Utilizing Semantic Interpretation of Junctions for 3D-2D Pose Estimation

Florian Pilz[1], Yan Shi[1], Daniel Grest[1], Nicolas Pugeault[2], Sinan Kalkan[3], and Norbert Krüger[4]

[1] Medialogy Lab, Aalborg University Copenhagen, Denmark
[2] School of Informatics, University of Edinburgh, United Kingdom
[3] Bernstein Center for Computational Neuroscience, University of Göttingen, Germany
[4] Cognitive Vision Group, University of Southern Denmark, Denmark

Abstract. In this paper we investigate the quality of 3D-2D pose estimates using hand labeled line and point correspondences. We select point correspondences from junctions in the image, allowing to construct a meaningful interpretation about how the junction is formed, as proposed in e.g. [1], [2], [3]. We make us of this information referred as the semantic interpretation, to identify the different types of junctions (i.e. L-junctions and T-junctions). T-junctions often denote occluding contour, and thus do not designate a point in space. We show that the semantic interpretations is useful for the removal of these T-junction from correspondence sets, since they have a negative effect on motion estimates. Furthermore, we demonstrate the possibility to derive additional line correspondences from junctions using the semantic interpretation, providing more constraints and thereby more robust estimates.

1 Introduction

The knowledge about the motion of objects in a scene is crucial for many applications such as driver assistant systems, object recognition, collision avoidance and motion capture in animation. One important class of motion is the 'Rigid Body Motion' (RBM) which is defined as a continuous movement of the object, such that the distance between any two points of the object remains fixed at all times. The mathematical structure of this motion has been studied for a long while (see e.g., [4]). However, the problem of visual based motion estimation is far from being solved. Different methods for RBM estimation have been proposed [5] and can be separated into feature based, optic flow based and direct methods, where this work concerns a feature based method. In feature based RBM estimation, image features (e.g., junctions [6] or lines [7]) are extracted and their are correspondences defined. The process of extracting and matching correspondences suffers from high ambiguity, and is even more severe than the correspondence problem for stereopsis, since the epipolar constraint is not directly applicable. A Rigid Body Motion consisting of translation $\boldsymbol{t}$ and rotation $\boldsymbol{r}$ is described by six parameters, three for the translation $\boldsymbol{t} = (t_1, t_2, t_3)$ and three for rotation

G. Bebis et al. (Eds.): ISVC 2007, Part I, LNCS 4841, pp. 692–701, 2007.

$\boldsymbol{r} = (r_1, r_2, r_3)$. This allows for the formulation of the transformation between a visual entity in one frame, and the same entity in the next frame.

$$RBM^{(\boldsymbol{t},\boldsymbol{r})}(e) = e' \tag{1}$$

The problem of computing the RBM from correspondences between 3D objects and 2D image entities is referred as 3D-2D pose estimation [8,9]. The 3D entity (3D object information) needs to be associated to a 2D entity (2D knowledge of the same object in the next image) by the perspective projection P.

$$P(RBM^{(\boldsymbol{t},\boldsymbol{r})}(e)) = e' \tag{2}$$

There exist approaches (in the following called projective approaches) that formalize constraints directly on equation (2) (see e.g., [10]). An alternative is, instead of formalizing the pose estimation problem in the image plane, to associate a 3D entity to each 2D entity: A 2D image point together with the optical center of the camera spans a 3D line (see figure 1a) and an image line together with the optical center generates a 3D plane (see figure 1b). In case of a 2D point p we denote the 3D line that is generated in this way by $\boldsymbol{L}(p)$. Now the RBM estimation problem can be formulated for 3D entities

$$RBM^{(\boldsymbol{t},\boldsymbol{r})}(\boldsymbol{p}) \in \boldsymbol{L}(p) \tag{3}$$

where $\boldsymbol{p}$ is the 3D Point. Such a formulation in 3D has been applied by, e.g., [11,9], coding the RBM estimation problem in a twist representation that can be computed iteratively on a linearized approximation of the RBM.

In this paper we study how multiple correspondence types of visual entities influence RBM estimation performance. We identify T-junctions by making use of the semantic interpretation [3] and study their effect on RBM estimation performance. Furthermore we make use of the semantic interpretation to derive additional line correspondences for junctions.

