Monocular Simultaneous Multi-Body Motion Segmentation and Reconstruction from Perspective Views

Reza Sabzevari and Davide Scaramuzza

Abstract— **In this paper, we tackle the problem of mapping multiple 3D rigid structures and estimating their motions from perspective views through a car-mounted camera. The proposed method complements conventional localization and mapping algorithms (such as Visual Odometry and SLAM) to estimate motions of other moving objects in addition to the vehicle's motion. We present a theoretical framework for robust estimation of multiple motions and structures from perspective images. The method is based on the factorization of the projective trajectory matrix without explicit estimation of projective depth values. We exploit the epipolar geometry of calibrated cameras to generate several hypotheses for motion segments. Once the hypotheses are obtained, they are evaluated in an iterative scheme by alternating between estimation of 3D structures and estimation of multiple motions. The proposed framework does not require any knowledge about the number of motions and is robust to noisy image measurements. The method is evaluated on street-level sequences from a car-mounted camera. A benchmark dataset is also used to compare the results with previous works, although most of the related works use synthetic scenes simulating desktop environments.**

I. INTRODUCTION

This paper addresses the problem of simultaneous estimation of multiple motions of rigid objects and their 3D structure from 2D image correspondences. Such 2D points belong to a sequence of images captured by a car-mounted camera under *perspective* camera model. The method presented in this article has strong ties to *Structure from Motion* (SfM), and consequently it provides complementary capabilities to related problems in robotics, i.e. *Visual Odometry* (VO) and SLAM.

A. Motivations

During the last decade, VO and SLAM have been widely studied and recent advancements in VO have resulted in fairly accurate estimation of vehicles' motion with respect to the environment [1]. Using VO, the vehicle can localize itself with respect to the static parts of the scene while all the moving parts are treated as outliers. The VO pipeline can provide a sparse reconstruction of the environment as well as the positions of the vehicle with respect to the sparse map. The missing parts in this pipeline are localization of other moving objects and estimation of their motions as well as reconstructing the 3D structures of these objects. Such additions to VO would be beneficial to driver assistance systems in estimating the motions of other vehicles and pedestrians. This paper proposes a framework that provides the aforementioned missing parts in navigation algorithms that are widely used in robotics.

The authors are with the Robotics and Perception Group, University of Zurich, Switzerland—http://rpg.ifi.uzh.ch

This research was supported by the Hasler Foundation—project number 13027—and the Swiss National Science Foundation through the National Center of Competence in Research Robotics.

B. Related Work

The problem of estimating multiple motions and structures from 2D correspondences of multiple images is known as *Multi-body Structure from Motion* or *motion segmentation and estimation*. The works on multi-body SfM problem can be categorized into two major groups: *i)* approaches for affine camera model, and *ii)* methods that can also be applied to perspective camera model. Solving this problem under perspective camera model is more challenging, because it requires estimation of the perspective depth scales, which is also a challenging problem by itself. Murakami et al. [2] studied the conditions that a projective factorization is feasible without estimating the projective depth values, and showed that is possible only under strict assumptions; i.e. all the camera coordinates lie on a single plane that is parallel to all image planes.

There are several works in the literature that provide rather good motion segmentation and estimation under affine camera model. The seminal work of Costeira and Kanade [3] formulated the multi-body SfM as a factorization problem. The framework proposed by them is for the affine and orthographic camera models, as discussed in details in Section II. Since this method is purely based on algebraic calculations, it is very sensitive to noise and outliers. Yan and Pollefeys contribution in [4] can fairly handle outliers but since it relies on local estimation of subspaces, it cannot handle the cases where two or more parts of the scene have the same motion but are not spatially correlated.

On the contrary, the multi-body SfM problem has not been well studied for perspective images. Vidal et al. [5] proposed a geometric approach to estimate multiple structures and motions from two perspective views. This work was then extended to three views in [6], but since both methods are geometric based approaches, they are not robust to noise and, thus, cannot be used for real-world applications. Schindler et al. [7] proposed a method for n-view multi-body SfM based on model selection. Their method uses 2-view geometry and, by linking motion segments between multiple pairs of frames, it propagates the initial segmentation to n-views. Differently, Li et al. [8] proposed a factorization approach to identify multiple rigid motions in perspective images. The method is based on an initial estimation of projective depth scales and consequently is not robust to noise. The details of this approach are discussed in Section II.

The opportunities that multi-body SfM provides to the robotics applications, especially to VO, is rarely investigated in the literature. Furthermore, most of the experiments in the literature are based on synthetic datasets. One of the few works in this context was done by Vidal in [9], who applied subspace clustering techniques to motion segmentation in perspective images. However, the motion segmentation was applied on optical flow information of an outdoor sequence to segment the motions but not for estimating the motions.

C. Contributions

This article proposes a theoretical framework to estimate motions and 3D structures of multiple rigid-body objects from perspective views collected by a car-mounted camera. The method is based on matrix factorization and assumes that the camera is calibrated. Unlike other factorization methods, which are based on a good initialization of motion segments, our method generates several hypotheses for initial motion-segments, which makes it more robust to noise. Experiments on a real-world dataset show that, despite the presence of many outliers in case of change of illumination in outdoor environments, the method provides accurate motion segmentation and reliable reconstruction. Furthermore since the proposed framework estimates motions and structures in an incremental way (does not need all the image frames at once), it is intrinsically compatible with similar problems in robotics (e.g. SLAM and VO) and can be easily integrated in such approaches.

D. Structure of the Paper

In section II, the theoretical background of single-body and multi-body SfM both for affine and perspective views is discussed. The discussion in the next section is mainly based on the factorization approaches to different variations of SfM problem for rigid motions. In Section III, the proposed framework and its theoretical concepts are thoroughly described. Results of the proposed approach on a street-level dataset as well as comparisons with other methods on a benchmark dataset are presented in Section IV.

II. Structure and Motion: a Factorization Approach

Structure from motion can be considered as the simultaneous solution for two dual problems: *i)* recovering an unknown structure from known camera positions, *ii)* determining of viewer's positions or camera motion from a set of known 2D correspondences. In general, 3D structure and camera motion can be estimated by applying epipolar geometry between every pair of images or using the multi-view geometry. The inter-image relations are linked by the fact that a unique shape is projected onto the images captured from different views. Since the extracted features are sparse 2D image points, the estimated 3D structure is also a sparse 3D point cloud.

