ISPRS Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences, Volume II-3/W5, 2015
ISPRS Geospatial Week 2015, 28 Sep – 03 Oct 2015, La Grande Motte, France

JOINT 3D ESTIMATION OF VEHICLES AND SCENE FLOW

M. Menze[a], C. Heipke[a], A. Geiger[b]

[a] Institute of Photogrammetry and GeoInformation, Leibniz Universität Hannover, Germany - (menze, heipke)@ipi.uni-hannover.de
[b] Perceiving Systems Department, Max Planck Institute for Intelligent Systems, Tübingen, Germany - andreas.geiger@tue.mpg.de

KEY WORDS: Scene Flow, Motion Estimation, 3D Reconstruction, Active Shape Model, Object Detection

ABSTRACT:

Three-dimensional reconstruction of dynamic scenes is an important prerequisite for applications like mobile robotics or autonomous driving. While much progress has been made in recent years, imaging conditions in natural outdoor environments are still very challenging for current reconstruction and recognition methods. In this paper, we propose a novel unified approach which reasons jointly about 3D scene flow as well as the pose, shape and motion of vehicles in the scene. Towards this goal, we incorporate a deformable CAD model into a slanted-plane conditional random field for scene flow estimation and enforce shape consistency between the rendered 3D models and the parameters of all superpixels in the image. The association of superpixels to objects is established by an index variable which implicitly enables model selection. We evaluate our approach on the challenging KITTI scene flow dataset in terms of object and scene flow estimation. Our results provide a prove of concept and demonstrate the usefulness of our method.

1. INTRODUCTION

3D reconstruction of dynamic scenes is an important building block of many applications in mobile robotics and autonomous driving. In the context of highly dynamic environments, the robust identification and reconstruction of individually moving objects are fundamental tasks as they enable save autonomous navigation of mobile platforms and precise interaction with surrounding objects. In image sequences, motion cues are amongst the most powerful features for separating foreground objects from the background. While approaches for monocular optical flow estimation have matured since the seminal work of (Horn and Schunck, 1980) 35 years ago, they still struggle with real world conditions such as non-lambertian surfaces, variable illumination conditions, untextured surfaces and large displacements. Apart from more sophisticated regularizers, stereo information provides a valuable source of information as it can be used to further constrain the problem. Furthermore, depth information allows for a more meaningful parametrization of the problem in 3D object space. Recent algorithms for scene flow estimation leverage this fact (Vogel et al., 2013, Vogel et al., 2014) and provide promising segmentations of the images into individually moving objects (Menze and Geiger, 2015).

In this paper, we build upon the method of (Menze and Geiger, 2015) but go one step further: Instead of simply decomposing the scene into a set of individually moving regions which share a common rigid motion, we decompose the scene into 3D objects and in addition to the rigid motion also model their pose and shape in 3D. Towards this goal, we incorporate a deformable 3D model of vehicles into the scene flow estimation process. More specifically, we exploit the Eigenspace-based representation of (Zia et al., 2011) which has previously been used in the context of pose estimation from a single image. Given two stereo pairs as input, our model jointly infers the number of vehicles, their shape and pose parameters, as well as a dense 3D scene flow field. The problem is formalized as energy minimization on a conditional random field encouraging projected object hypotheses to agree with the estimated motion and depth. A representative result is shown in Fig. 1 which depicts scene flow estimates projected to disparity and optical flow as well as the result of model-based reconstruction.

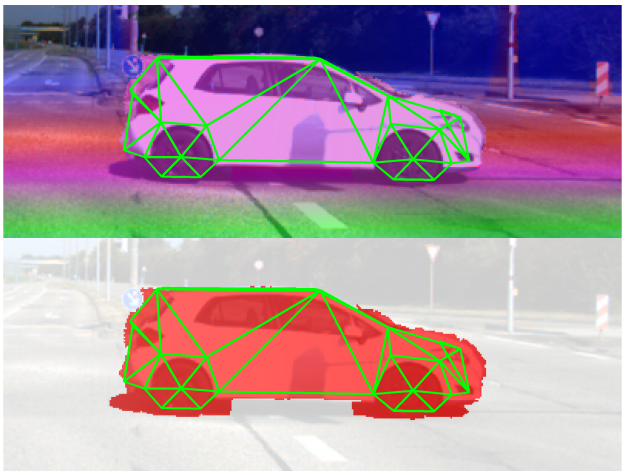


Figure 1. **Result of Model-based Reconstruction.** The green wire-frame representation is superimposed to the inferred disparity (top) and optical flow map (bottom).

