 3-vector $\mathbf{e} \neq 0$ such that $F_i^T \mathbf{e} = 0$, and therefore $\forall j$ $F_{ji}\mathbf{e} = 0$, implying, in contradiction, that the camera centers are all collinear.

To establish the sufficient condition, let F be an n-view fundamental matrix that satisfies conditions 1 and 2. Condition 1 (along with (iii)) implies that $F_{ij} = U_i V_j^T + V_i U_j^T$. Next, $F_{ii} = 0$ implies that $U_i V_i^T$ is skew symmetric, and so $\forall i$ either $rank(U_i) = 2$ or $rank(V_i) = 2$. Next, as we show in the supplementary material, $rank(F_i) = 3$ implies WLOG that $\forall i, rank(V_i) = 3$ and $rank(U_i) = 2$. This and the skew-symmetry of $U_i V_i^T$ imply that $V_i^{-1} U_i$ is skew-symmetric. Denote this matrix by $T_i = [\mathbf{t}_i]_\times$, we obtain $F_{ij} = V_i(T_i - T_j)V_j^T$, establishing that F is consistent. Finally, $\{\mathbf{t}_i\}_{i=1}^n$ are not all collinear, since, otherwise $\exists i$ and $\exists \mathbf{e} \neq 0$ such that $\forall j$ $F_{ji}\mathbf{e} = 0$, implying that $F_i^T \mathbf{e} = 0$, contradicting the full rank of F_i.

Theorem 1 also provides a practical tool for projective reconstruction. Given a set of (possibly noisy) pairwise fundamental matrices we can use constrained, low-rank optimization to recover a matrix that satisfies conditions 1-2. Then, we can use the obtained n-view fundamental matrix to recover the underlying camera matrices. This is summarized in the following corollary.

Corollary 1. *Projective reconstruction: Let $F \in \mathbb{S}^{3n}$ be a consistent n-view fundamental matrix, then it is possible to explicitly determine camera matrices $P_1, ..., P_n$ that are consistent with F.*

Proof. The claim is justified by the following construction, which relies on the proof of Theorem 1.

1. Since F satisfies condition 1 we can use its eigen-decomposition to express it as $F = XX^T - YY^T$, and then construct U and V using (2).
2. Now, WLOG, $rank(V_i) = 3$ and $rank(U_i) = 2$ for all $i = 1, ..., n$ (or else U and V should be interchanged).
3. We next define $T_i = V_i^{-1} U_i$. T_i is skew symmetric, and we denote $T_i = [\mathbf{t}_i]_\times$.
4. By construction $F_{ij} = V_i(T_i - T_j)V_j^T$, implying that $P_i = [V_i^{-T} | -V_i^{-T}\mathbf{t}_i]$ form a consistent set of camera matrices.

This construction is unique up to a 4×4 projective transformation. □

Let F be a consistent n-view fundamental matrix. Clearly, if we scale differently any of its blocks, $\lambda_{ij} F_{ij}$ with $\lambda_{ij} \neq 0$, then in general F ceases to be consistent. In particular, it maintains condition 2 of Theorem 1, but its rank is no longer 6. Note, however, that we can scale each block-row (and by symmetry column) of F differently and maintain both the conditions of the theorem; i.e., let $S = diag(s_1 I, .., s_n I)$ with $s_i \neq 0$, then SFS is consistent if and only if F is consistent. (Such scaling is equivalent to scaling the projective camera matrices.) This, in fact, implies that we need to determine only $\binom{n}{2} - n$ of the scale factors and set the rest of the n scales arbitrarily.

Up to this point we considered n-view matrices that include all $\binom{n}{2}$ pairwise fundamental matrices. In real applications, however, often only a subset of the pairwise matrices can be computed. Additionally, the estimated fundamental matrices are improperly scaled and may suffer from large inaccuracies. In these cases we may want to reconstruct the cameras from partial subsets of fundamental matrices. Indeed, our algorithm, presented later in Sec. 3, recovers a consistent set of camera matrices from triplets of images, allowing us to handle missing fundamental matrices and to remove outliers. The following theorem establishes that, with proper intersection, camera matrices are determined uniquely (up to the usual 4×4 projective ambiguity) from consistent sub-matrices. We first need the following definition:

Definition 3. *Let $F \in \mathbb{R}^{3n \times 3n}$ and let $F^1, ..., F^k$ be block sub-matrices of F, with $F^i \in \mathbb{R}^{3m_i \times 3m_i}$, $3 \leq m_i \leq n$. $\{F^1, ..., F^k\}$ is called a **consistent cover** of F if each F^i forms a consistent multi-view fundamental matrix and each diagonal element of F is contained in at least one of $F^1, ..., F^k$.*