Consider a set of p 2D point correspondences in multiple views accumulated in a matrix $\mathtt{W}$. Given matrix $\mathtt{W}$, the SfM problem consists of solving simultaneously for the position of points in 3-D space, denoted as $\mathtt{S}$, and the relative pose of the cameras representing the motion, denoted as $\mathtt{M}$. A set of popular approaches estimate $\mathtt{M}$ and $\mathtt{S}$ matrices via factorization approaches using solely the collection of such 2D image point correspondences (i.e. matrix $\mathtt{W}$) [10], [11] and [3].

A. Single Motion, Rigid Object and Affine Camera

For the affine camera model, rigid SfM problem can be formulated in the mathematical context of bilinear matrix factorization. So, the 2D image trajectories used by SfM can be described by bilinear matrix models [11]. In more detail, by defining the image coordinate of a point i in frame g as the vector $\mathbf{w}_{gi} = [x_{gi}\ y_{gi}]^T$, we may write the measurement matrix $\mathtt{W}$ that gathers the coordinates of all the points in all the views as:

$$\mathtt{W} = \begin{bmatrix} \mathbf{w}_{11} & \dots & \mathbf{w}_{1p} \\ \vdots & \ddots & \vdots \\ \mathbf{w}_{f1} & \dots & \mathbf{w}_{fp} \end{bmatrix} = [\overline{\mathbf{w}}_1\ \dots\ \overline{\mathbf{w}}_p], \tag{1}$$

where f is the number of frames ($g = 1 \dots f$), p is the number of points ($i = 1 \dots p$) and vector $\overline{\mathbf{w}}_i$ is column i of matrix $\mathtt{W}$ that represents image coordinates for i-th point in f views. In case of a rigid object, the camera motion matrices $\mathtt{M}_g$ and the 3D points $\mathbf{s}_i$ can be expressed as:

$$\mathtt{M}_g = \left[\begin{array}{ccc|c} R_{g1} & R_{g3} & R_{g5} & t_{g1} \\ R_{g2} & R_{g4} & R_{g6} & t_{g2} \end{array}\right] \quad \text{and} \quad \mathbf{s}_i = \begin{bmatrix} X_i \\ Y_i \\ Z_i \\ 1 \end{bmatrix}, \tag{2}$$

where $\mathtt{M}_g \in \mathbb{R}^{2\times 4}$ is the projection matrix containing rotation and translation components and $\mathbf{s}_i$ is a 4-vector containing the coordinate of the i-th point in 3D space. The recovered structure $\mathbf{s}_i$ and motion $\mathtt{M}_g$ are up to a nonsingular linear transformation and Euclidean upgrade is possible via a 4×4 transformation matrix. So, a 2D point i in a frame g is given by $\mathbf{w}_{gi} = \mathtt{M}_g\,\mathbf{s}_i$.

We can collect all the image measurements and their respective bilinear components $\mathtt{M}_g$ and $\mathbf{s}_i$ in a global matrix form. Thus, the factorization model of image trajectories can be formulated as

$$\mathtt{W}_{2f\times p} = \mathtt{M}_{2f\times 4}\ \mathtt{S}_{4\times p}, \tag{3}$$

where the bilinear components $\mathtt{M}$ and $\mathtt{S}$ are defined as:

$$\mathtt{M} = \begin{bmatrix} \mathtt{M}_1 \\ \vdots \\ \mathtt{M}_f \end{bmatrix} \quad \text{and} \quad \mathtt{S} = \begin{bmatrix} \mathbf{s}_1 & \cdots & \mathbf{s}_p \end{bmatrix}. \tag{4}$$

In general, the rank of $\mathtt{W}$ is constrained to be $\text{rank}\{\mathtt{W}\} \le r$ where $r \ll \min\{2\times f,\ p\}$. For case of affine camera model the rank of matrix $\mathtt{W}$ would be at most four. In practice, the image measurements cannot be noise free, which increases the rank of matrix $\mathtt{W}$. So, the rank-four constraint should be enforced in the factorization.

B. Single Motion, Rigid Object and Perspective Camera

In case of the perspective camera model, vector $\mathbf{w}_{gi}$ in Eq. (1) denotes the homogeneous coordinates of i-th point in g-th image frame that is scaled by the projective depth λ_{gi}, such that

$$\mathbf{w}_{gi} = \lambda_{gi} \; [x_{gi} \;\; y_{gi} \;\; 1]^\top = [u_{gi} \;\; v_{gi} \;\; \lambda_{gi}]^\top . \tag{5}$$

Consequently, Eq. (3) becomes:

$$\mathtt{W}_{3f\times p} = \mathtt{M}_{3f\times 4} \;\; \mathtt{S}_{4\times p} \, . \tag{6}$$

The rank-4 factorization of Eq. (6) is possible if the depth scales λ_{gi} are known. Using epipolar geometry, Sturm and Triggs [10] proposed a method to estimate λ_{gi} up to a global scale factor. This can be achieved by estimating the fundamental matrices $\mathtt{F}_{gg'}$ and, consequently, the epipoles $\mathbf{e}_{gg'}$ that relate every pair of consecutive frames g and g'. These two elements ($\mathtt{F}_{gg'}$ and $\mathbf{e}_{gg'}$) can be estimated in a *least-squares* manner using the 8-point algorithm [12]. Thus, the relation between depth scales λ_{gi} and $\lambda_{g'i}$ in two consecutive frames will be as:

$$\lambda_{gi} = \frac{\left(\mathbf{e}_{gg'} \times \mathbf{w}_{gi}\right)^\top \left(\mathtt{F}_{gg'}\mathbf{w}_{g'i}\right)}{\left\|\mathbf{e}_{gg'} \times \mathbf{w}_{gi}\right\|^2} \, \lambda_{g'i} \, . \tag{7}$$