The paper is structured as following: In section 2, we describe the formulation of the constraint equations that are used in the 3D-2D pose estimation algorithm. In section 3, we introduce the scenario in which our technique is applied. In section 4 we describe the concept of a semantic interpretation for junction and explain in more detail how this applies to the 3D-2D pose estimation problem.

2 Constraint Equations

We now want to formulate constraints between 2D image entities and 3D object entities, where a 2D image point together with the optical center of the camera spans a 3D-line (see fig. 1a) and an image line together with the optical center generates a 3D-plane (see fig. 1b).

A 3D line. $\mathbf{L}$ can be expressed as two 3D vectors $\mathbf{r}, \mathbf{m}$. The vector $\mathbf{r}$ describes the direction and $\mathbf{m}$ describes the moment which is the cross product of a point

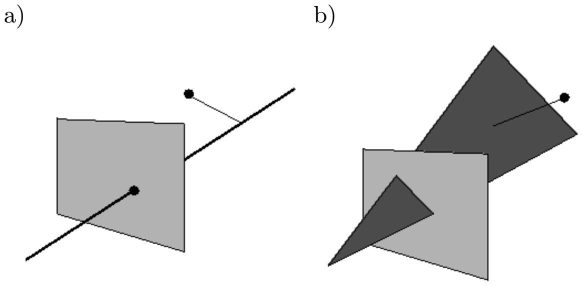


Fig. 1. Geometric Interpretation of constraint equations[12], a) Knowing the camera geometry a 3D-line can be generated from an image point and the optical center of the camera. The 3D- -point-3D-line constraint realizes the shortest Euclidian distance between the 3D-point and the 3D-line. b) From an image line a 3D-plane can be generated. The 3D-point-3D-plane constraint realizes the shortest Euclidian distance between the 3D-point and the 3D-plane.

$\mathbf{p}$ on the line and the direction $\mathbf{m} = \mathbf{p} \times \mathbf{r}$. The vectors $\mathbf{r}$ and $\mathbf{m}$ are called Plücker coordinates. The null space of the equation $\mathbf{x} \times \mathbf{r} - \mathbf{m} = \mathbf{0}$ is the set of all points on the line, and can be expressed in matrix form (see eq. 4). The creation of the 3D-line $\mathbf{L}$ from the 2D-point p together with the 3D-point $\mathbf{p}$ allows the formulation of the 3D-point-3D-line constraint [13] (eq.5),

$$\mathbf{F}^{\mathbf{L}}(\mathbf{x}) = \begin{pmatrix} 0 & r_x & -r_y & -m_x \\ -r_z & 0 & r_x & -m_y \\ r_y & -r_x & 0 & -m_z \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = \begin{pmatrix} 0 \\ 0 \\ 0 \end{pmatrix} \tag{4}$$

$$\mathbf{F}^{\mathbf{L}(p)} \left((I_{4\times 4} + \alpha\tilde{\xi})\mathbf{p} \right) = 0. \tag{5}$$

where $\tilde{\xi}$ is the matrix representing the linearisation of the RBM and α is a scale value 7 [12]. Note, that the value $||\mathbf{F}^L(\mathbf{x})||$ can be interpreted as the Euclidian distance between the point (x_1, x_2, x_3) and the closest point on the line to (x_1, x_2, x_3) [14,9]. Note that although we have 3 equations for one correspondence the matrix is of rank 2 resulting in 2 constraints.

A 3D-plane. $\mathbf{P}$ can be expressed by defining the components of the unit normal vector $\boldsymbol{n}$ and a scalar (Hesse distance) δ_h. The null space of the equation $\boldsymbol{n} \cdot \mathbf{x} - \delta_h = \mathbf{0}$ is the set of all points on the plane, and can be expressed in matrix form (see eq. 6). The creation of the 3D-plane $\mathbf{P}$ from the 2D-line l together with the 3D-point $\mathbf{p}$ allows the formulation of the 3D-point-3D-plane constraint (eq.7) [13].

$$F^{\mathbf{P}}(\mathbf{x}) = \begin{pmatrix} n_1 & n_2 & n_3 & -\delta_h \end{pmatrix} \begin{pmatrix} x_1 \\ x_2 \\ x_3 \\ 1 \end{pmatrix} = 0 \tag{6}$$

$$F^{\mathbf{P}(p)}\Big((I_{4\times 4}+\alpha\tilde{\xi})\mathbf{p}\Big)=0. \tag{7}$$

Note that the value $||F^P(\mathbf{x})||$ can be interpreted as the Euclidian distance between the point (x_1, x_2, x_3) and the closest point on the plane to (x_1, x_2, x_3) [14,9]. These 3D-point-3D-line and 3D-point-3D-plane constraints result in a system of linear equations, which solution becomes optimized iteratively (for details see [12]).