Theorem 2. *Let $F \in \mathbb{R}^{3n \times 3n}$ and let $F^1, ..., F^k$ form a consistent cover of F. If for all $2 \leq m \leq k$, there exists $l < m$ such that F^l, F^m overlap in at least one fundamental matrix, then there exists a unique n-view consistent fundamental matrix $\bar{F}$ (up to n scale factors) whose blocks $\bar{F}_{ij} = \lambda_{ij} F_{ij}$ with some $\lambda_{ij} \neq 0$ for all F_{ij} that belong to any of $F^1, ..., F^k$.*

Proof. We prove this by induction on k. We begin with $k = 2$. By Corollary 1, F^1, F^2 define two sets of camera matrices $\mathcal{P}^1, \mathcal{P}^2$ that are consistent with respect to F^1, F^2, respectively. Since F^1 and F^2 share a fundamental matrix F_{ij}, F_{ij} corresponds to a pair of cameras in $\mathcal{P}^1$ and a second pair in $\mathcal{P}^2$ so that the two pairs are equal up to a projective homography ([8], p. 254). Consequently, $\mathcal{P}^2$ can be mapped to the projective frame of $\mathcal{P}^1$ to form a set of n camera matrices [20], that in turn determine a unique (up to n global scale factors) consistent n-view matrix, $\bar{F}$. Now, each fundamental matrix, in either F^1 or F^2, corresponds to two cameras from this set of n cameras and hence has exactly the same entries as in $\bar{F}$ up to scale.

This argument can now be repeated inductively to prove the theorem for all $k > 2$. □

3. Method

Given images $I_1, ..., I_n$, we assume a standard robust method (e.g., RANSAC) is used to estimate the pairwise fundamental matrices, where we denote by $\Omega = \{\hat{F}_{ij}\}$ the set of estimated fundamental matrices. In general, only a subset of the $\binom{n}{2}$ pairwise fundamental matrices are estimated, due to, e.g., occlusion, large motion, or changes in brightness, and the available estimates are noisy. An additional complication is that to make these fundamental matrices consistent they must each be scaled by an unknown factor to fit the global coordinate frame. Our aim therefore is to find a consistent n-view matrix $F \in \mathbb{S}^{3n}$ that is as similar as possible to the measured fundamental matrices. Straightforward optimization of this problem is difficult as it yields a nonlinear optimization formulation with rank constraints, as in [22], which required initialization by a high quality method.

Below we introduce a novel method that utilizes global optimization and yet circumvents the need to recover the scale factors. Our method works by enforcing the consistency constraints of Theorem 1 on 3-view fundamental matrices and by maintaining their intersections, as is required by Theorem 2, to obtain a global reconstruction. This yields an optimization problem, with simple and efficient formulation, that directly uses the measured data without the need to incorporate any initialization. We solve this optimization using the alternating direction method of multipliers (ADMM) [3], where each step in the ADMM has a closed form solution.

We avoid recovering scale factors by enforcing consistency for image triplets. As we explained in Sec. 2, we only need to determine $\binom{n}{2} - n$ of the scale factors, while the rest can be set arbitrarily. A consequence of this is that there are no scale factors for $n = 3$, as proved below.

Corollary 2. *A consistent 3-view fundamental matrix is invariant to arbitrarily scaling its constituent fundamental matrices.*

Proof. Let F be a consistent 3-view fundamental matrix whose blocks are defined as $F_{ij} = V_i(T_i - T_j)V_j^T$, and let $\widetilde{F}$ be a 9×9 matrix whose blocks are defined to be $\widetilde{F}_{ij} = s_{ij} F_{ij}$ where $s_{ij} \neq 0$ are arbitrary scale factors. Without loss of generality we can assume that the number of negative scale factors is even (otherwise we can multiply the entire matrix by -1). Therefore, $s_1 = (\frac{s_{12}s_{13}}{s_{23}})^{\frac{1}{2}}$, $s_2 = (\frac{s_{23}s_{12}}{s_{13}})^{\frac{1}{2}}$, and $s_3 = (\frac{s_{13}s_{23}}{s_{12}})^{\frac{1}{2}}$ determine real values such that $s_1 s_2 = s_{12}$, $s_1 s_3 = s_{13}$, and $s_2 s_3 = s_{23}$. Let $\widetilde{V}_i = s_i V_i$ for $i = 1 \ldots 3$, we get that

$$\begin{aligned} \widetilde{F}_{ij} = s_{ij} V_i (T_i - T_j) V_j^T &= s_i V_i (T_i - T_j) s_j V_j^T \\ &= \widetilde{V}_i (T_i - T_j) \widetilde{V}_j^T \end{aligned} \quad (3)$$