By writing Eq. (7) for every pair of corresponding image points and every pair of consecutive image frames, the depth values can be recovered recursively up to an arbitrary initial value for λ_{1i}. In practice, the image measurements are noisy, and relying only on geometric estimations will not provide enough robustness. The robustness can be increased by iteratively alternating between two steps: *i)* rank-4 estimation of matrices of structure $\mathtt{S}$ and motion $\mathtt{M}$ (given an initial estimate for depth values λ_{gi}), *ii)* estimating the depth values that improve the previous estimations of structure and motion [13]. In more detail, if the depth values are initialized as $\lambda_{gi} = 1$, then the best rank-4 estimation of $\mathtt{W}$ will be:

$$\begin{aligned} \mathtt{W}_{3f\times p} &\approx \tilde{\mathtt{M}}_{3f\times 4} \, \tilde{\mathtt{S}}_{4\times p}, \\ \tilde{\mathtt{W}} &= \tilde{\mathtt{M}} \, \tilde{\mathtt{S}}, \end{aligned} \tag{8}$$

where $\tilde{\mathtt{S}}$ and $\tilde{\mathtt{M}}$ are the best rank-4 estimations for structure and motion, respectively, and $\tilde{\mathtt{W}}$ is an approximation of $\mathtt{W}$ given by $\tilde{\mathtt{S}}$ and $\tilde{\mathtt{M}}$. Once the estimations for motion and structure are obtained, the depth values are estimated as:

$$\lambda_{gi} = \left\|\mathbf{w}_{ig} - \tilde{\mathbf{w}}_{ig}\right\|, \tag{9}$$

where $\tilde{\mathbf{w}}_{ig}$ is an approximation of $\mathbf{w}_{ig}$ given by Eq. (8). In [13] the convergence of such iterative scheme for estimation of depth scale as well as structure and motion is proved.

C. Multiple Motions of Rigid Objects

If the 2D image correspondences belong to the motions of multiple objects, the image measurement matrix $\mathtt{W}$ that envelopes all the image correspondences belonging to several motions can be written as:

$$\mathtt{W} = [\mathtt{W}_1 | \mathtt{W}_2 | \ldots | \mathtt{W}_n] , \tag{10}$$

where n is the number of motions and $\mathtt{W}_j$, $j = 1 \ldots n$, is the matrix containing 2D point correspondences belonging to the j-th motion. Basically, matrix $\mathtt{W}$ is the horizontal concatenation of $\mathtt{W}_j$ matrices, each containing p_j points that comply with motion j, where $p = \sum_{j=1}^{n} p_j$ is the total number of points for all the motions. So, the camera motion matrix $\mathtt{M}$ and the structure matrix $\mathtt{S}$ can be written as:

$$\mathtt{M} = [\mathtt{M}_1 | \mathtt{M}_2 | \ldots | \mathtt{M}_n] \;\; \text{and} \;\; \mathtt{S} = \begin{bmatrix} \mathtt{S}_1 & \mathtt{0} & \ldots & \mathtt{0} \\ \mathtt{0} & \mathtt{S}_2 & \ldots & \mathtt{0} \\ \vdots & \vdots & \ddots & \vdots \\ \mathtt{0} & \mathtt{0} & \ldots & \mathtt{S}_n \end{bmatrix} . \tag{11}$$

In this case, the generic SfM equation, $\mathtt{W} = \mathtt{M}\,\mathtt{S}$, can be rewritten as:

$$[\mathtt{W}_1 | \ldots | \mathtt{W}_n] = [\mathtt{M}_1 | \ldots | \mathtt{M}_n] \cdot \begin{bmatrix} \mathtt{S}_1 & & \\ & \ddots & \\ & & \mathtt{S}_n \end{bmatrix} . \tag{12}$$

For the affine camera model, $\mathtt{W} \in \mathbb{R}^{2f\times p}$ and $\mathtt{W}_j \in \mathbb{R}^{2f\times p_j}$ contain image coordinates. So, the camera motion matrix $\mathtt{M}$ belongs to $\mathbb{R}^{2f\times 4n}$, which represents the horizontal concatenation of individual motion matrices $\mathtt{M}_j \in \mathbb{R}^{2f\times 4}$. Similarly, the structure matrix $\mathtt{S} \in \mathbb{R}^{4n\times p}$ is the diagonal concatenation of individual structures $\mathtt{S}_j \in \mathbb{R}^{4\times p_j}$. To recover multiple structures and motions, the sparse structure of $\mathtt{S}$ is employed and, using Eq. (12), the image measurement matrix $\mathtt{W}$ is factorized in a way that minimizes the noise in zero areas of matrix $\mathtt{S}$. This can be achieved by iteratively alternating between estimating two components: *i)* the 3D structures by maximizing the sparsity of matrix $\mathtt{S}$, *ii)* the motion matrices by minimizing the reprojection error and discarding the points from matrices $\mathtt{W}$ and $\mathtt{S}$ that cause large reprojection errors. Costeira and Kanade [3] proposed this factorization method for the orthographic and affine camera models. The main flaw of this method is that a good initialization—usually close to the final solution—of motion segmentation is required, otherwise, most of the points will be discarded as they violate the sparsity constraint.

For the case of perspective camera model, matrix $\mathtt{W}$ belongs to $\mathbb{R}^{3f\times p}$ and matrix $\mathtt{W}_j \in \mathbb{R}^{3f\times p_j}$, both holding homogenous image coordinates scaled by depth values λ_{gi}. Consequently, matrix $\mathtt{M}$ is a $3f \times 4n$ matrix which contains individual motion matrices $\mathtt{M}_j \in \mathbb{R}^{3f\times 4}$. Li et al. [8] proposed an approach for projective factorization of multiple rigid motions based on depth estimation of Strum and Triggs [10]. In their method, an initial motion segmentation as well as initial depth estimation are required. An iterative refinement stage alternates between estimating the depth values and motion segments. Once the motion segments and depth values are converged, motion and structure for each motion-segment are estimated via factorization.

III. METHOD

In this section, the proposed methodology for estimating relative motion and structure of independently moving objects is discussed. Given f perspective views of p points belonging to rigid objects moving under n classes of motions, the goal is to segment these points based on their motions, estimate the motions, and recover the position of these points in the 3D coordinate.

