On Exploiting Occlusions in Multiple-view Geometry

Y. Ma — Dept. of Electrical and Computer Engineering, University of Illinois at Urbana-Champaign, Urbana, IL 61801

P. Favaro — Dept. of Electrical Engineering, Washington University, St. Louis, MO 63130

A. Duci — Dipartimento di Matematica, Scuola Normale, Pisa, Italy

S. Soatto — Dept. of Computer Science, UCLA, Los Angeles, CA 90095

Technical Report # UCLA CSD-TR030015

Keywords: geometry, multiple-view matrix, rank constraint

Abstract

Occlusions are commonplace in man-made and natural environments; they often result in photometric features where a line terminates at an occluding boundary, resembling a "T". We show that the 2-D motion of such T-junctions in multiple views carries non-trivial information on the 3-D structure of the scene and its motion relative to the camera. We show how the constraint among multiple views of T-junctions can be used to reliably detect them and differentiate them from ordinary point features. Finally, we propose an integrated algorithm to recursively and causally estimate structure and motion in the presence of T-junctions along with other point-features.

1 Introduction

The term "T-junction" commonly refers to a point on the image where a line segment terminates on another line, as illustrated in Fig. 1. The significance of T-junctions is that they often correspond to *occlusions*: the terminating segment lies on a surface

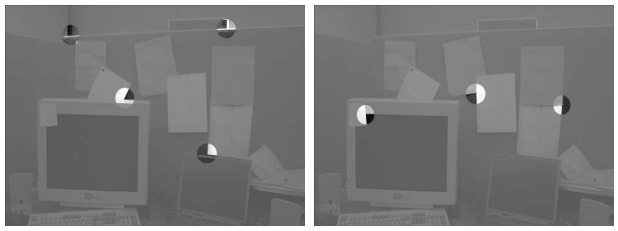

Figure 1: *Examples of proper (left) and false (right) T-junctions.*

in space that is occluded by another surface, whose occluding boundary contributes to form the "T" (Fig. 1 left). When this occurs we say that the T-junction is *proper*. When this does not occur, and both line segments lie on the same surface in space, we have a *false* T-junction, which is simply a radiance profile that happens to resemble a "T" (Fig. 1 right).

In order to distinguish a proper T-junction from a false one, one needs multiple views of the scene taken from different vantage points, as we describe in Sect. 1.2. Nevertheless, most photometric feature detectors, for instance the popular Harris' corner detector [5], happily detect both types because of the rich irradiance profile. While false T-junctions do not cause any problem if fed to a multiple-view geometry reconstruction algorithm (they are just ordinary point features), proper T-junctions do not correspond to any point physically attached to the (rigid) scene, and therefore they adversely affect the estimate of the multiple-view geometry.

Traditionally, T-junctions are detected based on their photometric signature and excised as outliers. This, however, presents two problems: first, detecting T-junctions based on local photometry is unreliable, because true T-junctions are defined based on their global geometric relationship among multiple views. Second, discarding T-junctions as outliers is tantamount to throwing away information. In this paper, we show that T-junctions carry non-trivial information on the structure of the scene and its motion relative to the camera, we propose algorithms to reliably detect T-junctions based on their multiple-view geometry, and we propose an integrated algorithm to recursively and causally estimate structure and motion in the presence of T-junctions along with other point-features.

1.1 Relation to prior work and contribution of this paper

Detection of T-junctions from one single image based on local photometric information has been the subject of numerous studies in the field of edge and contour detection. The interested reader can consult [12, 3, 13, 15] and references therein. Our work, instead, addresses the role of (proper) T-junction in multiple-view geometry: it is exactly the difficulty in reliably detecting T-junctions in single images that motivates us to extend the analysis to multiple images, where proper T-junctions can be detected and differentiated from other features, including false T-junctions.

Detection on T-junctions based on the local (instantaneous) motion was addressed in [2] based on a probabilistic model. We work in a deterministic, wide-baseline, multiple-view scenario that complements that of Black and Fleet [2]. Naturally, our work relates on the expansive literature on multiple-view geometry, which we cannot review in detail given the limited space. We refer the reader to the recent monographs [4, 6] and references therein for an accurate account of results and credits. Recently, [1, 7] generalized the geometry to points moving on lines. Matrix rank conditions have been an effective tool to study the geometry of multiple views of points and lines [16, 9, 14, 11]. This paper shows that the same framework can be used to study T-junctions.

In this paper we first show that T-junctions carry non-trivial information on the structure of the scene and its motion relative to the camera (Sect. 2). We then show how the geometric constraints can be used to reliably detect T-junctions based on their motion and geometry (Sect. 3). Once we have unraveled the geometry of T-junctions, in order to arrive at a robust, causal inference scheme to estimate structure and motion, we implement a robust version of the extended Kalman filter, as proposed by [10], in Section 4.2, and document its performance in Section 5.

1.2 Notation and definition of T-junction

Consider two lines in space, ℓ_1 and ℓ_2, each represented by a base point (a point on the line) $\boldsymbol{X}_1, \boldsymbol{X}_2 \in \mathbb{R}^3$, and a vector $\boldsymbol{V}_1, \boldsymbol{V}_2 \in T\mathbb{R}^3$, such that $\ell_i = \{\boldsymbol{X} = \boldsymbol{X}_i + \rho \boldsymbol{V}_i, \rho \in \mathbb{R}\},\ i = 1, 2$. In homogeneous coordinates, we write $\bar{\boldsymbol{X}}_i = [\boldsymbol{X}_i^T, 1]^T \in \mathbb{R}^4$ and $\bar{\boldsymbol{V}}_i = [\boldsymbol{V}_i^T, 0]^T \in \mathbb{R}^4,\ i = 1, 2$. Note that points have a "1" appended, whereas vectors have a "0". We often forgo the bar in the homogeneous coordinates when the dimension is clear from the context.