The remainder of this paper is structured as follows. We first provide a brief summary of related work in Section 2. and a detailed formal description of the proposed method in Section 3. In Section 4. we present results for dynamic scenes on the novel KITTI scene flow dataset proposed by (Menze and Geiger, 2015). We conclude the paper in Section 5.

2. RELATED WORK

In this section, we provide a brief overview over the state-of-the-art in scene flow estimation as well as related work on integrating 3D models into reconstruction.

Scene flow estimation has first been addressed by (Vedula et al., 1999, Vedula et al., 2005) who define scene flow as a flow field describing the 3D motion at every point in the scene. Like in classical optical flow estimation (Horn and Schunck, 1980), the problem is often formulated in a coarse-to-fine variational setting (Basha et al., 2013, Huguet and Devernay, 2007, Pons et al., 2007, Valgaerts et al., 2010, Wedel et al., 2011, Vogel et al., 2011)

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
Editors: A. Yilmaz, C. Mallet, M. Ying Yang, and Y. Vizilter
doi:10.5194/isprsannals-II-3-W5-427-2015

and local regularizers are leveraged to encourage smoothness in depth and motion. As in optical flow estimation, this approach eventually fails to recover large displacements of small objects. Following recent developments in optical flow (Yamaguchi et al., 2013, Nir et al., 2008, Wulff and Black, 2014, Sun et al., 2013) and stereo (Yamaguchi et al., 2014, Bleyer et al., 2011, Bleyer et al., 2012), Vogel et al. (Vogel et al., 2013, Vogel et al., 2014) proposed a slanted-plane model which assigns each pixel to an image segment and each segment to one of several rigidly moving 3D plane proposals, thus casting the task as a discrete optimization problem. Fusion moves are leveraged for solving binary subproblems with quadratic pseudo-boolean optimization (QPBO) (Rother et al., 2007). Their approach yields promising results on challenging outdoor scenes as provided by the KITTI stereo and optical flow benchmarks (Geiger et al., 2012). More recently, (Menze and Geiger, 2015) noticed that many structures in the visual world move rigidly and thus decompose the scene into a small number of rigidly moving objects and the background. They jointly estimate the segmentation as well as the motion of the objects and the 3D geometry of the scene. In addition to segmenting the objects according to their motion (which doesn't guarantee instances to be separated), in this paper, we propose to also estimate their shape and pose parameters. Thus, we infer a parametrized reconstruction of all moving vehicles in the 3D scene jointly with the 3D scene flow itself.

3D Models have a long history in supporting 3D reconstruction from images (Szeliski, 2011). Pioneering work, e.g. by (Debevec et al., 1996) made use of shape primitives to support photogrammetric modelling of buildings. While modelling generic objects, like buildings, is a very challenging task, there are tractable approaches to formalizing the geometry of objects with moderate intra-class variability, like faces and cars. A notable example is the active shape model (ASM) proposed by (Cootes et al., 1995) where principal component analysis of a set of annotated training examples yields the most important deformations between similar shapes. (Bao et al., 2013) compute a mean shape of the observed object class along with a set of discrete anchor points. Using HOG features, they adapt the mean shape to a newly observed instance of the object by registering the anchor points. (Güney and Geiger, 2015) leverage semantic information to sample CAD shapes with an application to binocular stereo matching. (Dame et al., 2013) use an object detector to infer the initial pose and shape parameters for an object model which they then optimize in a variational SLAM framework. Recently, (Prisacariu et al., 2013) proposed an efficient way to compress prior information from CAD models with complex shape variations using Gaussian Process Latent Variable Models. (Zia et al., 2013a, Zia et al., 2013b, Zia et al., 2015) revisited the idea of the ASM and applied it to a set of manually annotated CAD models to derive detailed 3D geometric object class representations. While they tackle the problem of object recognition and pose estimation from single images, in this paper, we make use of such models in the context of 3D scene flow estimation.