Therefore $\widetilde{F} = SFS$ with $S = diag(s_1 I, s_2 I, s_3 I)$, and hence it is consistent. □

This corollary implies that for 3-view fundamental matrices consistency is invariant under *any* choice of scale factors. Our optimization formulation relies on this observation to avoid the need to estimate the scale factors during optimization. In particular, we introduce a global optimization scheme that enforces the consistency of triplets of views, while simultaneously enforcing the compatibility of the different triplets. In the rest of this section we first formulate our optimization problem and discuss how to solve it with ADMM. Then we discuss how to select minimal subsets of triplets to speed up the optimization and finally show how the results of our optimization can be used to reconstruct the n cameras.

3.1. Optimization

Our input set of estimated fundamental matrices $\{\hat{F}_{ij}\}$ determines a *viewing graph* $G = (V, E)$ with nodes $v_1, ..., v_n$, corresponding to the n cameras, and $e_{ij} \in E$ if $\hat{F}_{ij}$ belongs to the collection of the estimated fundamental matrices. Let τ denote a collection of m 3-cliques in G, $m \leq \binom{n}{3}$. The collection τ may include all the 3-cliques in G, or a subset, as we explain in Sec. 3.3. We index the elements of τ by $k = 1, ..., m$, where $\tau(k)$ denotes the k^{th} triplet. The selection of τ induces a partial selection of estimated fundamental matrices, Ω, that participate in the optimization process. In our construction, if $\hat{F}_{ij} \in \Omega$ then $\hat{F}_{ij}^T = \hat{F}_{ji} \in \Omega$.

We define the measurement matrix $\hat{F} \in \mathbb{S}^{3n}$ to include all $\hat{F}_{ij} \in \Omega$ in their corresponding 3×3 block while setting the rest of the blocks to $0_{3\times 3}$. In the optimization process, we look for a matrix $F \in \mathbb{S}^{3n}$ that is as close as possible to $\hat{F}$ under the constraint that its 9×9 blocks, induced by $\{\tau(k)\}_{k=1}^m$ and denoted as $\{F_{\tau(k)}\}_{k=1}^m$, are consistent. In general, such an F is inconsistent (since its scale factors are incompatible across triplets) and incomplete, but, based on Theorem 2, it uniquely determines the corresponding projective cameras.

We next introduce our constrained optimization problem

$$
\begin{aligned}
\min_{F} \quad & \sum_{k=1}^{m} ||F_{\tau(k)} - \hat{F}_{\tau(k)}||_F^2 \\
\text{s.t.} \quad & F = F^T \\
& F_{ii} = 0_{3\times 3} \qquad i = 1, \dots, n \\
& rank(F_{\tau(k)}) = 6 \qquad k = 1, \dots, m.
\end{aligned}
\tag{4}
$$

Solving (4) is challenging due to the rank constraints. As mentioned above, we approach this problem using ADMM. To that end, $2m$ auxiliary matrix variables of size 9×9 are added: m variables duplicating $\{F_{\tau(k)}\}_{k=1}^m$, denoted $\{B_k\}_{k=1}^m$, and m Lagrange multipliers $\{\Gamma_k\}_{k=1}^m$, yielding the objective

$$
\begin{aligned}
\max_{\Gamma_k} \min_{F, B_1, \dots, B_m} \quad & \sum_{k=1}^{m} L(F_{\tau(k)}, B_k, \Gamma_k) \\
\text{s.t.} \quad & F = F^T \\
& F_{ii} = 0_{3\times 3} \qquad i = 1, ..., n \\
& rank(B_k) = 6 \qquad k = 1, ..., m,
\end{aligned}
\tag{5}
$$

where

$$L(F_{\tau(k)}, B_k, \Gamma_k) = \alpha||\hat{F}_{\tau(k)} - F_{\tau(k)}||_F^2 + ||B_k - F_{\tau(k)} + \Gamma_k||_F^2.$$

We initialize the auxiliary variables at $t = 0$ with

$$B_k^{(0)} = \hat{F}_{\tau(k)}, \quad \Gamma_k^{(0)} = 0$$

and then alternate between the following three steps, where at each step we update the values of the variables at iteration t given their values at $t - 1$.