A "T-junction" is defined as a point on a plane that is the intersection of the projection of two lines onto it (see Figure 1). For perspective projection, a T-junction is represented by homogeneous coordinates $\boldsymbol{x} = [x, y, 1]^T \in \mathbb{R}^3$ that satisfy the following equation[1]

$$\boldsymbol{x} \sim \Pi[\bar{\boldsymbol{X}}_1 + \rho \bar{\boldsymbol{V}}_1] \sim \Pi[\bar{\boldsymbol{X}}_2 + \gamma \bar{\boldsymbol{V}}_2] \tag{1}$$

for some $\rho, \gamma \in \mathbb{R}$, where $\Pi = [R, T]$ with[2] $T \in \mathbb{R}^3$ and $R \in SO(3)$ for the case of calibrated geometry, otherwise $R \in GL(3)$. In order to avoid trivial statements, we require that the two lines be non-intersecting (otherwise their intersection provides a trivial incidence relation), and non-parallel (otherwise the T-junction degenerates to an entire line). That is, we require that

$$\ell_1 \cap \ell_2 = \emptyset, \quad \boldsymbol{V}_1 \times \boldsymbol{V}_2 \neq 0 \tag{2}$$

where $\times$ denotes cross product between vectors.

Consider now a collection of m images of a scene taken from different vantage points, and assume that one can establish the correspondence of a T-junction (represented by the two lines ℓ_1 and ℓ_2) among the different views. That is, one can measure $\boldsymbol{x}_1, \boldsymbol{x}_2, \ldots, \boldsymbol{x}_m \in \mathbb{R}^2$ that satisfy

$$\boldsymbol{x}_i \sim \Pi_i[\boldsymbol{X}_1 + \rho_i \boldsymbol{V}_1] \sim \Pi_i[\boldsymbol{X}_2 + \gamma_i \boldsymbol{V}_2],\ i = 1, \ldots, m \tag{3}$$

[1] We here use "$\sim$" to denote equality up to scale.

[2] $SO(3)$ is the space of matrices that are orthogonal with unit determinant $\{R | RR^T = I, det(R) = 1\}$.

where the structure of the scene, represented by $\boldsymbol{X}_1, \boldsymbol{X}_2, \boldsymbol{V}_1, \boldsymbol{V}_2$, and the motion of the camera $\Pi_i = [R_i, T_i]$ as well as the scales ρ_i, γ_i are all unknown. The first question that arises naturally is whether knowledge of the T-junction measurements $\boldsymbol{x}_i$ provide any useful information on the unknown structure of the scene and its motion relative to the camera. This question is addressed in the next section. In what follows $\widehat{\boldsymbol{u}} \in \mathbb{R}^{3\times 3}$ denotes the skew-symmetric matrix defined by $\widehat{\boldsymbol{u}}\boldsymbol{v} = \boldsymbol{u} \times \boldsymbol{v},\ \forall\, \boldsymbol{u}, \boldsymbol{v} \in \mathbb{R}^3$.

2 Multiple-view geometry of T-junctions

In this section we show that the multiple-view constraints for a T-junction can be characterized by the rank of a multiple-view matrix which has a special null space. The rank of the matrix and its null space precisely distinguish a proper T-junction from regular point features (false T-junction), either fixed or moving freely on a line.

2.1 Multiple-view rank conditions for T-junctions

The following claim unravels how the information on scene structure and camera motion is related through multiple images of T-junctions.

Lemma 1 *Let $\boldsymbol{x}_1, \ldots, \boldsymbol{x}_m$ denote the position of a T-junction in m distinct views, related by changes of coordinates $[R_1, T_1]$, $\ldots$, $[R_m, T_m]$. Let ℓ_1, ℓ_2 be two lines in space represented by $\{\boldsymbol{X}_1, \boldsymbol{V}_1\}$ and $\{\boldsymbol{X}_2, \boldsymbol{V}_2\}$ respectively, that satisfy the conditions (2). Define the following matrix*

$$W \doteq \begin{bmatrix} \boldsymbol{x}_1^T R_1 & \boldsymbol{x}_1^T \widehat{T}_1 R_1 \\ \boldsymbol{x}_2^T R_2 & \boldsymbol{x}_2^T \widehat{T}_2 R_2 \\ \vdots & \vdots \\ \boldsymbol{x}_m^T R_m & \boldsymbol{x}_m^T \widehat{T}_m R_m \end{bmatrix} \in \mathbb{R}^{m\times 6}. \tag{4}$$

Then we have that rank$(W) \leq 4$.

Proof: ¿From the definition of W, we have

$$\begin{aligned} \boldsymbol{x} &\sim R\boldsymbol{X}_1 + \rho R\boldsymbol{V}_1 + T, \\ \boldsymbol{x} &\sim R\boldsymbol{X}_2 + \gamma R\boldsymbol{V}_2 + T. \end{aligned}$$

Applying the hat operator on both sides, we get

$$\begin{aligned} \widehat{\boldsymbol{x}} &\sim R\widehat{\boldsymbol{X}_1}R^T + \rho R\widehat{\boldsymbol{V}_1}R^T + \widehat{T}, \\ \widehat{\boldsymbol{x}} &\sim R\widehat{\boldsymbol{X}_2}R^T + \gamma R\widehat{\boldsymbol{V}_2}R^T + \widehat{T}. \end{aligned}$$

Multiply on both sides $\boldsymbol{x}^T$ to the left and R to the right,[3] it yields

$$\begin{aligned} 0 &= \boldsymbol{x}^T R\widehat{\boldsymbol{X}_1} + \rho \boldsymbol{x}^T R\widehat{\boldsymbol{V}_1} + \boldsymbol{x}^T \widehat{T} R, \\ 0 &= \boldsymbol{x}^T R\widehat{\boldsymbol{X}_2} + \gamma \boldsymbol{x}^T R\widehat{\boldsymbol{V}_2} + \boldsymbol{x}^T \widehat{T} R. \end{aligned}$$