In more detail, consider set $\mathcal{P} = \{P_1, \dots P_p\}$ containing indices for p point trajectories, such that:

$$\mathcal{P} = \bigcup_{j=1}^{n} \mathcal{P}_j, \tag{13}$$

where $\mathcal{P}_j$ is the set of point trajectories that obey motion j. Thus, set $\mathcal{P}_j$ will include p_j columns of matrix $\mathtt{W}$ (see Eq. (1)) such that:

$$\begin{aligned} \mathcal{P}_j &= \left\{ \overline{\mathbf{w}}_j^{(1)}, \dots, \overline{\mathbf{w}}_j^{(p_j)} \right\}, \\ \mathtt{W}_j &= \left[\overline{\mathbf{w}}_j^{(1)}, \dots, \overline{\mathbf{w}}_j^{(p_j)} \right], \end{aligned} \tag{14}$$

where matrix $\mathtt{W}_j$ contains all columns of matrix $\mathtt{W}$ that had a similar motion among f frames.

So, finding subsets of $\mathcal{P}$, holding Eq. (13), results in a motion segmentation hypothesis. Given ψ hypotheses for motion segments, they should be evaluated by calculating the reprojection error with respect to all the estimated motions and structures. In the evaluation phase, matrices $\mathtt{W}$, $\mathtt{S}$, and $\mathtt{M}$ as in Eq. (12) are formed with respect to one of the motion hypotheses. After initializing these matrices, the reprojection error is minimized by iteratively alternating between estimating the structures $\mathtt{S}$ while fixing matrix $\mathtt{M}$ and estimating the motions $\mathtt{M}$ while fixing matrix $\mathtt{S}$. The reprojection error for all motion segmentation hypotheses are calculated and the ones with the smallest reprojection error is reported as the best motion segmentation hypothesis describing the trajectory matrix $\mathtt{W}$. The outline of our algorithm is presented in Alg. (1).

Algorithm 1 Outline of Simultaneous Motion Segmentation and Reconstruction

Input: *2D image correspondences*
Output: *Motions and structures of rigid bodies*

1: Generate hypotheses for motion segments (*see Alg. (2)*)
2: Compute reprojection error for every motion segmentation hypothesis (*see Alg. (3)*)
3: **return** The structures and motions for the hypothesis with the smallest reprojection error

A. Generating Hypotheses for Motion Segments

In this section, our approach for generating hypotheses for segmenting p point trajectories into n motions is discussed. Such hypotheses for motion segments is used to initialize the algorithm. To generate a hypothesis, a set of sample points from the trajectory matrix $\mathtt{W}$ that have been moved similarly among the f frames are selected. Then, by estimating the motion from these sample points, other points in the trajectory matrix are evaluated to identify those that comply with the same motion as the sample set. This process is repeated with the reminders of matrix $\mathtt{W}$ in a multi-RANSAC scheme, as presented in [14], fitting a single-motion rigid-body SfM model to identify the motion segments.

Given set $\mathcal{P}$, we would like to sample k points from this set that represent a unique motion, where k is the minimum number of points that represent a motion. Given the fact that the points closer to each other are more probable to be from the same structure, such minimal set is generated in a way that neighboring points are selected with higher probability. So, if point P_i has already been selected, then point P_i' will be selected with the following probability:

$$P(P_i|P_i') = \begin{cases} \frac{1}{\zeta} \exp -\frac{\|P_i - P_i'\|^2}{\sigma^2} & \text{if } P_i \neq P_i' \\ 0 & \text{if } P_i = P_i' \end{cases}, \tag{15}$$

where ζ is the normalization constant and σ will be selected heuristically [14].

Considering Eq. (15), for each hypothesis, k points from set $\mathcal{P}$ are selected, and a new trajectory matrix $\mathtt{W}_j^{(s)}$ is constructed using these points. Afterwards, the set of point trajectories encapsulated in matrix $\mathtt{W}_j^{(s)}$ should be verified to see if they represent a unique motion. To that end, for every pair of frames, point correspondences in matrix $\mathtt{W}_j^{(s)}$ are triangulated and the perspective camera motion as well as the 3D structure will be estimated by enforcing the epipolar constraints [15], such that:

$$(\mathbf{w}_{g'i} \otimes \mathbf{w}_{gi})\ \mathtt{E} = 0, \tag{16}$$

where vectors $\mathbf{w}_{g'i}$ and $\mathbf{w}_{gi}$ are normalized vectors in homogeneous coordinates (i.e. $\mathbf{w}_{gi} = [x_{gi} \quad y_{gi} \quad 1]^\top$) representing two corresponding points on a pair of frames and matrix $\mathtt{E}$ is the essential matrix.[1]

By writing Eq. (16) for all the k points, we will have a linear system of equations, of which the essential matrix $\mathtt{E}$ will be the null space of $(\mathbf{w}_{g'i} \otimes \mathbf{w}_{gi})$. Once the essential matrix is estimated, the motion matrix $\mathtt{M}$ will be given by *singular value decomposition* of $\mathtt{E}$ as:

$$\begin{aligned} \mathtt{E} &= \hat{\mathtt{T}}\, \mathtt{R}, \\ \hat{\mathtt{T}} &= \begin{bmatrix} 0 & -t_z & t_y \\ t_z & 0 & -t_x \\ -t_y & t_x & 0 \end{bmatrix}, \quad \mathbf{t} = [t_x \ t_y \ t_z]^\top, \\ \mathtt{M} &= [\mathtt{R} \mid \mathbf{t}], \end{aligned} \tag{17}$$

where $\mathtt{R} \in \mathbb{R}^{3\times3}$ and $\mathtt{M} \in \mathbb{R}^{3\times4}$. The estimated motion matrix $\mathtt{M}$ in Eq. (17) represents the relative motion (including rotation

[1] Operator $\otimes$ denotes the *Kronecker product*.

and translation) between two frames up to an arbitrary scale factor.