(i) **Solving for F.**

$$
\begin{aligned}
\operatorname*{argmin}_{F} \quad & \sum_{k=1}^{m} \alpha||\hat{F}_{\tau(k)} - F_{\tau(k)}||_F^2 \\
& \quad + ||B_k^{(t-1)} - F_{\tau(k)} + \Gamma_k^{(t-1)}||_F^2 \\
\text{s. t.} \quad & F = F^T \\
& F_{ii} = 0_{3\times 3} \qquad i = 1, ..., n.
\end{aligned}
\tag{6}
$$

In practice, we explicitly maintain the equality constraints over F, i.e., F is symmetric with zero block diagonal. Therefore, at each iteration t we can solve only for the triangular upper part of F, i.e., $\{F_{ij} | \hat{F}_{ij} \in \Omega, i < j\}$. This yields an unconstrained convex quadratic objective in these variables, and hence it admits a closed form solution. Let N_{ij} be the number of 3-cliques in τ that include the edge e_{ij}. Then, for each such triplet $\tau(k)$ we denote the variables corresponding to the i, j block as $B_k(i, j)$, $\Gamma_k(i, j)$, and $\hat{F}_{\tau(k)}(i, j)$. This yields the following update rule

$$
\begin{aligned}
F_{ij}^{(t)} = \quad & \frac{1}{N_{ij}(1+\alpha)} \sum_{k=1}^{N_{ij}} B_k^{(t-1)}(i, j) + \Gamma_k^{(t-1)}(i, j) \\
& \qquad + \alpha \hat{F}_{\tau(k)}(i, j).
\end{aligned}
\tag{7}
$$

(ii) **Solving for B_k.** For all $k = 1, \dots, m$

$$
\begin{aligned}
& B_k^{(t)} = \operatorname*{argmin}_{B_k} ||B_k - F_{\tau(k)}^{(t)} + \Gamma_k^{(t-1)}||_F^2 \\
& \text{s.t} \quad rank(B_k) = 6.
\end{aligned}
\tag{8}
$$

Here, the closed form solution is

$$B_k^{(t)} = SVP(F_{\tau(k)}^{(t)} - \Gamma_k^{(t-1)}, 6) \tag{9}$$

where $SVP(A, p)$ denotes the singular value projection of the matrix A to rank p.

(iii) **Updating Γ_k.** For all $k = 1, \dots, m$

$$\Gamma_k^{(t)} = \Gamma_k^{(t-1)} + B_k^{(t)} - F_{\tau(k)}^{(t)}. \tag{10}$$

Note that the constraints in (4) cover only a subset of the constraints in Theorem 1, and in particular they do not restrict the sign pattern of the eigenvalues of $F_{\tau(k)}$, the rank 2 of F_{ij}, or the rank 3 of F_i. Our experiments, however, indicate that with the amounts of noise prevalent in existing datasets, and by removing collinear triplets, our solutions always satisfy these constraints to a good numerical precision.

3.2. Camera recovery

We use the optimized matrix F to recover the corresponding projective cameras. Since F is generally inconsistent we cannot use the eigen-decomposition method described in Corollary 1. However, the 9×9 sub-matrices $F_{\tau(k)}$ are consistent, and by construction they form a triplet cover of G. It is therefore straightforward to traverse the graph G_τ and, using Theorem 2, apply a homography to each three cameras corresponding to $F_{\tau(k)}$, $k = 1, ..., m$, to bring them all to a common projective frame. Note that this process is exact, since all $F_{\tau(k)}$ are consistent.

3.3. Constructing triangle cover

Eq. (4) enforces the consistency of 3-view sub-matrices. In principle, this formulation can be applied to all 3-cliques of G (although outlier and collinear triplet removal may be needed). It is however more efficient (and suffices for reconstruction) to apply this to a subset of the triplets, provided the triplets produce a solvable viewing graph.

Given a viewing graph $G = (V, E)$, we consider further a graph $G_\tau = (V_\tau, E_\tau)$ whose nodes $v' \in V_\tau$ represent 3-clique in G and an edge $e'_{kl} \in E_\tau$ exists if triangles v'_k and v'_l share two images. We call G_τ a *triplet cover* of G if every node $v_i \in V$ belongs to at least one vertex in V_τ.

Our objective is to find a small and reliable triplet cover of G. We do this heuristically as follows. We associate a weight w_{ij} with each edge $e_{ij} \in E$, where w_{ij} counts the number of inliers of pairwise correspondences, identified for $\hat{F}_{ij}$. We then find N_G edge-disjoint maximal spanning trees for G, and use this set to produce a triplet cover for G, denoted G_τ, in a similar way to [13]. Next, we prune G_τ greedily, removing triplets whose cameras are collinear and triplets whose consistency scores are low. To measure collinearity we divide the distance between the two epipoles in each image by their average distance from the image center and average these ratios over the three images. Denote this measure by l_k we remove triplets with $l_k < \delta_1$. We further measure consistency as $c_k = \|F_{\tau(k)} - \hat{F}_{\tau(k)}\|_F$, where $\hat{F}_{\tau(k)}$ is the measured 3-view fundamental matrix associated with the triplet v'_k and $F_{\tau(k)}$ is its closest consistent triplet calculated using our ADMM optimization for only this triplet. (This can be done very efficiently.) Finally, we sort the remaining triplets by their stability scores, defined as $s_k = l_k^{\delta_2}/c_k$, and greedily remove triplets of low score while maintaining the connectivity of G_τ and its cover of G, see an illustration in Fig. 1.