Now multiplying $\boldsymbol{V}_1$ to the first equation and $\boldsymbol{V}_2$ to the second we get

$$\begin{aligned} 0 &= \boldsymbol{x}^T R\widehat{\boldsymbol{X}_1}\boldsymbol{V}_1 + \boldsymbol{x}^T \widehat{T} R\boldsymbol{V}_1, \\ 0 &= \boldsymbol{x}^T R\widehat{\boldsymbol{X}_2}\boldsymbol{V}_2 + \boldsymbol{x}^T \widehat{T} R\boldsymbol{V}_2. \end{aligned}$$

Obviously, since the following two vectors

$$\boldsymbol{U}_1 = \begin{bmatrix} \widehat{\boldsymbol{X}_1}\boldsymbol{V}_1 \\ \boldsymbol{V}_1 \end{bmatrix}, \quad \boldsymbol{U}_2 = \begin{bmatrix} \widehat{\boldsymbol{X}_2}\boldsymbol{V}_2 \\ \boldsymbol{V}_2 \end{bmatrix}$$

[3] In this step sufficiency is lost.

are always in the null space of $[\boldsymbol{x}^T R,\ \boldsymbol{x}^T \widehat{T} R]$, they are in the null space of W. Therefore, we have

$$\text{rank}(W) \leq 4.$$

■

If we choose the first camera frame to be the world frame, $[R_1, T_1] = [I, 0]$, then the above matrix W simplifies to

$$W = \begin{bmatrix} \boldsymbol{x}_1^T & 0 \\ \boldsymbol{x}_2^T R_2 & \boldsymbol{x}_2^T \widehat{T}_2 R_2 \\ & \vdots \\ \boldsymbol{x}_m^T R_m & \boldsymbol{x}_m^T \widehat{T}_m R_m \end{bmatrix}. \tag{5}$$

Multiplying on the right by the full rank matrix $\mathbb{R}^{6\times 7}$

$$\begin{bmatrix} \boldsymbol{x}_1 & \widehat{\boldsymbol{x}}_1 & 0 \\ 0 & 0 & I \end{bmatrix}. \tag{6}$$

we obtain the matrix $W' \in \mathbb{R}^{m\times 7}$

$$W' = \begin{bmatrix} \boldsymbol{x}_1^T \boldsymbol{x}_1 & 0 & 0 \\ \boldsymbol{x}_2^T R_2 \boldsymbol{x}_1 & \boldsymbol{x}_2^T R_2 \widehat{\boldsymbol{x}}_1 & \boldsymbol{x}_2^T \widehat{T}_2 R_2 \\ & \vdots & \\ \boldsymbol{x}_m^T R_m \boldsymbol{x}_1 & \boldsymbol{x}_m^T R_m \widehat{\boldsymbol{x}}_1 & \boldsymbol{x}_m^T \widehat{T}_m R_m \end{bmatrix}. \tag{7}$$

Note that the matrix W' has the same rank as W since we multiplied it by a full rank matrix on the right. Therefore,

$$\text{rank}(W') \leq 4. \tag{8}$$

Let us define the multiple-view matrix (as a sub-matrix of W')

$$M_T \doteq \begin{bmatrix} \boldsymbol{x}_2^T R_2 \widehat{\boldsymbol{x}}_1 & \boldsymbol{x}_2^T \widehat{T}_2 R_2 \\ \boldsymbol{x}_3^T R_3 \widehat{\boldsymbol{x}}_1 & \boldsymbol{x}_3^T \widehat{T}_3 R_3 \\ \vdots & \vdots \\ \boldsymbol{x}_m^T R_m \widehat{\boldsymbol{x}}_1 & \boldsymbol{x}_m^T \widehat{T}_m R_m \end{bmatrix} \in \mathbb{R}^{(m-1)\times 6}, \tag{9}$$

where the subscript "T" indicates T-junction.

Theorem 2 *Under the same conditions of Lemma 1, the m images of a T-junction satisfy the following rank condition*

$$\textit{rank}(M_T) \leq 3. \tag{10}$$

It is not difficult to see that

$$\boldsymbol{U}_0 = \begin{bmatrix} \boldsymbol{x}_1 \\ 0 \end{bmatrix}, \quad \boldsymbol{U}_1 = \begin{bmatrix} \lambda_1 \boldsymbol{V}_1 \\ \boldsymbol{V}_1 \end{bmatrix}, \quad \boldsymbol{U}_2 = \begin{bmatrix} \lambda_2 \boldsymbol{V}_2 \\ \boldsymbol{V}_2 \end{bmatrix} \in \mathbb{R}^6 \tag{11}$$

for some $\lambda_1, \lambda_2 \in \mathbb{R}$ are three linearly independent vectors in the three-dimensional null space of M_T.

2.2 Relations to other rank conditions

A T-junction can be viewed as the intermediate case between a fixed point as the intersection of two lines and a point which can move freely on one straight line, as shown in Figure 2. It is then reasonable to expect that the rank condition for the T-junction is related to those of the other two cases. This is indeed true. Our derivation in the previous section directly applies to the case when the point $\boldsymbol{X}$ moves on only one line. In this case, we have

$$\text{rank}(M_T) \leq 4, \text{ with } \left\{ \begin{bmatrix} \lambda_1 \boldsymbol{V}_1 \\ \boldsymbol{V}_1 \end{bmatrix} \right\} \in \text{null}(M_T). \tag{12}$$

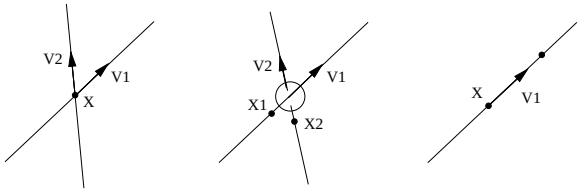

Figure 2: Left: A point $\boldsymbol{X}$ as the intersection of two lines. Middle: A T-junction from two lines. Right: A point that moves on one line.