A calibrated camera can be described by an image plane perpendicular to the Z-axis and distanced 1 from the origin [15]. So, Eq. (5) turns to $\mathbf{w}_{gi} = [x_{gi} \;\; y_{gi} \;\; 1]$ and there is no need to estimate the depth scales as in Eq. (9). Furthermore, if the image points are normalized to lie on the unit sphere (i.e. $x_{gi}^2 + y_{gi}^2 + z_{gi}^2 = 1$) instead of the image plane, the 3D structure can be estimated up to a general scale factor. Thus, the estimation of 3D structure and camera motion can be formulated as a bilinear factorization problem (see Eq. (6)).

All k points in trajectory matrix $\mathtt{W}_j^{(s)}$ agree on a unique motion if the reprojection error of the estimated structure is less than a threshold ε, such that:

$$\|\mathtt{W}_j^{(s)} - (\mathtt{M}_j \;\; \mathtt{S}_j^{(s)})\| < \varepsilon, \tag{18}$$

where matrix $\mathtt{M}_j$ is the estimated camera motion and matrix $\mathtt{S}_j^{(s)}$ contains the estimated structure. If Eq. (18) does not hold, sampling points from matrix $\mathtt{W}$ continues until a set of points that have a similar motion is identified.

Once a motion is identified, other points in set $\mathcal{P}$ will be verified to check whether they comply with the identified motion using:

$$\mathtt{S}_{\bar{j}} = \mathtt{M}_j^{\top} \;\; \mathtt{W}_{\bar{j}}, \tag{19}$$

where matrices $\mathtt{S}_{\bar{j}}$ and $\mathtt{W}_{\bar{j}}$ represent the points that are in set $\mathcal{P} - \mathcal{P}_j$.

To generate a motion segmentation hypothesis, this process will be repeated until all the points in $\mathcal{P}$ (or the columns of trajectory matrix $\mathtt{W}$) are associated to a motion-segment. Alg. (2) shows the algorithm for generating ψ motion segmentation hypothesis.

Algorithm 2 Generating hypotheses for motion segmentation

Input: *2D image correspondences*
Output: *Several motion segmentation hypotheses*

for $c = 1$ to ψ **do**
  ▷ *% generate ψ hypotheses for motion segments%*
  while $\mathcal{P} \neq \varnothing$ **do**
    j=1 ▷ *% j represents the motion index%*
    while (reprojection error $> \varepsilon$) **do**
      ▷ *% reject invalid hypotheses%*
      Sample k points from set $\mathcal{P}$ and form $\mathtt{W}_j^{(s)}$
      Estimate $\mathtt{M}_j$ and $\mathtt{S}_j^{(s)}$
      ▷ *% using epipolar geometry%*
      Calculate the reprojection error
    end while
    Remove points $\mathtt{W}_j^{(s)}$ from $\mathtt{W}$
    Estimate structure for $\mathtt{W}$ with respect to $\mathtt{M}_j$
    Remove points from $\mathcal{P}$ that comply with $\mathtt{M}_j$
    j=j+1
  end while
end for

B. Evaluating Motion Segmentation Hypotheses

From every hypothesis, an initial estimate for motion segments and 3D structures is given. This helps to form matrices $\mathtt{W}$, $\mathtt{M}$ and $\mathtt{S}$ as in Eq. (12). Once these matrices are formed, the estimation of structures and motion-segments is refined iteratively. This can be achieved by alternatively estimating the structures matrix $\tilde{\mathtt{S}}$ while fixing motions and estimating the motions matrix $\tilde{\mathtt{M}}$ while fixing structures, where matrices $\tilde{\mathtt{M}}$ and $\tilde{\mathtt{S}}$ are defined in Eq. (8).

Considering (8) and (12), given the multiple motions matrix $\tilde{\mathtt{M}}$, estimation of multiple structures matrix $\tilde{\mathtt{S}}$ can be formalized as an optimization problem that solves a linear system of equations.

In more detail, Eq. (8) can be rewritten in form of $\mathtt{A}\mathbf{x} = \mathbf{b}$, such as:

$$\widehat{\mathtt{M}} \;\; \overrightarrow{\mathtt{S}} = vec(\tilde{\mathtt{W}}), \tag{20}$$

where matrix $\widehat{\mathtt{M}} \in \mathbb{R}^{3fp \times 4np}$ contains $4np_j$ columns for every motion in a block-diagonal way, and is defined as:

$$\widehat{\mathtt{M}} = \begin{bmatrix} \widehat{\mathtt{M}}_1 & & & \\ & \widehat{\mathtt{M}}_2 & & \\ & & \ddots & \\ & & & \widehat{\mathtt{M}}_n \end{bmatrix}_{3fp \times 4np},$$

$$\widehat{\mathtt{M}}_j = \begin{bmatrix} \breve{\mathtt{M}}_1^{(j)} \;\; \breve{\mathtt{M}}_2^{(j)} \;\; \ldots \;\; \breve{\mathtt{M}}_f^{(j)} \end{bmatrix}^{\top}_{3fp_j \times 4np_j}, \qquad j = 1 \ldots n, \tag{21}$$

$$\breve{\mathtt{M}}_g^{(j)} = \begin{bmatrix} \tilde{\mathtt{M}}_g & & \\ & \ddots & \\ & & \tilde{\mathtt{M}}_g \end{bmatrix}_{3p_j \times 4np_j}, \qquad g = 1 \ldots f,$$

$$\tilde{\mathtt{M}}_g = \left[\tilde{\mathtt{M}}_{g1} \mid \tilde{\mathtt{M}}_{g2} \mid \ldots \mid \tilde{\mathtt{M}}_{gn}\right], \qquad \tilde{\mathtt{M}}_g \in \mathbb{R}_{3 \times 4n},$$

and $\overrightarrow{\mathtt{S}}$ is a column-wise vectorization of matrix $\tilde{\mathtt{S}}$, such that:

$$\overrightarrow{\mathtt{S}}_{4np \times 1} = \left[\mathbf{s}_1 \, \mathbf{0}_a \, \mathbf{s}_2 \, \ldots \, \mathbf{0}_a \, \mathbf{s}_{p_1} \, \mathbf{0}_{a'} \, \ldots \, \mathbf{0}_a \, \mathbf{s}_{p_j} \, \mathbf{0}_{a'} \, \ldots \, \mathbf{0}_a \, \mathbf{s}_{p_n}\right]^{\top}, \tag{22}$$

where $\mathbf{0}_a$ and $\mathbf{0}_{a'}$ are vectors of a and a' zeros, where $a = 4(n-1)$ and $a' = 4n$. Finally, $vec(\tilde{\mathtt{W}})$ is the column-wise vectorization of $\tilde{\mathtt{W}}$. Structure of these matrices is shown in Fig. (1).