3 Ego-motion Estimation from Stereo Image Sequences

We apply the pose estimation algorithm for estimating the motion in stereo sequences. Here we do not have any model knowledge about the scene. Therefore the 3D entities need to be computed from stereo correspondences. We provide hand picked 2D-point and line correspondences in two consecutive stereo frames. From stereo correspondences (fig.2) we compute a 3D-point. The corresponding 2D-point or line in the next left frame (fig.3 a-f) provides either a 3D-line or 3D-plane. For each, one constraint can be derived (see eq.(5 and (7).

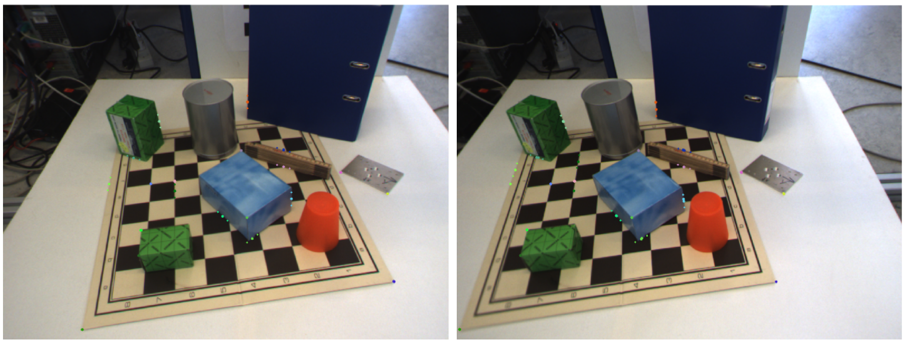


Fig. 2. Stereo image pair (left and right) used for computation of 3D entities

A 3D-point-2D-line correspondence leads to one independent constraint equation and a point-point (3D-point-2D-point) correspondence leads to two independent constraint equations [12]. Therefore, it is expected that 3D-point-2D-point correspondences produce more accurate estimates with smaller sets than 3D-point-2D-line correspondences [15].

4 Using Semantic Interpretation of Junctions

A junction is represented by a set of rays corresponding to the lines that intersect, each defined by an orientation [1] (see fig. 4). In [3] a more elaborate description can be found, where methods for junction detection and extraction

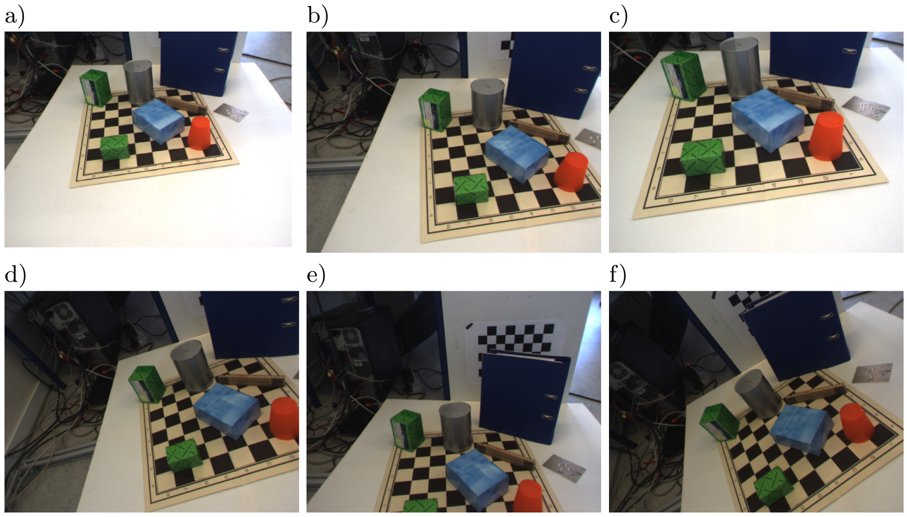


Fig. 3. Image Set: a) translation on x-axis (150 mm), b) translation on y-axis (150 mm), c) translation on z axis (150mm), d) rotation around x-axis (+20 deg), e) rotation around y-axis (-15 deg), f) rotation around z-axis (+30 deg)