In the T-junction case, we have

$$\text{rank}(M_T) \leq 3, \text{ with } \left\{ \begin{bmatrix} \lambda_1 \boldsymbol{V}_1 \\ \boldsymbol{V}_1 \end{bmatrix}, \begin{bmatrix} \lambda_2 \boldsymbol{V}_2 \\ \boldsymbol{V}_2 \end{bmatrix} \right\} \in \text{null}(M_T). \tag{13}$$

In the case when two lines intersect at a fixed point, we may choose the base point for each line to be the intersection $\boldsymbol{X}$ and then we have $\rho = \gamma = 0$. This implies that the derivation in the previous section holds for any vector $\boldsymbol{V}_1$ or $\boldsymbol{V}_2 \in \mathbb{R}^3$. Therefore, in this case, we have

$$\text{rank}(M_T) \leq 2, \text{ with } \left\{ \begin{bmatrix} \lambda \boldsymbol{V} \\ \boldsymbol{V} \end{bmatrix}, \forall \boldsymbol{V} \in \mathbb{R}^3 \right\} \in \text{null}(M_T). \tag{14}$$

This condition yields

$$[\lambda \boldsymbol{x}_i^T R_i \widehat{\boldsymbol{x}}_1 + \boldsymbol{x}_i^T \widehat{T}_i R_i] \boldsymbol{V} = 0, \quad \forall \boldsymbol{V} \in \mathbb{R}^3 \tag{15}$$

for $i = 1, 2, \ldots, m$. Note that $R_i \widehat{\boldsymbol{x}}_1 = \widehat{R_i \boldsymbol{x}_1} R_i$, we obtain the equation

$$\lambda \boldsymbol{x}_i^T \widehat{R_i \boldsymbol{x}_1} + \boldsymbol{x}_i^T \widehat{T}_i = 0, \quad \forall i. \tag{16}$$

This is equivalent to

$$\lambda \widehat{\boldsymbol{x}}_i R_i \boldsymbol{x}_1 + \widehat{\boldsymbol{x}}_i T_i = 0, \quad \forall i. \tag{17}$$

Therefore, the above condition (14) is equivalent to the known rank condition for a fixed point

$$\text{rank} \begin{bmatrix} \widehat{\boldsymbol{x}_2} R_2 \boldsymbol{x}_1 & \widehat{\boldsymbol{x}_2} T_i \\ \widehat{\boldsymbol{x}_3} R_3 \boldsymbol{x}_1 & \widehat{\boldsymbol{x}_3} T_i \\ \vdots & \vdots \\ \widehat{\boldsymbol{x}_m} R_m \boldsymbol{x}_1 & \widehat{\boldsymbol{x}_m} T_i \end{bmatrix} \leq 1. \tag{18}$$

Notice that all the above constraints are non trivial. If we take measurements of a point moving randomly in space, the corresponding matrix M_T has in general[4] rank ≤ 5 (it cannot possibly have rank 6 because the vector $\boldsymbol{U}_0$ is always in the kernel of M_T).

2.3 Constraints from the rank conditions for T-junctions

What kind of constraints does the above rank condition for a T-junction give rise to? The rank condition (10) requires that any 4×4 minor of the matrix M_T has determinant 0. It is easy to notice that any such a minor involves up to 5 different images.

However, one should notice further that the nature of the rank condition for a T-junction is *different* from the conventional rank conditions studied in [11]. The rank conditions themselves are only *necessary* but not *sufficient* for the m images of a T-junction. The reason is because the null space of the matrix M_T here needs to have the special structure described in (11). From the equation

$$M_T \boldsymbol{U}_1 = 0, \tag{19}$$

[4]This is true up to a set (of measure zero) of degenerate camera motions.

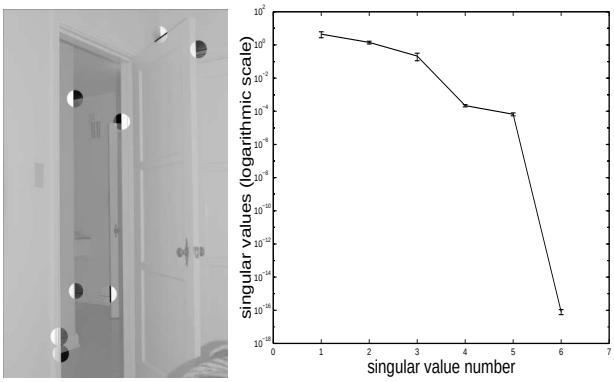

Figure 3: Verification of the rank constraint. The plot on the right shows the mean and standard deviation of the 6 singular values. Notice that there is a drop after the third singular value as expected from the multiple-view matrix rank constraint. The last singular value is always zero by construction.

we obtain

$$\frac{\boldsymbol{x}_2^T R_2 \widehat{\boldsymbol{x}}_1 \boldsymbol{V}_1}{\boldsymbol{x}_2^T \widehat{T}_2 R_2 \boldsymbol{V}_1} = \frac{\boldsymbol{x}_3^T R_3 \widehat{\boldsymbol{x}}_1 \boldsymbol{V}_1}{\boldsymbol{x}_3^T \widehat{T}_3 R_3 \boldsymbol{V}_1} = \cdots = \frac{\boldsymbol{x}_m^T R_m \widehat{\boldsymbol{x}}_1 \boldsymbol{V}_1}{\boldsymbol{x}_m^T \widehat{T}_m R_m \boldsymbol{V}_1} \tag{20}$$

since the ratio is exactly λ_1. A similar set of equations can be obtained from $M_T \boldsymbol{U}_2 = 0$. The above equations can also be written in another way

$$\boldsymbol{V}_1^T (\widehat{\boldsymbol{x}}_1 R_i^T \boldsymbol{x}_i \boldsymbol{x}_j^T \widehat{T}_j R_j - R_i^T \widehat{T}_i \boldsymbol{x}_i \boldsymbol{x}_j^T R_j \widehat{\boldsymbol{x}}_1) \boldsymbol{V}_1 = 0 \tag{21}$$

for all $2 \leq i < j \leq m$. Define[5]

$$S_{ij} \doteq \widehat{\boldsymbol{x}}_1 R_i^T \boldsymbol{x}_i \boldsymbol{x}_j^T \widehat{T}_j R_j - R_i^T \widehat{T}_i \boldsymbol{x}_i \boldsymbol{x}_j^T R_j \widehat{\boldsymbol{x}}_1, \quad \in \mathbb{R}^{3\times 3}$$

the above equation is simplified to $\boldsymbol{V}_1^T S_{ij} \boldsymbol{V}_1 = 0$. Similarly, we have $\boldsymbol{V}_2^T S_{ij} \boldsymbol{V}_2 = 0$.