4. Experiments

4.1. Structure from motion pipeline

We used our method to construct a projective SFM pipeline. The pipeline obtains pairwise fundamental matrices computed with RANSAC. As described in Sec. 3, we first construct a reliable triangle cover. We next use ADMM to solve (4), obtaining consistent 3-view submatrices, and use them to construct camera matrices. In postprocessing we use projective bundle adjustment (BA) to improve our camera recovery and 3D point reconstruction. Note that, similar to global Euclidean SFM methods [2, 7, 11, 18, 27], we only apply BA once at the end of the pipeline. For this step we triangulate point tracks and apply projective bundle adjustment to the camera matrices and the point matches. Similar to [21], after convergence, 3D points are re-triangulated and a few additional iterations of BA are applied for better convergence. Our pipeline is summarized in Alg. 1.

Algorithm 1: Projective SFM Pipeline

Input : Fundamental matrices $\Omega = \{\hat{F}_{ij}\}$
Viewing Graph: $G = (V, E)$
Tracks of point matches (for BA)
Output: Projective reconstruction of Cameras and Points
$G_\tau \leftarrow$Select a triplet cover for G (Sec. 3.3)
Form the n-view measurement matrix $\hat{F} \leftarrow \{\hat{F}_{ij}\}$
Solve (4) using ADMM:
Initialize $B_k^{(0)} = \hat{F}_{\tau(k)}, \;\; \Gamma_k^{(0)} = 0, t = 1$
for $t = 1, ..., N_{it}$ **do**
  Update $F^{(t)}$ using (7)
  For $k = 1, ..., m$, update $B_k^{(t)}$ using (9)
  For $k = 1, ..., m$, update $\Gamma_k^{(t)}$ using (10)
end
For $k = 1, ..., m$, retrieve camera matrices for triplet $\tau(k)$ (Corollary 1)
$\mathcal{P} \leftarrow$ Bring cameras to a global projective frame of reference (Sec. 3.2)
$\mathcal{X} \leftarrow$ Triangulate points from points tracks
$\mathcal{P}, \mathcal{X} \leftarrow$ Refine solution using Bundle Adjustment
Return $\mathcal{P}, \mathcal{X}$

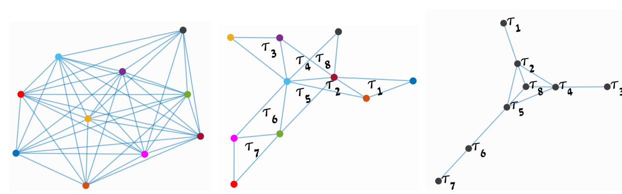


Figure 1: Building a triplet cover for the House dataset (10 cameras). From left to right, the viewing graph, the final viewing graph, and the corresponding triplet cover.

4.2. Implementation details

Before optimization we normalize the input fundamental matrices $\hat{F}_{ij}$, as in [9], by $\hat{F}^n_{ij} = N_i^{-T} \hat{F}_{ij} N_j^{-1}$, where $N_i \in$

Table 1: Reprojection error and run time obtained in our experiments.