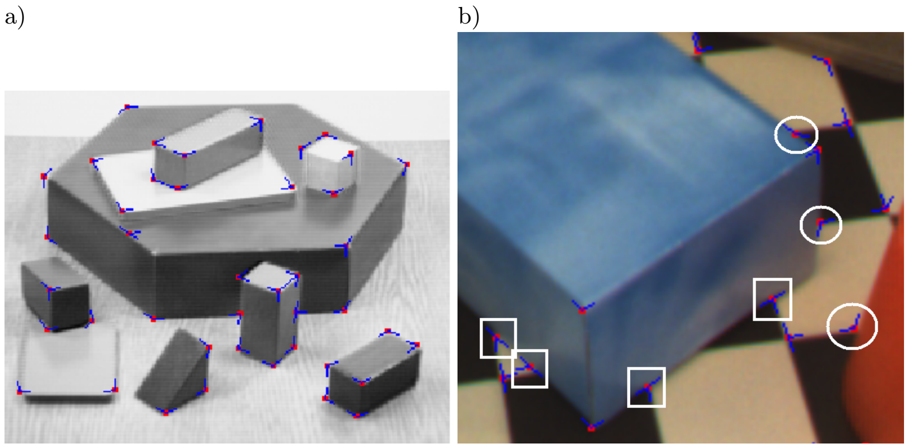


Fig. 4. a) Detected junctions with extracted semantic interpretation [3]. b) Examples of T-junctions (partly detected) - rectangles: valid T-junctions, circles: invalid T-junctions due depth discontinuity.

their semantic information are proposed. Since 3D-point-2D-point correspondences are junctions often, the additional semantic information can be extracted and utilized in the context of 3D-2D pose estimation. Feature-based motion estimation makes assumptions on the visual entities used as correspondences. One kind of uncertainty arises through the assumption, that correspondences