Now, we can solve Eq. (20) to estimate the structures. The equations belonging to non-zero values of $\overrightarrow{\mathtt{S}}$ can be used to create systems of equations to estimate structures in a *least-squares* sense. To that end, every non-zero block of $\overrightarrow{\mathtt{S}}$—representing a moving structure—forms an independent linear system of equations which can be solved individually. Note that, it is also possible to exploit the sparsity of vector $\overrightarrow{\mathtt{S}}$ as an additional constraint in the optimization process (as in [3]) and solve Eq. (20) for all the structures and motions simultaneously.

To estimate the motions, we can rewrite Eq. (8) as:

$$(\tilde{\mathtt{S}}^{\top} \otimes \mathtt{I}_{3f})\ \ vec(\tilde{\mathtt{M}}) = vec(\tilde{\mathtt{W}}), \tag{23}$$

where $\mathtt{I}_{3f}$ is a $3f \times 3f$ identity matrix, $vec(\tilde{\mathtt{M}})$ is column-wise vectorization of $\tilde{\mathtt{M}}$.

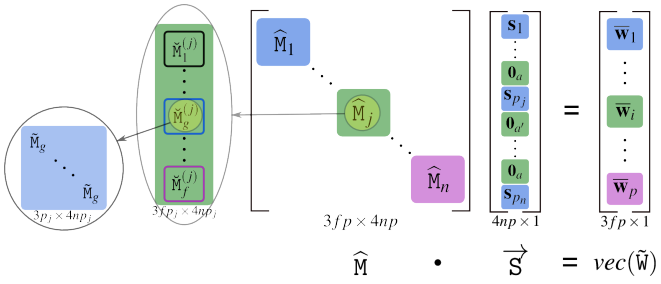

Fig. 1. Structure of matrices in Eq. (20) for p points having n motions in f frames.

Using Eq. (20) and Eq. (23) the algorithm alternates between estimating multiple structures and multiple motions until they converge. The algorithm to identify the best motion segmentation hypothesis is outlined in Alg. (3).

Algorithm 3 Evaluate hypotheses

Input: *Motion segmentation hypothesis*
Output: *Reprojection error for each hypothesis*

1: **for all** motion hypotheses **do**
2: Generate multi-body SfM matrices $\mathtt{W}$, $\mathtt{S}$ and $\mathtt{M}$ ▷ *% with respect to Eq. (12)%*
3: **repeat**
4: Given motion matrix $\mathtt{M}$ Solve Eq. (20) for $\mathtt{S}$
5: Given structure matrix $\mathtt{S}$ Solve Eq. (23) for $\mathtt{M}$
6: **until** Convergence
7: **return** Reprojection error
8: **end for**

IV. EXPERIMENTS

A popular street-level dataset—*KITTI dataset*[2]—is used for the experiments. This dataset was originally created to benchmark VO algorithms [16]. It consists of several sequences collected by a perspective car-mounted camera driving in urban areas. The whole dataset was visually inspected and several sequences containing different types of multiple motions were extracted.

Although the motion model could be constrained, in the experiments, the motion-segments are modeled as a 6-degree of freedom motion. Thus, the 8-point algorithm [12] is used to estimate the unconstrained motions. In this case, at least 8 points are required ($k = 8$), but in the constrained case, fewer points would be enough to identify individual motions.

The case of pedestrians is not studied, because pedestrian's motion cannot be considered as a rigid motion. Furthermore, in most cases, there are not sufficient and stable features on pedestrians to be considered as individual bodies.

[2]http://www.cvlibs.net/datasets/kitti/

The input to our pipeline is the sequence of images. First, feature points are extracted from the images and matched between consecutive frames. In our experiments, SIFT features [17] are used to detect feature points and to track the features among the frames two-way matching scheme is used to reduce the rate of outliers. The feature matches are then passed to the algorithm, which automatically rejects the outliers during motions' hypotheses generation stage. The outputs of the algorithm are the estimated structures and motions.

Fig. (2) to (5) show samples from different scenarios and the results obtained by our algorithm. As the sequences are from a car-mounted camera, in all these figures the camera is moving forward and consequently the static parts of the scene are identified as individual motions. In Fig. (2) the camera equipped car is moving forward and passing another car. So, in addition to the motion belonging to the static parts of the scene, the relative backward motion of the front car is identified as a different motion. Fig. (3) shows the case when a vehicle is moving perpendicularly to the camera motion. Another scenario is presented in Fig. (4), which shows a vehicle coming from the opposite direction while turning left. In this experiment, although there are some false negatives from the car, most of the points are segmented correctly. Fig. (5) shows the case where a car is coming straight from the opposite direction and its corresponding motion is segmented from the camera motion. Table I shows the amount of reprojection error and percentage of misclassified points for the sequences presented in Fig. (2) to (5). The reprojection error for each point on every frame is calculated as $\|\mathbf{w}_{ig} - \tilde{\mathbf{w}}_{ig}\|$, and the classification error is defined as:

$$\text{Classification Error} = 100 \cdot \frac{\text{No. of misclassified points}}{\text{Total No. of points}}.$$

Our algorithm is also tested on the *Checkerboard* sequence from the *Hopkins 155* dataset [18], which consists of 104 sub-sequences.[3] For this sequence, median and mean reprojection errors among all the sub-sequences are 1.03 and 0.77 pixels, respectively. Since the sequences are collected by a hand held camera, in some cases where the camera is almost static, the camera motion is not large enough to be identified as an individual motion-segment by our algorithm. In these cases, the camera motion will be joined with the most similar one in the scene (such as Fig. (6)) or will not be detected at all. In such cases, by ignoring the small camera motion, the mean and median classification errors are 0.35% and 0.23%, but considering camera motion, the classification error raises up to 23.5%. Fig. (6) shows the result of our algorithm on a sample of *Checkerboard* sequence. For this sample, although the algorithm cannot identify the camera movement as an individual motion segment, the other two motions are segmented with zero classification error. In the literature, according to [19], the best performance obtained for the *Checkerboard* sequence has mean classification error 1.24% and median classification error 0.0%.