In order to eliminate the two unknowns[6] in $\boldsymbol{V}_1$ and arrive at expressions that do not depend on $\boldsymbol{V}_1$, we need three independent equations of the form $\boldsymbol{V}_1^T S_{ij} \boldsymbol{V}_1 = 0$ which typically involve at least 5 images. This conforms to the result that if a point is moving on a straight-line, 5 images are needed in order to obtain effective constraints [1].

However, here a T-junction lies simultaneously on two straight-lines and the equations

$$\boldsymbol{V}_1^T S_{ij} \boldsymbol{V}_1 = 0, \quad \boldsymbol{V}_2^T S_{ij} \boldsymbol{V}_2 = 0 \tag{22}$$

usually are not totally unrelated. For instance, if we assume[7] $\boldsymbol{V}_1^T \boldsymbol{V}_2 = 0$, there are only a total of three unknowns in both $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$. Then, we can use the following four equations

$$\begin{aligned} \boldsymbol{V}_1^T S_{23} \boldsymbol{V}_1 = 0, \quad \boldsymbol{V}_1^T S_{34} \boldsymbol{V}_1 = 0, \\ \boldsymbol{V}_2^T S_{23} \boldsymbol{V}_2 = 0, \quad \boldsymbol{V}_2^T S_{34} \boldsymbol{V}_2 = 0 \end{aligned}$$

to eliminate both $\boldsymbol{V}_1$ and $\boldsymbol{V}_2$. Obviously, the resulting constraint will only involve collections of 4 images.

2.4 Testing the rank constraints

In order to validate the analysis experimentally, we have generated 6 views of T-junctions seen from a moving vantage point (Fig. 2.4). The plot on the right shows the numerical value of the rank for each of the T-junctions, displayed as a mean and standard deviation. As one can see, the numerical rank - although not strictly equal to 3 due to noise in feature localization - drops significantly beyond 3, and can therefore be easily determined by thresholding techniques. This experiment is shown

[5] Enforcing $S_{ij} = 0$ leads to the well-known trilinear constraint.

[6] Recall that the vector $\boldsymbol{V}_1$ is defined up to scale.

[7] In the case that a T-junction is caused by a pair of mutually orthogonal lines.

only for illustrative purposes. We do not propose using the rank condition "cold turkey" to estimate structure and motion, especially because real scenes are rarely comprised entirely of T-junctions. Instead, we wish to integrate the rank condition in a general multiple-view inference scheme in order to exploit information coming from the T-junctions rather than discarding them as outliers. We discuss this in the next sections.

3 T-junction detection

The results in section (2) can be used in a straightforward manner to classify point features into rigid points, T-junctions and outliers. By using the filtering scheme described below in the next sections, we assume that a sufficiently large subset of the detected point features are ordinary features or false T-junctions. Our procedure consists in collecting measurements of features for 5 or more frames with the corresponding estimated camera motions, and then building the multiple-view matrix associated to each of the features.

We classify a feature as outlier if the rank of the multiple-view matrix is 5 or 4, as a T-junction if the rank is 3, and as a rigid point if the rank is 2. Due to noise or degenerate motions, the classification may be prone to errors. However, this is not an issue in the proposed robust filtering scheme (see section 4.2), since measurements that are not modeled properly are detected as outliers and automatically weighted accordingly.

Once T-junctions are detected, their structure can be reconstructed from the constraints in eq. (21) by minimizing the following cost functionals:

$$\begin{aligned} \tilde{V}_1 &= \arg\min_{V_1} \sum_{i=2..m} \left(\frac{V_1^T S_{1i} V_1}{\|V_1\|^2} \right)^2 \\ \tilde{V}_2 &= \arg\min_{V_2} \sum_{i=2..m} \left(\frac{V_2^T S_{1i} V_2}{\|V_2\|^2} \right)^2 + \alpha \left(\frac{\tilde{V}_1^T V_2}{\|V_2\|} \right)^2 \end{aligned} \tag{23}$$

where $\alpha \in \mathbb{R}$ is a tuning parameter. Notice that the second minimization also forces the direction V_2 to be transversal to V_1. We perform the minimization by using a simple gradient descent technique. The estimated structure can be used to initialize the structure and motion filter that we are going to present in the next section.

4 Structure and motion estimation from mixtures of rigid features and T-junctions

We consider the (general) class of structure from motion estimation algorithms that are divided into the following two steps: (i) Select and track features points on the image plane, and (ii) use the 2D trajectories of the tracked points to infer both their structure and the camera motion. The advantage of splitting the task into these two steps resides in considerably reducing the number of required computations and in simplifying the design of each algorithm. However, splitting the task into two steps has also some drawbacks. In order to estimate camera motion, the second step needs to have an a-priori model for structure. Therefore, when we feed it with measurements that are not generated by the same model, what we call *outliers*, the estimation process can produce erratic results. Since a feature tracker is based solely on matching the photometry of the features, and not their 3D geometry, it has no way to detect these outliers.

As a solution to this problem, we propose to use a filtering scheme that accounts for multiple structure models, which will be introduced in section 4.1. To be able to select the most appropriate model during the estimation process, we work in the framework of *robust* Kalman filtering, which will be presented in section 4.2.