3. METHOD

Our aim is to jointly estimate optimal scene flow parameters for each pixel in a reference image and a parametrized reconstruction of individually moving vehicles as shown in Fig. 2. The proposed algorithm works on the classical scene flow input consisting of two consecutive stereo image pairs of calibrated cameras. We define the first image from the left camera as the reference view. Following the state-of-the-art, we approximate 3D scene geometry with a set of planar segments which are derived from superpixels in the reference view (Yamaguchi et al., 2013). Like

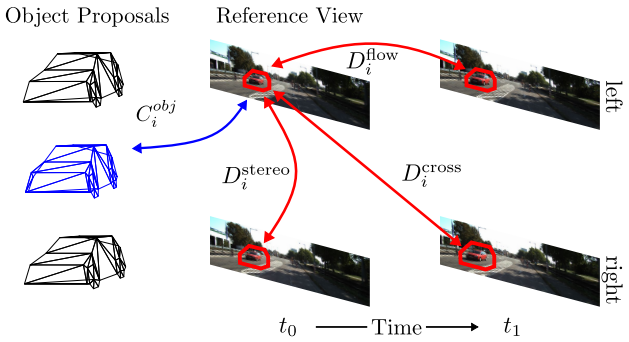


Figure 2. **Data and Shape Terms.** Each superpixel $\mathbf{s}_i$ in the reference view is matched to corresponding image patches in the three remaining views. Its shape and motion are encouraged to agree with the jointly estimated 3D object model.

(Menze and Geiger, 2015), we assume a finite number of rigidly moving objects in the scene. It is important to note that using this formulation, the background can be considered as yet another object. The only difference is that we do not estimate a 3D model for the background component.

In this section, we first give a formal definition of our model and the constituting energy terms for data, shape and smoothness. Then, the employed active shape model and the inference algorithm are explained in detail.

3.1 Problem statement

Let $\mathcal{S}$ and $\mathcal{O}$ denote the set of superpixels and objects, respectively. Each superpixel $\mathbf{s}_i \in \mathcal{S}$ is associated with a region $\mathcal{R}_i$ in the image and a random variable $(\mathbf{n}_i, l_i)^T$ where $\mathbf{n}_i \in \mathbb{R}^3$ describes a plane in 3D ($\mathbf{n}_i^T \mathbf{x} = 1$ for points $\mathbf{x} \in \mathbb{R}^3$ on the plane) and $l_i \in \{1, \dots, |\mathcal{O}|\}$ is a label assigning the superpixel to an object. Each object $\mathbf{o}_k \in \mathcal{O}$ is associated with a random variable $(\xi_k, \gamma_k, \mathbf{R}_k, \mathbf{t}_k)^T$ comprising its state. $\xi_k \in \mathbb{R}^3$ determines the pose, i.e. the position (2D coordinates in the ground plane) and the orientation of the object in terms of its heading angle. $\gamma_k \in \mathbb{R}^2$ contains the parameters determining the shape of the 3D model. $\mathbf{R}_k \in SO(3)$ and $\mathbf{t}_k \in \mathbb{R}^3$ describe the rigid body motion of object $\mathbf{o}_k$ in 3D, i.e. the rotation and translation relating the poses of the object at subsequent time steps. Each superpixel $\mathbf{s}_i$ is associated with an object via l_i. Thus, the superpixel inherits the rigid motion parameters of the respective object $(\mathbf{R}_{l_i}, \mathbf{t}_{l_i}) \in SE(3)$. In combination with the plane parameters $\mathbf{n}_i$, this fully determines the 3D scene flow at each pixel inside the superpixel.