[3]http://www.vision.jhu.edu/data/hopkins155/

Fig. 2. Forward-Backward: car-mounted camera is moving forward passing another car which is also moving forward (moving backward with respect to the camera). Different colors represent the segmented trajectories.

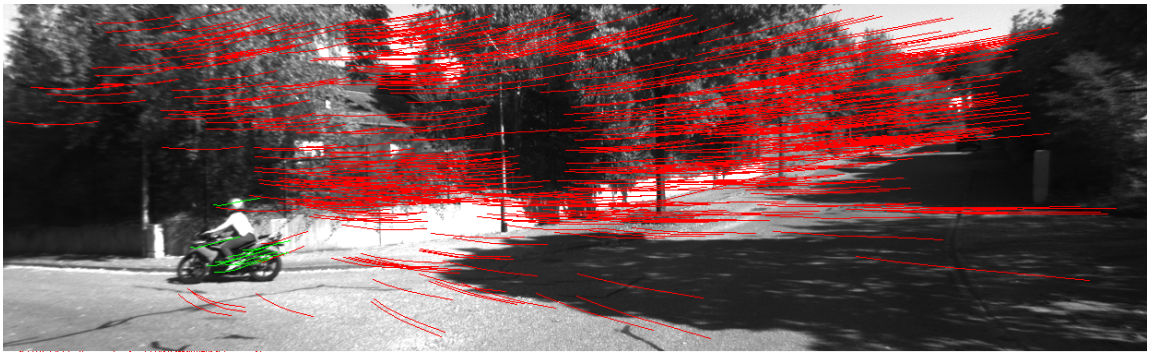

Fig. 3. Forward-Perpendicular: car-mounted camera is moving forward and a motorbike is driving perpendicularly to the car's motion. Different colors represent the segmented trajectories.

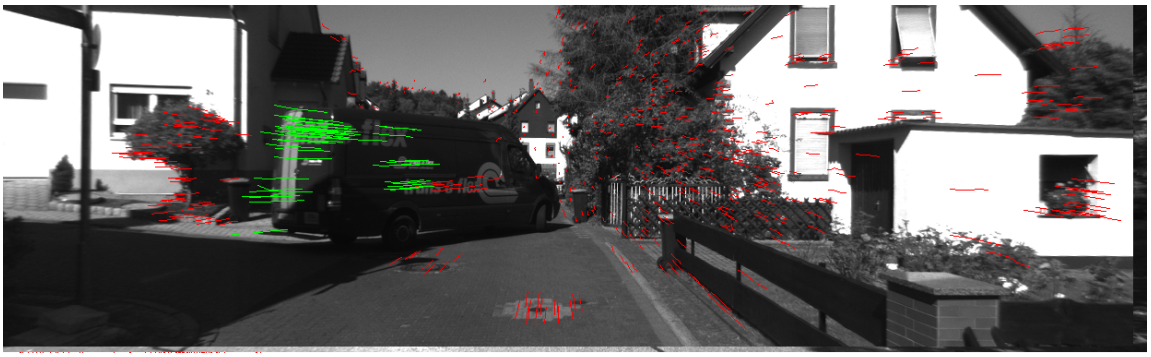

Fig. 4. Forward-Backward Curve: car-mounted camera is moving forward and another car is coming backward from the opposite direction and turning left. Different colors represent the segmented trajectories.

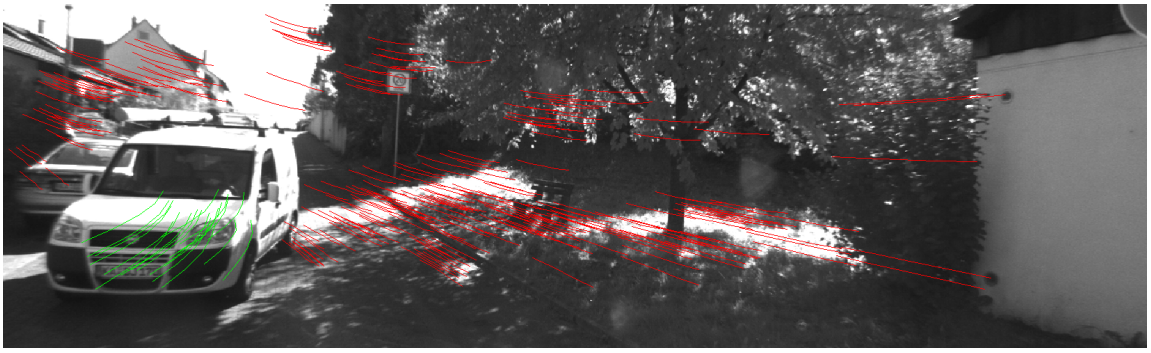

Fig. 5. Forward-Backward: car-mounted camera is moving forward and another car is coming from the opposite direction. Different colors represent the segmented trajectories.