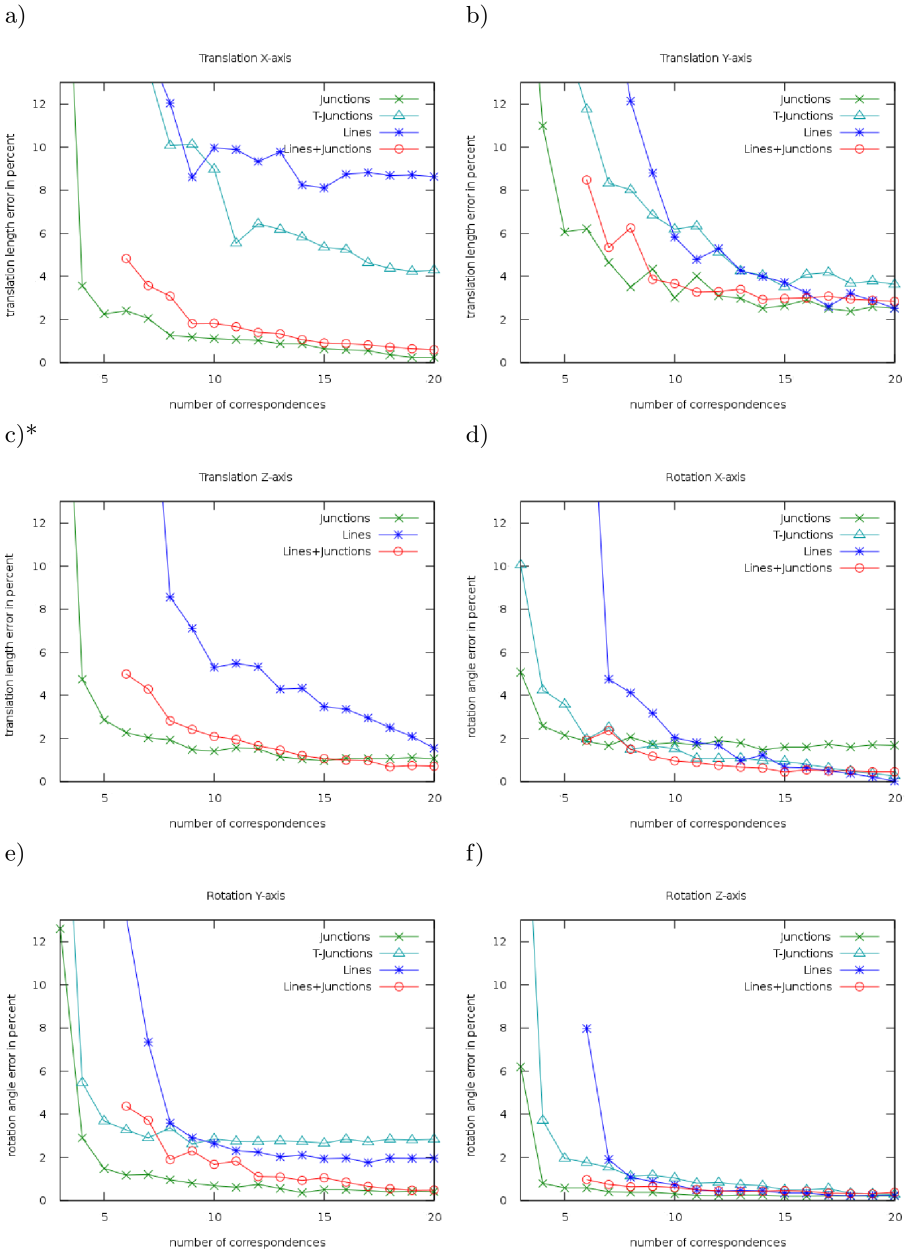
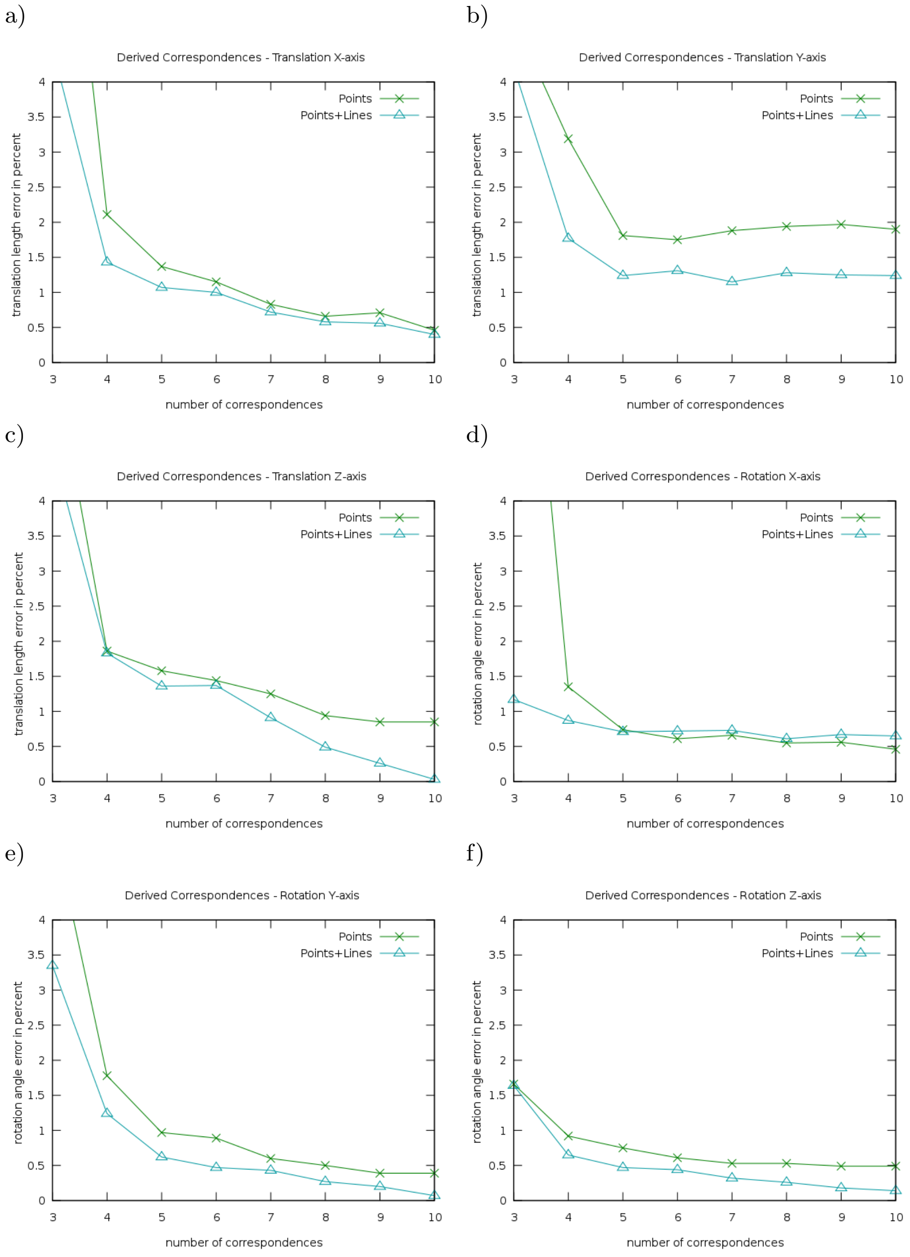