Dataset	#points	#Images	Error(pixels)				Time(s)			
			Ours	PPSFM	Sengupta	Var-Pro	Ours	PPSFM	Sengupta	Var-Pro
Dino 319	319	36	**0.4314**	0.5042	0.6134	0.6157	**3.07**	3.24	46.40	5.48
Dino 4983	4983	36	**0.4205**	0.4442	0.5795	0.5961	**4.80**	15.19	50.35	384.01
Corridor	737	11	**0.2596**	0.276	0.2765	0.2741	**1.12**	1.31	21.05	45.16
House	672	10	**0.3399**	0.3687	0.5984	0.3719	0.95	**0.70**	18.10	55.19
Gustav Vasa	4249	18	**0.1564**	0.1687	0.2591	0.1671	**3.18**	7.06	32.93	326.48
Folke Filbyter	21150	40	**0.258**	–	–	26.6054	**6.07**	–	–	2.33E+04
Park Gate	9099	34	**0.3109**	0.3447	0.3288	0.5489	**11.33**	25.85	62.6	1600
Nijo	7348	19	**0.3901**	0.4412	0.4173	0.417	**5.70**	10.33	32.22	148.74
Drinking Fountain	5302	14	**0.2806**	0.3125	0.2942	0.2942	**2.85**	5.80	26.35	82.07
Golden Statue	39989	18	**0.223**	0.24	0.2368	0.2272	**7.07**	20.45	86.8	3890
Jonas Ahls	2021	40	**0.1845**	0.2108	0.1979	0.197	**5.29**	15.22	46.64	90.81
De Guerre	13477	35	**0.2609**	0.2891	0.2728	0.2715	**13.79**	27.92	103.70	282.87
Dome	84792	85	**0.2354**	0.2507	1.8991	0.2413	**109.83**	171.83	970	3.78E+04
Alcatraz Courtyard	23674	133	**0.5162**	0.5592	5.3641	0.5366	**65.04**	113.17	537	3210
Alcatraz Water Tower	14828	172	**0.4704**	0.5972	0.5003	3.0353	87.11	**68.31**	539	1080
Cherub	72784	65	**0.7408**	0.7921	–	Out of memory (16GB)	**34.87**	81.42	–	–
Pumpkin	69335	195	**0.5959**	–	–	Time Limit (12H)	**130.74**	–	–	–
Sphinx	32668	70	**0.3366**	0.3669	0.3508	0.3486	**23.36**	49.05	191	2.53E+04
Toronto University	7087	77	0.5417	0.2588	0.2557	**0.2556**	**25.65**	100.27	779.5	335.9708
Sri Thendayuthapani	88849	98	0.6113	0.3517	**0.3204**	Out of memory (16GB)	**248.72**	418.21	1070	–
Porta san Donato	25490	141	**0.3992**	0.4352	4.5186	0.4155	**78.22**	87.84	771	2060
Buddah Tooth	27920	162	**0.5957**	0.8583	1.7853	0.6245	**62.17**	84.44	792	5530
Tsar Nikolai I	37857	98	**0.2897**	0.309	0.3021	0.3013	**62.71**	95.68	422	8700
Smolny Cathedral	51115	131	**0.4639**	0.5079	0.4773	Out of memory (16GB)	**197.72**	264.04	640.87	–
Skansen Kronan	28371	131	0.4424	0.4414	0.4477	**0.4291**	**102.75**	165.77	1000	2610

$\mathbb{R}^{3\times3}$ normalizes the location of interest points in image I_i to have zero mean and unit variance. In data sets where the point matches distribute non-isotropically, we normalize the variance separately in each axis. After the optimization F_{ij} is denormalized by $N_i^T F_{ij} N_j$.

For the optimization, we set $\alpha = 0.001$ (in (6)) and perform $N_{it} = 1000$ iterations of ADMM updates. For the triplet selection procedure (Sec. 3.3) we set $N_G = 5$ and $\delta_1 = 0.03$. We set δ_2 by the following condition, if the average non-collinearity measure, l_k, exceeds 0.5 we set $\delta_2 = 0$. Otherwise, it means the data is highly collinear, and so we set $\delta_2 = 1.2$.

To produce 3D points from multiview tracks we used the Matlab linear triangulation code of [8]. We implemented projective bundle adjustment using the Ceres [1] non linear least squares optimization package and used Huber loss (with parameter 0.1) for robustness. We performed up to 100 iterations of bundle adjustment. Our code is implemented in Matlab on an Intel processor i-7 7700 with 16GB RAM.

4.3. Results

We tested the method on several projective datasets from [17] and VGG [26] and compared the results to state-of-the-art projective reconstruction pipelines, including:

P²SfM [15]. This recent method solves for projective structure incrementally by solving linear least squares systems that incorporate constraints on the sought projective depths. The paper demonstrated both superior re-projection accuracy and running time, compared to existing methods.

VarPro [10]. This method first applies affine bundle adjustment followed by projective BA. It further uses variable projection to improve the solution of BA.

Sengupta et. al [22]. Similar to our method, this method applies rank constraints to n-view matrices, but it explicitly recovers the scale factors. As the authors acknowledge, their algorithm is sensitive, and so it was suggested as a refinement to a calibrated method [18]. To avoid calibration, for a fair comparison, we initialized the method with P²SfM [15]. Moreover, as the method does not suggest a way to produce projective reconstruction, we used Corollary 1 to obtain camera matrices from its output.

To compare all the methods, we ran them with the code supplied by the authors. For [22] we counted only its running time (excluding the running time of the initialization method). For [10] we set a time limit of 12 hours. We ran all the methods on the same computer under the same conditions.