4.1 Structure and motion representation

Camera motion is represented by a time-varying translation vector $T(t) \in \mathbb{R}^3$ and rotation matrix $R(t) \in SO(3)$. Camera motions transform the coordinates of a point $\boldsymbol{X}$ in space via $R(t)\boldsymbol{X} + T(t)$. Associated to each motion there is a velocity, represented by a vector of linear velocity $V(t)$ and a skew-symmetric matrix $\widehat{\omega}(t)$ of rotational velocity. Under such a velocity, motion evolves according to $T(t+1) = e^{\widehat{\omega}(t)}T(t) + V(t)$; $R(t+1) = e^{\widehat{\omega}(t)}R(t)$. The exponential of a skew-symmetric matrix can be computed using Rodrigues formula: $e^{\widehat{\omega}} = I + \frac{\widehat{\omega}}{\|\widehat{\omega}\|}\sin(\|\widehat{\omega}\|) + \frac{\widehat{\omega}^2}{\|\widehat{\omega}\|^2}(1 - \cos(\|\widehat{\omega}\|))\ \forall \|\widehat{\omega}\| \neq 0$, otherwise $e^{\widehat{0}} = I$.

As mentioned in previous sections, the measurements on the image plane $\{x_i\}_{i=1..m}$ may have been generated by points on a rigid structure, by T-junctions, or, more in general, may be moving entirely independently of the scene.

We model a rigid point $\boldsymbol{X} \in \mathbb{R}^3$ as the product of a $2D$ homogeneous coordinates $\boldsymbol{x} \in \mathbb{R}^3$ with a positive scalar $\lambda \in \mathbb{R}_+$, i.e.

$$\boldsymbol{X} = \boldsymbol{x}\lambda. \tag{24}$$

This choice has the advantage that $\boldsymbol{x}$ can be measured directly on the image plane, and it leaves one with estimating only the scalar λ.

As the camera moves in time, the measurement equation corresponding to a rigid point becomes:

$$\boldsymbol{x}(t)\lambda(t) = \Pi(t)\boldsymbol{x}\lambda \tag{25}$$

where $\boldsymbol{x} = \boldsymbol{x}(0)$ and $\lambda = \lambda(0)$.

T-junctions are instead modeled by using the normalized directions $\boldsymbol{V}_1 \in \mathbb{R}^3$ and $\boldsymbol{V}_2 \in \mathbb{R}^3$, and two points $\boldsymbol{X}_1 \in \mathbb{R}^3$ and $\boldsymbol{X}_2 \in \mathbb{R}^3$ on the lines ℓ_1 and ℓ_2 respectively. To keep the representation minimal, instead of using any two points on the lines, we take the ones that can be factorized as the product of a measurement in homogeneous coordinates $\boldsymbol{x} \in \mathbb{R}^3$ (the measurement at time 0) and two scalars β_1 and β_2, i.e.

$$\begin{array}{rcl} \boldsymbol{X}_1 & = & \boldsymbol{x}\beta_1 \\ \boldsymbol{X}_2 & = & \boldsymbol{x}\beta_2. \end{array} \tag{26}$$

In this case we use the following measurement equation

$$\begin{array}{rcl} \boldsymbol{x}(t)\lambda(t) & = & ((R(t)\boldsymbol{V}_1) \times (R(t)\boldsymbol{x}\beta_1 + T(t)) \times \\ & & \times (R(t)\boldsymbol{V}_2) \times (R(t)\boldsymbol{x}\beta_2 + T(t))) \,. \end{array} \tag{27}$$

4.2 A robust Kalman filter implementation

We make the assumption that (both linear and rotational) accelerations are a Brownian motion in time. This assumption and the structure models in the previous section result in the following state and measurements equations:

$$\left\{ \begin{array}{rcll} \lambda_i(t+1) & = & \lambda_i(t) & \forall i = 1...N_R \\ \beta_{1,i}(t+1) & = & \beta_{1,i}(t) & \forall i = 1...N_T \\ \beta_{2,i}(t+1) & = & \beta_{2,i}(t) & \forall i = 1...N_T \\ \boldsymbol{V}_{1,i}(t+1) & = & \boldsymbol{V}_{1,i}(t) & \forall i = 1...N_T \\ \boldsymbol{V}_{2,i}(t+1) & = & \boldsymbol{V}_{2,i}(t) & \forall i = 1...N_T \\ T(t+1) & = & e^{\widehat{\omega}(t)}T(t) + V(t) & \\ R(t+1) & = & e^{\widehat{\omega}(t)}R(t) & \\ V(t+1) & = & V(t) + n_V & \\ \omega(t+1) & = & \omega(t) + n_\omega & \\ \boldsymbol{x}_i^{rgd}(t) & = & \pi\left(R(t)\boldsymbol{x}_i^{rgd}\lambda_i + T(i)\right) & \\ & & \forall i = 1...N_R & \\ \boldsymbol{x}_i^{jct}(t) & = & \pi\left((R(t)\boldsymbol{V}_{1,i}) \times \left(R(t)\boldsymbol{x}_i^{jct}\beta_{1,i} + T(t)\right) \times \right. & \\ & & \left. \times (R(t)\boldsymbol{V}_{2,i}) \times \left(R(t)\boldsymbol{x}_i^{jct}\beta_{2,i} + T(t)\right)\right) & \\ & & \forall i = 1...N_T & \end{array} \right. \tag{28}$$

where π is the perspective projection defined as $\pi([X_1\ X_2\ X_3]^T) = [\frac{X_1}{X_3}\ \frac{X_2}{X_3}]^T$. N_R is the number of rigid features and N_T the number of T-junctions, so that $m = N_R + N_T$. $(\cdot)^{rgd}$ denotes measurements of rigid point features, while $(\cdot)^{jct}$ denotes measurements of T-junctions.