Given the left and right input images of two consecutive stereo frames at t_0 and t_1, our goal is to infer the 3D geometry, i.e. the plane parameters $\mathbf{n}_i$ of each superpixel and its object label l_i together with the rigid body motion, the pose and the shape parameters of each object. We specify our model as a conditional random field (CRF) in terms of the following energy function

$$E(\mathbf{s}, \mathbf{o}) = \sum_{i \in \mathcal{S}} [\underbrace{\varphi_i(\mathbf{s}_i, \mathbf{o})}_{\text{data}} + \underbrace{\kappa_i(\mathbf{s}_i, \mathbf{o})}_{\text{shape}}] + \sum_{i \sim j} \underbrace{\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j)}_{\text{smoothness}} \quad (1)$$

where $\mathbf{s} = \{\mathbf{s}_i | i \in \mathcal{S}\}$, $\mathbf{o} = \{\mathbf{o}_k | k \in \mathcal{O}\}$, and $i \sim j$ denotes the set of adjacent superpixels in $\mathcal{S}$. We use the same data term $\varphi(\cdot)$ and the same smoothness term $\psi(\cdot)$ as proposed in (Menze and Geiger, 2015), and add an additional shape term $\kappa(\cdot)$ to model the pose and shape of the objects in 3D. To make the paper self-contained, we will briefly review the data term before we provide the formal description of the novel shape term.

3.2 Data Term

Data fidelity of corresponding image points is enforced with respect to all four input images in a combined data term depending on shape and motion. Since both entities are encoded in different random variables, the data term is defined as a pairwise potential between superpixels and objects

$$\varphi_i(\mathbf{s}_i, \mathbf{o}) = \sum_{k \in \mathcal{O}} [l_i = k] \cdot D_i(\mathbf{n}_i, \mathbf{o}_k) \tag{2}$$

where l_i assigns superpixel i to a specific object and $[\cdot]$ denotes the Iverson bracket, which returns 1 if the condition in square brackets is satisfied and 0 otherwise. Thus, the actual data term $D_i(\mathbf{n}, \mathbf{o})$ is only evaluated with respect to the selected object. It comprises three components: A stereo, an optical flow and a cross term which relate the reference view (left image at t_0) to the three remaining images, as depicted in Fig. 2:

$$D_i(\mathbf{n}, \mathbf{o}) = D_i^{\text{stereo}}(\mathbf{n}, \mathbf{o}) + D_i^{\text{flow}}(\mathbf{n}, \mathbf{o}) + D_i^{\text{cross}}(\mathbf{n}, \mathbf{o})$$

Note that this term depends on the plane parameters $\mathbf{n}$ of the superpixel and the rigid motion parameters of the object $\mathbf{o}$. Each sub-term sums matching costs C of all pixels $\mathbf{p}$ inside the region $\mathcal{R}$ of superpixel i. As we assume that the geometry within a superpixel can be approximated by a local plane, we are able to warp pixels from the reference view to the other images using homographies computed from $\mathbf{n}$ and $\mathbf{o}$:

$$D_i^{\text{x}}(\mathbf{n}, \mathbf{o}) = \sum_{\mathbf{p} \in \mathcal{R}_i} C_{\text{x}}\big(\mathbf{p}, \underbrace{\mathbf{K}\left(\mathbf{R}_{\text{x}}(\mathbf{o}) - \mathbf{t}_{\text{x}}(\mathbf{o}) \cdot \mathbf{n}^T\right)\mathbf{K}^{-1}}_{3\times 3 \text{ homography}} \mathbf{p}\big)$$

The superscript of D indicates which image is compared to the reference view, with $\text{x} \in \{\text{stereo}, \text{flow}, \text{cross}\}$. Without loss of generality, the camera calibration matrix $\mathbf{K} \in \mathbb{R}^{3\times 3}$ is assumed to be the same for both cameras. The matching cost $C_{\text{x}}(\mathbf{p}, \mathbf{q})$ is a dissimilarity measure between a pixel at location $\mathbf{p} \in \mathbb{R}^2$ in the reference image and a pixel at location $\mathbf{q} \in \mathbb{R}^2$ in the target image.