Table 1 shows the mean reprojection error (in pixels) across all scene points and the total running time (in seconds) obtained with our method compared to P²SfM, VarPro, and Sengupta et al. It can be seen that our method achieved superior accuracies to all the other methods in 22 of the 25 data sets tested. Moreover, for certain complex scene structures (e.g., Pumpkin and Folk Filbyter) our method managed to reconstruct the scene, whereas all the other methods failed to obtain a reconstruction (we tried with many different hyper-parameters). In almost all cases our method was also faster, improving runtime in 22 out of 25 data sets. The best compared method (separately for each dataset) had a median of 4.8% reprojection error worse than our method and required an additional median runtime of 74% compared to our method. The quality of our reconstruction can also be appreciated by the recovered 3D point

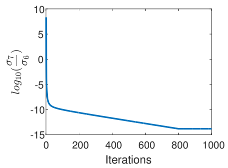


Figure 2: Enforcing rank constraints. The plot shows (in $\log_{10}$ scale) the ratio between the seventh and sixth singular value of $F_{\tau(k)}$ averaged for all triplets.

clouds shown in Fig 3.

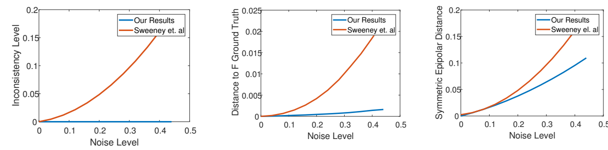
We believe the improved accuracy is partly explained by the effective enforcement of the rank 6 constraint for all triplets. We demonstrate this in Fig. 2, which shows the ratio between the 7^{th} and 6^{th} singular values of $F_{\tau(k)}$ averaged over all triplets for the House model. Our optimization reduces this ratio to near machine precision, indicating that indeed rank 6 is achieved in all runs.

4.4. Graph consistency simulations

Since our optimization (4) only enforces a subset of the consistency constraints in Theorem 1 we next test the consistency of our recovered fundamental matrices (with no bundle adjustment) in synthetic experiments. We generated 10 camera matrices and 15,000 three dimensional points. We then projected the points and perturbed them by Gaussian noise. Finally, we selected 15 matching triplets at random and used them to compute fundamental matrices, which we gave to our algorithm.

The results are shown in Fig. 4. We evaluated consistency using the symmetric epipolar distance associated with the term $\mathbf{e}_{ik}^T F_{ij} \mathbf{e}_{jk}$, where $\mathbf{e}_{ik}$ denotes the projected location of camera k onto camera i. Denote this distance by S_{ijk}, then $S_{ijk} + S_{jki} + S_{kij} = 0$ implies that cameras i, j, k are consistent ([8], p. 384). Additionally, we show the quality of recovering the ground truth fundamental matrices (measured by average Frobenius norm), and the average symmetric epipolar distance of the ground truth matches. It can be seen that our method maintains consistency under all error levels while achieving high quality recovery of

Figure 3: Visualization of the result of our projective structure from motion pipeline after applying self-calibration of [4] . From left to right: Tsar Nikolai I, Sphinx, Dome.



(a) (b) (c)

Figure 4: Synthetic experiments. Average symmetric epipolar distance of corresponding epipoles (left), error in fundamental matrix recovery, compared to ground truth (average Frobenius norm, middle), and symmetric epipolar distance for ground truth matches. Our method (in blue) is compared against [24].

fundamental matrices compared to ground truth. We compare our results with [24]'s consistency optimization (Sec. 3 therein).

5. Conclusion

We considered the problem of recovering projective camera matrices from collections of fundamental matrices. We derived a complete algebraic characterization of n-view fundamental matrices in the form of conditions that are both necessary and sufficient to enable the recovery of camera matrices. We further introduced an algorithm that uses this characterization for global recovery of camera matrices from measured fundamental matrices. Our algorithm is efficient and requires no initialization. We tested the algorithm on a large number of datasets and compared it to existing, state-of-the-art methods, showing both improved accuracy and runtime.

In future work we plan to explore ways to relate our algebraic constraints with the question of solvability of viewing graphs. In addition, we plan to similarly characterize collections of essential matrices.

Acknowledgment This research was supported in part by the Minerva foundation with funding from the Federal German Ministry for Education and Research.

References

[1] S. Agarwal, K. Mierle, and Others. Ceres solver. `http://ceres-solver.org`.

[2] M. Arie-Nachimson, S. Z. Kovalsky, I. Kemelmacher-Shlizerman, A. Singer, and R. Basri. Global motion estimation from point matches. In *2012 Second Int. Conf. on 3D Imaging, Modeling, Processing, Visualization & Transmission*, pages 81–88. IEEE, 2012.