To infer the state parameters, we implement a *robust* EKF (extended Kalman filter) (see [10] for more details). The main difference between the robust EKF and a traditional EKF is that the distribution of the state conditioned over the measurements, usually modeled by a normal distribution, is considered contaminated by outliers. Thus, instead of obtaining the MAP (maximum a posteriori) estimate, we seek for the M-estimate of the state. Following [8] (1981, p.71), we choose the "least informative" probability density and obtain that at each estimation step the measurement covariance $R_n = diag([r_1 \dots r_{2m}])$ changes as:

$$\left\{ \begin{array}{rcll} r_i & = & n & \text{if } \frac{|e_i|}{\sqrt{n}} \leq c \\ r_i & = & \frac{\sqrt{n}|e_i|}{c} & \text{if } \frac{|e_i|}{\sqrt{n}} > c \end{array} \right. \quad \forall i = 1..2m \tag{29}$$

where e_i is the innovation (i.e. the difference between the actual measurement and the prediction of the measurement) of the $i-th$ measurement, m is the number of feature measurements, n is the measurement noise variance (identical for all points), and c is a constant threshold usually set to 1.5. In other words, the robust EKF detects an outlier by testing the innovation. If the innovation is above a certain threshold (which is tied to the maximum degree of contamination of the normal distribution), the corresponding measurement is weighted so that it does not affect the state estimate.

The general structure and motion estimation scheme then proceed as follows:

- Initialize the filter with $N_T = 0$ (i.e. no T-junctions are modeled in the filter, but only rigid features)
- During motion estimation the robust scheme automatically detects outliers and re-weights the measurement covariance R_n accordingly, by using eq. (29)
- T-junctions are detected among the outliers as explained in section 3, and the corresponding state and measurement equations are inserted in the filter.

5 Experiments

The purpose of this set of experiments is to show that outliers need to be accounted for, as they have catastrophic effects on the estimation process, and that T-junctions are carriers of information, and rather than being discarded, they should be exploited. To this aim, we implemented three filters: one is the traditional EKF, where outliers are not accounted for; the second is the robust EKF, where outliers are detected and re-weighted accordingly (i.e. "discarded"), but where T-junctions are not explicitly modeled (i.e. N_T is always 0). The third is the T-junction EKF, which is as the robust EKF, but where T-junctions are instead used in the estimation process.

The synthetic scene is composed of 30 points of which 20 are rigid features, and 10 are T-junctions. The camera rotates around the points, with center of rotation approximately on the center of mass of the structure (T-junctions do not define a center of mass). We rescale both translation and structure by fixing the depth coordinate of one of the rigid points to 1. In Figure 4 we show one instance of the experiments performed for each of the implementations. The true motion (translation and rotation) is superimposed to the estimated motions of each of the filters. In particular, the traditional EKF diverges almost immediately. Notice that the camera motion estimated by both the robust EKF and the T-junction EKF are very close to the ground truth motion. To better appreciate the difference between these two implementations, in Figure 5 we show in more detail the estimation error on the translation for the robust EKF (dotted) and the T-junction EKF (solid). Similarly, in Figure 5 we show in more detail the estimation error on the rotation. We plot the norm of the error between the estimate and the ground truth for each time instant. The RMS error over the whole motion for the robust EKF is of $1.5 \cdot 10^{-3}$ for translation and of $1.5 \cdot 10^{-3}$ for rotation, and for the T-junctions EKF is of $6.2 \cdot 10^{-4}$ and $7.8 \cdot 10^{-4}$ respectively. This shows that using T-junctions doubles the performance of the filter. Thus, we can conclude that measurements of T-junctions are beneficial to the camera motion estimation once they are properly modeled, and therefore should be used rather than being discarded.

Unlike synthetic sequences, in real sequences we do not have accurate ground truth available. In order to generate a controlled experiment, we collect a sequence, and then play it forward and backward in order to guarantee that through the composite sequence the camera returns exactly at the initial position. We then evaluate the repositioning error (error in position and orientation of the camera relative to $T = 0$ and $R = I$). We do so for 10 experiments with forward translation, sideways translation, and roto-translation about a fixed point in space. The results of these preliminary experiments are summarized in Table 1 and show that our algorithm is very promising for applications in uncontrolled environments.

6 Conclusions

T-junctions are commonplace in man-made and natural environments, and cannot be distinguished from rigid features only from their photometric information. On the one hand, not accounting for T-junctions, may result in catastrophic consequences for camera motion estimation. As we showed in this paper, T-junctions should not be discarded as outliers, as they carry non-trivial information on the 3-D structure of the scene and its motion relative to the camera. We analyzed T-junctions in the context of the multiple-view geometry, defined the multiple-view matrix for T-junctions, and derived the corresponding rank constraint. We showed how the constraint among multiple views of T-junctions can be used to reliably detect them and differentiate them from ordinary point features. Finally, we proposed a scheme in the framework of robust Kalman filtering to recursively and causally estimate structure and motion in the presence of T-junctions along with other point-features.

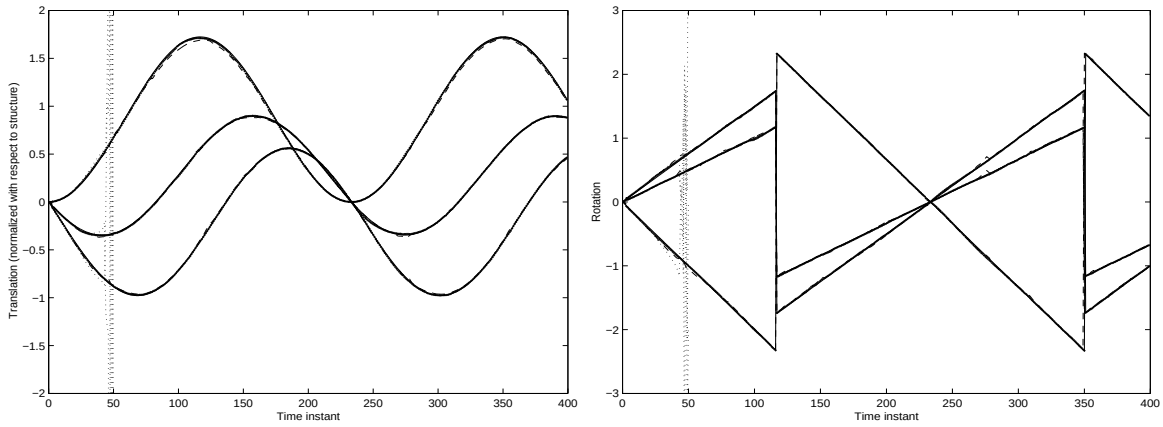

Figure 4: Ground truth camera motion components superimposed to the estimates of the three implementations. The evolution of the three components of translation (normalized with respect to the structure) in time is shown on the top figure, while the evolution of the three components of the rotation are shown on the bottom figure. The traditional EKF (dashed line) diverges after 50 frames. The robust EKF (dotted) and the T-junction EKF (dotted-dashed) are both very close to the ground truth as expected. To better appreciate the difference, we show their corresponding estimation error in Figure 5 and Figure 6.