In this work, we evaluate two types of features and define $C_{\text{x}}(\mathbf{p}, \mathbf{q})$ as the weighted sum of matching costs based on dense Census features (Zabih and Woodfill, 1994) and sparse disparity and optical flow observations:

$$C_{\text{x}}(\mathbf{p}, \mathbf{q}) = \theta_{1,\text{x}}\, C_{\text{x}}^{\text{dense}}(\mathbf{p}, \mathbf{q}) + \theta_{2,\text{x}}\, C_{\text{x}}^{\text{sparse}}(\mathbf{p}, \mathbf{q})$$

The dense matching cost is computed as the truncated Hamming distance between Census features. Pixels leaving the target image are penalized with a truncation value. As precomputed disparity estimates (Hirschmüller, 2008) and optical flow features (Geiger et al., 2011) are not available for every pixel, we calculate $C_{\text{x}}^{\text{sparse}}$ only at locations for which observations exist. More specifically, we define $C_{\text{x}}^{\text{sparse}}$ as the robust l_2 distance between the warped pixel $\pi_{\text{x}}(\mathbf{p})$ and the expected pixel $\mathbf{q}$

$$C_{\text{x}}^{\text{sparse}}(\mathbf{p}, \mathbf{q}) = \begin{cases} \rho_{\tau_1}\left(\|\pi_{\text{x}}(\mathbf{p}) - \mathbf{q}\|_2\right) & \text{if } \mathbf{p} \in \Pi_{\text{x}} \\ 0 & \text{otherwise} \end{cases}$$

where $\rho_{\tau_i}(x)$ denotes the robust truncated penalty function $\rho_{\tau_i}(x) = \min(|x|, \tau_i)$ with threshold τ_i and $\pi_{\text{x}}(\mathbf{p})$ denotes the pixel $\mathbf{p}$, warped according to the set of sparse feature correspondences. Π_{x} is the set of pixels in the reference image for which correspondences have been established. For more details, we refer the reader to (Menze and Geiger, 2015).

3.3 Shape and Pose Consistency Term

Our novel shape consistency term enforces consistency between the 3D plane of superpixel $\mathbf{s}_i$ and the pose and shape of the referenced object. Similarly to the data term, we can take advantage of the fact that this term decomposes into computationally tractable pairwise potentials between superpixels and objects:

$$\begin{aligned}\kappa_i(\mathbf{s}_i, \mathbf{o}) = \sum_{k \in \mathcal{O}} \big([l_i = k] \cdot S_i(\mathbf{n}_i, \mathbf{o}_k) \\ + [l_i \neq k \wedge k > 1] \cdot O_{ik}(\mathbf{o}_k)\big)\end{aligned} \tag{3}$$

Here, $S_i(\mathbf{n}_i, \mathbf{o}_k)$ enforces consistency between the shape of object $\mathbf{o}_k$ and the 3D plane described by $\mathbf{n}_i$. In analogy with the data term, shape consistency is evaluated with respect to the object associated with the superpixel via l_i. We define the penalty function S_i as

$$S_i(\mathbf{n}, \mathbf{o}) = \begin{cases} C^{\text{bg}} & \text{if } \mathbf{o} \text{ is background} \\ C_i^{\text{obj}}(\mathbf{n}, \mathbf{o}) & \text{otherwise} \end{cases}$$

where C^{bg} denotes a constant penalty for superpixels associated with the background, and $C_i^{\text{obj}}(\mathbf{n}, \mathbf{o})$ denotes the sum of the truncated absolute differences between the 3D model of object $\mathbf{o}_k$ projected to a disparity map (see Section 3.5) and the disparities induced by the 3D plane $\mathbf{n}_i$. Differences are computed for all pixels inside $\mathcal{R}_i$ which coincide with the projection of $\mathbf{o}_k$. Remaining, uncovered pixels are penalized with a multiple of C^{bg}. Note that in contrast to the data term D_i this term evaluates the consistency between the deformed shape model and the reconstructed superpixels.

The second part of Eq. 3 is the occlusion penalty O_{ik}. It penalizes a possible overlap between parts of a foreground model and superpixels that are assigned to a different object via the arguments of the leading Iverson bracket. The overlap penalty itself is chosen to be proportional to the overlap of the projected model of object $\mathbf{o}_k$ with the superpixel $\mathbf{s}_i$. This term is crucial to avoid object models from exceeding the true object boundaries.