TABLE I

REPROJECTION AND SEGMENTATION ERRORS FOR SEQUENCES FROM *KITTI* DATASET (FIG. (2) TO (5))

Sequence Related to	Number of Frames	Number of Points	Number of Misclassified Points	Mean Reprojection Error (pixels)	Median Reprojection Error (pixels)	Segmentation Error (%)
Fig. (2)	5	193	0	1.63	1.43	0
Fig. (3)	5	608	3	1.69	1.54	0.49
Fig. (4)	5	573	9	2.14	1.67	1.57
Fig. (5)	5	283	0	2.31	2.26	0

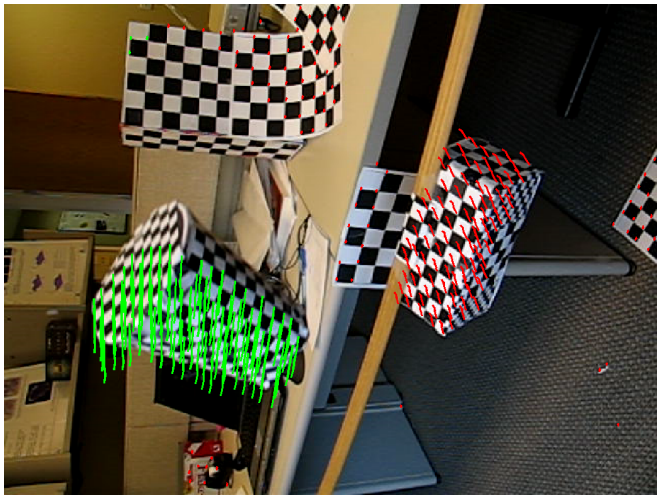

Fig. 6. A sample of *Checkerboard* sequence from *Hopkins 15*: the obtained classification error was zero (ignoring near-constant camera motion); median and mean reprojection error are 0.93 and 1.05 pixels, respectively. The difference between camera motion (points on the up-right checkerboard) and another motion in the scene was not large enough to be distinguished.

In our experiments, parameter σ (in Eq. (15)) is in domain of $[0.05, 0.3]$ and the evaluation of feature point classification is done visually.

V. CONCLUSIONS AND FUTURE PERSPECTIVE

This paper provided a theoretical framework for estimating 3D structures of rigid objects, together with the motion associated to each structure, from perspective images. The motivation of this research was to complement visual navigation algorithms, like VO and SLAM, with the capability of considering both static and dynamic parts of the scene for localization and mapping. The experiments were designed to be compatible with the motivation. Thus, street-level sequences were mainly used for the experiments, although a benchmark dataset was also used to compare the performance of our method with previous works.

Although the current implementation of our algorithm does not run in real-time, this can be improved by an implementation for parallel processing or by using motion constraints. Since generating and evaluating each hypothesis is independent of other hypotheses, these two processes can be highly parallelized. Moreover, using motion constraints (e.g. planar motion), the 6-degree of freedom motion models used in this work can be reduced to fewer degrees of freedom, which results in a considerable boost in timing.

REFERENCES

[1] D. Scaramuzza and F. Fraundorfer, "Visual odometry: Part I: The first 30 years and fundamentals," *IEEE robotics & automation magazine*, vol. 18, no. 4, pp. 80–92, 2011.

[2] Y. Murakami, T. Endo, Y. Ito, and N. Babaguchi, "Depth-estimation-free condition for projective factorization and its application to 3d reconstruction," in *Asian Conference on Computer Vision (ACCV'13)*. Springer, 2013, pp. 150–162.

[3] J. P. Costeira and T. Kanade, "A multibody factorization method for independently moving objects," *International Journal of Computer Vision (IJCV'98)*, vol. 29, no. 3, pp. 159–179, 1998.

[4] J. Yan and M. Pollefeys, "A general framework for motion segmentation: Independent, articulated, rigid, non-rigid, degenerate and non-degenerate," in *European Conference on Computer Vision (ECCV'06)*. Springer, 2006, pp. 94–106.

[5] R. Vidal, Y. Ma, S. Soatto, and S. Sastry, "Two-view multibody structure from motion," *International Journal of Computer Vision (IJCV'06)*, vol. 68, no. 1, pp. 7–25, 2006.

[6] R. Vidal and R. Hartley, "Three-view multibody structure from motion," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'08)*, vol. 30, no. 2, pp. 214–227, 2008.

[7] K. Schindler, U. James, and H. Wang, "Perspective n-view multibody structure-and-motion through model selection," in *European Conference on Computer Vision (ECCV'06)*. Springer, 2006, pp. 606–619.

[8] T. Li, V. Kallem, D. Singaraju, and R. Vidal, "Projective factorization of multiple rigid-body motions," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*. IEEE, 2007, pp. 1–6.

[9] R. Vidal, "Multi-subspace methods for motion segmentation from affine, perspective and central panoramic cameras," in *IEEE International Conference on Robotics and Automation (ICRA'05)*. IEEE, 2005, pp. 1216–1221.

[10] P. Sturm and B. Triggs, "A factorization based algorithm for multi-image projective structure and motion," in *European Conference on Computer Vision (ECCV'96)*. Springer, 1996, pp. 709–720.

[11] C. Tomasi and T. Kanade, "Shape and motion from image streams under orthography: a factorization method," *International Journal of Computer Vision (IJCV'92)*, vol. 9, no. 2, pp. 137–154, 1992.

[12] R. Truesdalc, K. T. M. Micropaloeonr, and J. Fenner, "A computer algorithm for reconstructing a scene from two projections," *Nature*, vol. 293, p. 133, 1981.

[13] J. Oliensis and R. Hartley, "Iterative extensions of the sturm/triggs algorithm: Convergence and nonconvergence," *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI'07)*, vol. 29, no. 12, pp. 2217–2233, 2007.

[14] M. Zuliani, C. S. Kenney, and B. Manjunath, "The multiransac algorithm and its application to detect planar homographies," in *IEEE International Conference on Image Processing (ICIP'05)*, vol. 3. IEEE, 2005, pp. III–153.

[15] Y. Ma, S. Soatto, J. Kosecka, and S. Sastry, *An Invitation to 3-D Vision: From Images to Geometric Models*. springer, 2004.

[16] A. Geiger, P. Lenz, and R. Urtasun, "Are we ready for autonomous driving? the kitti vision benchmark suite," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'12)*. IEEE, 2012, pp. 3354–3361.

[17] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *International journal of computer vision (IJCV'04)*, vol. 60, no. 2, pp. 91–110, 2004.

[18] R. Tron and R. Vidal, "A benchmark for the comparison of 3-d motion segmentation algorithms," in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR'07)*. IEEE, 2007, pp. 1–8.

[19] R. Vidal, "Subspace clustering," *IEEE Signal Processing Magazine*, vol. 28, no. 2, pp. 52–68, 2011.