[3] S. Boyd, N. Parikh, E. Chu, B. Peleato, J. Eckstein, et al. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends in Machine learning*, 3(1):1–122, 2011.

[4] M. Chandraker, S. Agarwal, F. Kahl, D. Nistér, and D. Kriegman. Autocalibration via rank-constrained estimation of the

absolute quadric. In *Computer Vision and Pattern Recognition, 2007. CVPR'07. IEEE Conference on*, pages 1–8. IEEE, 2007.

[5] S. Christy and R. Horaud. Euclidean shape and motion from multiple perspective views by affine iterations. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 18(11):1098–1104, 1996.

[6] Y. Dai, H. Li, and M. He. Projective multiview structure and motion from element-wise factorization. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 35(9):2238–2251, 2013.

[7] T. Goldstein, P. Hand, C. Lee, V. Voroninski, and S. Soatto. Shapefit and shapekick for robust, scalable structure from motion. In *European Conf. on Computer Vision*, pages 289–304. Springer, 2016.

[8] R. Hartley and A. Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[9] R. I. Hartley. In defence of the 8-point algorithm. In *Int. Conf. on Computer Vision*, pages 1064–1070. IEEE, 1995.

[10] J. H. Hong, C. Zach, A. Fitzgibbon, and R. Cipolla. Projective bundle adjustment from arbitrary initialization using the variable projection method. In *European Conf. on Computer Vision*, pages 477–493. Springer, 2016.

[11] N. Jiang, Z. Cui, and P. Tan. A global linear method for camera pose registration. In *IEEE Int. Conf. on Computer Vision*, pages 481–488, 2013.

[12] R. Kennedy, L. Balzano, S. J. Wright, and C. J. Taylor. Online algorithms for factorization-based structure from motion. *Computer Vision and Image Understanding*, 150:139–152, 2016.

[13] M. Klopschitz, A. Irschara, G. Reitmayr, and D. Schmalstieg. Robust incremental structure from motion. In *Proc. 3DPVT*, volume 2, pages 1–8, 2010.

[14] N. Levi and M. Werman. The viewing graph. In *Computer Vision and Pattern Recognition*, volume 1, pages I–I. IEEE, 2003.

[15] L. Magerand and A. Del Bue. Practical projective structure from motion (p2sfm). In *2017 IEEE Int. Conf. on Computer Vision (ICCV)*, pages 39–47. IEEE, 2017.

[16] J. Oliensis and R. Hartley. Iterative extensions of the sturm/triggs algorithm: Convergence and nonconvergence. *IEEE Trans. on Pattern Analysis and Machine Intelligence*, 29(12):2217–2233, 2007.

[17] C. Olsson and O. Enqvist. Stable structure from motion for unordered image collections. In *Scandinavian Conf. on Image Analysis*, pages 524–535. Springer, 2011.

[18] O. Ozyesil and A. Singer. Robust camera location estimation by convex programming. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 2674–2683, 2015.

[19] M. Pollefeys, L. Van Gool, M. Vergauwen, F. Verbiest, K. Cornelis, J. Tops, and R. Koch. Visual modeling with a hand-held camera. *Int. Journal of Computer Vision*, 59(3):207–232, 2004.

[20] A. Rudi, M. Pizzoli, and F. Pirri. Linear solvability in the viewing graph. In *Asian Conf. on Computer Vision*, pages 369–381. Springer, 2010.

[21] J. L. Schonberger and J.-M. Frahm. Structure-from-motion revisited. In *IEEE Conf. on Computer Vision and Pattern Recognition*, pages 4104–4113, 2016.

[22] S. Sengupta, T. Amir, M. Galun, T. Goldstein, D. W. Jacobs, A. Singer, and R. Basri. A new rank constraint on multi-view fundamental matrices, and its application to camera location recovery. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4798–4806, 2017.

[23] P. Sturm and B. Triggs. A factorization based algorithm for multi-image projective structure and motion. In *European Conf. on computer vision*, pages 709–720. Springer, 1996.

[24] C. Sweeney, T. Sattler, T. Hollerer, M. Turk, and M. Pollefeys. Optimizing the viewing graph for structure-from-motion. In *IEEE Int. Conf. on Computer Vision*, pages 801–809, 2015.

[25] T. Ueshiba and F. Tomita. A factorization method for projective and euclidean reconstruction from multiple perspective views via iterative depth estimation. In *European Conf. on computer vision*, pages 296–310. Springer, 1998.

[26] O. VGG. Multiview datasets. `http://www.robots.ox.ac.uk/~vgg/data/`.

[27] K. Wilson and N. Snavely. Robust global translations with 1dsfm. In *European Conf. on Computer Vision*, pages 61–75. Springer, 2014.