Fig. 5. translation length error in percent: a) translation on x-axis (150 mm), b) translation on y-axis (150 mm), c) translation on z axis (150mm), * here the error for T-junctions is too high to be shown. d) rotation around x-axis (+20 deg), e) rotation around y-axis (-15 deg),f) rotation around z-axis (+30 deg).



Derived Correspondences - Translation X-axis

Derived Correspondences - Translation Y-axis

Derived Correspondences - Translation Z-axis

Derived Correspondences - Rotation X-axis

Derived Correspondences - Rotation Y-axis

Derived Correspondences - Rotation Z-axis

Fig. 6. translation length error in percent for derived correspondence: a) translation on x-axis (150 mm), b) translation on y-axis (150 mm), c) translation on z axis (150mm), d) rotation around x-axis (+20 deg), e) rotation around y-axis (-15 deg),f) rotation around z-axis (+30 deg) are temporally stable. Certain kinds of visual entities such as T-junctions (see Fig.4b) may be an indication of depth discontinuity, having a negative effect on motion estimates. The use of the semantic interpretation allows us to identify and remove constraints based on these T-junction correspondences. Temporally unstable correspondences introduce error in two ways. First, the computed 3D point is errorenous and second the corresponding 2D point does not reflect the motion between to consecutive frames. Furthermore, if a semantic interpretation of a junction is known, further constraints can be derived for one location. The idea is to use the information of the lines to build up additional 3D-point 2D-line correspondences. In the case of point and line correspondences, it allows deriving line correspondences from existing point correspondences (e.g., for an L-junction this leads to one point and two line correspondences). For testing, we handpick 10 point correspondences (see figure 2 and 3a-f), estimate the pose and compare it with the estimation based one the very same 10 point correspondences and their intersecting edges.

5 Results

The images are recorded with a calibrated camera system mounted on a robotic arm. The error is computed as the difference between the estimated translation length and rotation angle compared to the ground, truth provided by the robot. The data set consist of a set of stereo images (Fig.2) at different time steps and six images for different arm motion (Fig.3a-f). The process of manually picking correspondences introduces a small position error, where lines and points are not picked with the same accuracy. To reduce this error difference between line and point correspondences, we add noise ($n \in [0, 1]$) to the 2D correspondences pixel positions. In Fig. 5 and Fig. 6 the error between estimated motion and ground truth is shown for different types of correspondences. The experiments were repeated 50 times for random subsets of correspondences with with a specific set size as shown on the x-axis. Results show that the error for 2D-point (junction) correspondences decreases faster than for 2D-line (edge) correspondences (see fig. 5a-f). Moreover it shows that more than 8 point or 15 line correspondence do not further minimize the error. Motion estimates from correspondence sets based on a combination of junctions and edges converge to nearly the same error than junction correspondence sets (see Fig.5a-f). In cases where lines perform better (see Fig. 5d), the error for mixed correspondence sets converges to a value similar to the best of both individual correspondence sets. Moreover, estimations are independent of motion direction or length and provide even for minimal solution sets accurate estimates.

We can make the general observation that correspondences from T-junctions have a negative influence on motion estimates. For translational motions the error is related to the motion direction. T-junctions introduce a noticeably error for forward/backward motions (x-axis) (Fig. 5a), while the error for sideways motions (y-axis) is insignificant (Fig. 5b). For motions along the z-axis the error introduced by T-junctions is enormous (around 250%) and therefore not displayed in figure 5c. For rotational motions the error introduced by T-junctions in this scenario is minimal (see fig. 5d-f).