	Forward Translation	Sideways Translation	Roto-Translation
Translation error	0.0082	0.0040	0.0039
Rotation error	0.0042	0.0060	0.0045

Table 1: In the table we show the results for three experiments: forward motion, sideways motion and roto-translation motion. The translation is normalized with respect to the structure (we fixed the depth of a rigid point to 1). The values in the table are the norm of the repositioning error. The data set contains a mixture of 25 rigid points with 5 T-junctions, which leads a traditional EKF to divergence.

References

[1] S. Avidan and A. Shashua. Trajectory triangulation of lines: Reconstruction of a 3d point moving along a line from a monocular image sequence. In *Proceedings of CVPR*, 1999.

[2] Michael J. Black and David J. Fleet. Probabilistic detection and tracking of motion discontinuities. In *ICCV (1)*, pages 551–558, 1999.

[3] Thierry Blaszka and Rachid Deriche. Recovering and characterizing image features using an efficient model based approach. Technical Report RR-2422.

[4] O. Faugeras, Q.-T. Luong, and T. Papadopoulo. *Geometry of Multiple Images*. The MIT Press, 2001.

[5] C. Harris and M. Stephen. A combined corner and edge detection. *In M.M. Matthews, editor, Proceedings of the 4th ALVEY vision conference*, pages 147–51, 1988.

[6] R. Hartley and A. Zisserman. *Multiple View Geometry in Computer Vision*. Cambridge, 2000.

[7] K. Huang, Y. Ma, and R. Fossum. Generalized rank conditions in multiple view geometry and its applications to dynamical scenes. In *Proceedings of ECCV*, pages 201–16, 2002.

[8] P.J. Huber. *Robust Statistics*. Wiley, New York, 1981.

[9] M. Irani. Multi-frame optical flow estimation using subspace constraints. In *Proceedings of ICCV*, volume 1, pages 626–33, 1999.

[10] K.R. Koch and Y. Yang. Robust kalman filter for rank deficient observation models. *Journal of Geodesy*, 72(7-8):436–41, 1998.

[11] Y. Ma, K. Huang, and J. Kosecka. Rank deficiency condition of the multiple view matrix for mixed point and line features. In *Proceedings of ACCV*, 2002.

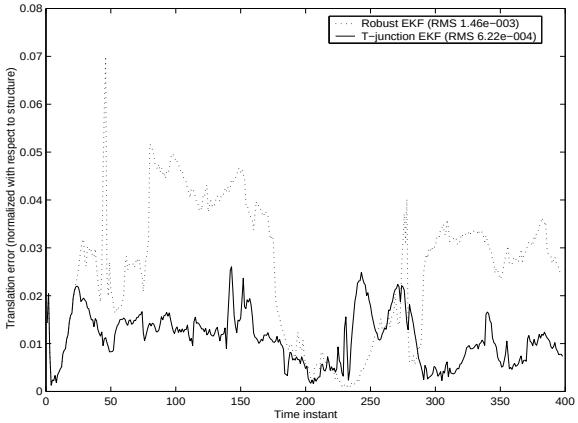

Figure 5: Estimation error of camera translation for a robust EKF (dotted) and the T-junction EKF (solid). We plot the norm of the RMS error for each time instant and compute the RMS error for the whole sequence. In both cases we can see that the T-junctions EKF improves the estimation of translation when compared to the robust EKF.

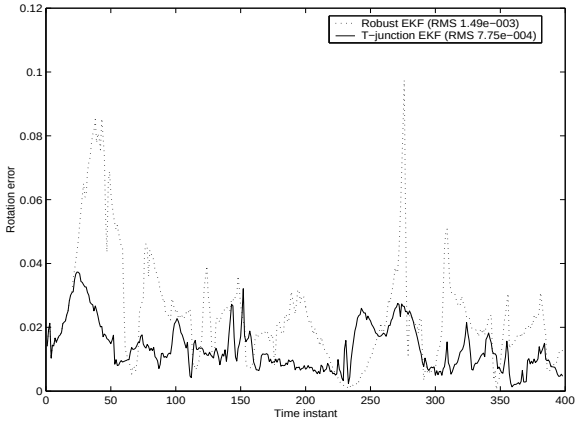

Figure 6: Estimation error of camera rotation for a robust EKF (dotted) and the T-junction EKF (solid). We plot the norm of the RMS error for each time instant and compute the RMS error for the whole sequence. In both cases we can see that the T-junctions EKF improves the estimation of rotation when compared to the robust EKF.

[12] L. Parida, D. Geiger, and R. Hummel. Junctions: Detection, classification and reconstruction. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 20(7):687–98, 1998.

[13] Pietro Perona. Steerable-scalable kernels for edge detection and junction analysis. In *European Conference on Computer Vision*, pages 3–18, 1992.

[14] L. Quan and T. Kanade. A factorization method for affine structure from line correspondences. In *Proceedings of the CVPR*, pages 803–808, 1996.

[15] E Simoncelli and H Farid. Steerable wedge filters for local orientation analysis. *IEEE Trans Image Proc*, 5(9):1377–1382, 1996.

[16] C. Tomasi and T. Kanade. Shape and motion from image streams under orthography. *Intl. Journal of Computer Vision*, 9(2):137–154, 1992.