If a semantic interpretation of junctions is available, further correspondences can be generated for these junctions. This provides more constraints and should therefore result in better RBM estimates. For a given L/Y junction the semantic interpretation provides the information of the intersecting lines. Each line is defined by its orientation and junctions position. In order to derive an additional line correspondence a point is constructed on the lines in some distance from the junction. Figure 6a-f shows that using the intersecting lines of a junction indeed results in more accurate estimates.

6 Discussion

The evaluation of line and point correspondences has shown, that 2D-point correspondences provide more valuable constraints than 2D-line, since they lead to more constraints. The results show that T-junctions may introduce errors and should therefore be avoided as point correspondences for motion estimation. The semantic interpretation is a way to identify and disregard these temporally inconsistent image entities, providing correspondence sets leading to more accurate and robust RBM estimations. It is questionnable whether line contraints derived from junctions provide additional information compared to those junctions. However, in the presence of noise, it is expected these additional constraints further reduce the estimation error. The results in Fig. 6 clearly confirm this expectation.

To summarize, the contribution shows that a combination of 2D-Points and 2D-Lines correspondences result in more accurate and robust motion estimations. The semantic interpretation of junctions (2D-points) allows us to disregard T-junctions and to derive additional line correspondences from junctions, providing more robust correspondences sets.

References

1. Parida, L., Geiger, D., Hummel, R.: Junctions: detection, classification and reconstruction. In: CVPR 1998. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 20, pp. 687–698 (1998)
2. Rohr, K.: Recognizing corners by fitting parametric models. International Journal of Computer Vision 9, 213–230 (1992)
3. Kalkan, S., Yan, S., Pilz, F., Krüger, N.: Improving junction detection by semantic interpretation. In: VISAPP 2007. International Conference on Computer Vision Theory and Applications (2007)
4. Ball, R.: The theory of screws. Cambridge University Press, Cambridge (1900)
5. Steinbach, E.: Data driven 3-D Rigid Body Motion and Structure Estimation. Shaker Verlag (2000)
6. Phong, T., Horaud, R., Yassine, A., Tao, P.: Object pose from 2-D to 3-D point and line correspondences. International Journal of Computer Vision 15, 225–243 (1995)

7. Krüger, N., Jäger, T., Perwass, C.: Extraction of object representations from stereo imagesequences utilizing statistical and deterministic regularities in visual data. In: DAGM Workshop on Cognitive Vision, pp. 92–100 (2002)
8. Grimson, W. (ed.): Object Recognition by Computer. MIT Press, Cambridge (1990)
9. Rosenhahn, B., Sommer, G.: Adaptive pose estimation for different corresponding entities. In: van Gool, L. (ed.) Pattern Recognition, 24th DAGM Symposium, pp. 265–273. Springer, Heidelberg (2002)
10. Araujo, H., Carceroni, R., Brown, C.: A fully projective formulation to improve the accuracy of lowe's pose–estimation algorithm. Computer Vision and Image Understanding 70, 227–238 (1998)
11. Rosenhahn, B., Perwass, C., Sommer, G.: Cvonline: Foundations about 2d-3d pose estimation. In: Fisher, R. (ed.) CVonline: On-Line Compendium of Computer Vision [Online] (2004), http://homepages.inf.ed.ac.uk/rbf/CVonline/
12. Krüger, N., Wörgötter, F.: Statistical and deterministic regularities: Utilisation of motion and grouping in biological and artificial visual systems. Advances in Imaging and Electron Physics 131, 82–147 (2004)
13. Wettegren, B., Christensen, L., Rosenhahn, B., Granert, O., Krüger, N.: Image uncertainty and pose estimation in 3d euclidian space. In: DSAGM 2005. Proceedings DSAGM 13th Danish Conference on Pattern Recognition and Image Analysis, pp. 76–84 (2005)
14. J.M., S.: Some remarks on the statistics of pose estimation. Technical Report SBU-CISM-00-25, South Bank University, London (2000)
15. Liu, Y., Huang, T., Faugeras, O.: Determination of camera location from 2-d to 3-d line and point correspondence. In: CVPR 1989. IEEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 12, pp. 28–37 (1989)