3.4 Smoothness Term

To encourage smooth surface shape and orientation as well as compact objects, the following smoothness potential is defined on the CRF:

$$\begin{aligned}\psi_{ij}(\mathbf{s}_i, \mathbf{s}_j) \quad = \quad & \theta_3\, \psi_{ij}^{\text{depth}}(\mathbf{n}_i, \mathbf{n}_j) + \\ & \theta_4\, \psi_{ij}^{\text{orient}}(\mathbf{n}_i, \mathbf{n}_j) + \theta_5\, \psi_{ij}^{\text{motion}}(\mathbf{s}_i, \mathbf{s}_j)\end{aligned} \tag{4}$$

The weights θ control the influence of the three constituting terms. First, regularization of depth is achieved by penalizing different disparity values d at shared boundary pixels $\mathcal{B}_{ij}$:

$$\psi_{ij}^{\text{depth}}(\mathbf{n}_i, \mathbf{n}_j) = \sum_{\mathbf{p} \in \mathcal{B}_{ij}} \rho_{\tau_2}\left(d(\mathbf{n}_i, \mathbf{p}) - d(\mathbf{n}_j, \mathbf{p})\right)$$

Second, the orientation of neighboring planes is encouraged to be similar by evaluating the difference of plane normals $\mathbf{n}$

$$\psi_{ij}^{\text{orient}}(\mathbf{n}_i, \mathbf{n}_j) = \rho_{\tau_3}\left(1 - |\mathbf{n}_i^T \mathbf{n}_j| / (\|\mathbf{n}_i\| \|\mathbf{n}_j\|)\right)$$

Finally, coherence of the assigned object indices is enforced by an orientation-sensitive Potts model:

$$\psi_{ij}^{\text{motion}}(\mathbf{s}_i, \mathbf{s}_j) = w(\mathbf{n}_i, \mathbf{n}_j) \cdot [l_i \neq l_j]$$

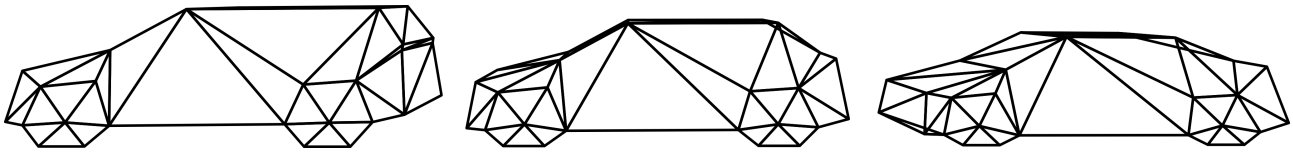


Figure 3. **3D Object Model.** Mean shape (center, $\gamma = (0, 0)$) and two instances illustrating the range of possible deformations with shape parameters $\gamma_{\text{left}} = (-1.0, -0.8)$ and $\gamma_{\text{right}} = (1.0, 0.8)$.

The weight $w(\cdot, \cdot)$ in the coherence term is defined as

$$w(\mathbf{n}_i, \mathbf{n}_j) = \exp\left(-\frac{\lambda}{|\mathcal{B}_{ij}|}\sum_{\mathbf{p}\in\mathcal{B}_{ij}}(d(\mathbf{n}_i, \mathbf{p}) - d(\mathbf{n}_j, \mathbf{p}))^2\right) \times |\mathbf{n}_i^T \mathbf{n}_j|/(\|\mathbf{n}_i\|\|\mathbf{n}_j\|)$$

and prefers motion boundaries that coincide with folds in 3D. Here, λ is the shape parameter of the penalty function which is normalized by the number of shared boundary pixels $|\mathcal{B}_{ij}|$.

3.5 3D Object Model

For encoding prior knowledge about the objects $\{\mathbf{o}_k | k \in \mathcal{O}\}$ and in order to restrict the high-dimensional space of possible shapes, we follow (Zia et al., 2013b) and use their 3D active shape model. In particular, we apply principal component analysis to a set of characteristic keypoints on manually annotated 3D CAD models. This results in a mean model over vertices as well as the directions of the most dominant deformations between the samples in the training set. In our CRF, the shape parameters γ_k of object $\mathbf{o}_k$ are optimized for consistency with the jointly estimated superpixels. The deformed vertex positions $\mathbf{v}$ are specified by a linear sub-space model

$$\mathbf{v}(\gamma_k) = \mathbf{m} + \sum_{i=\{1,2\}} \gamma_k^{(i)} \mathbf{e}_i \tag{5}$$

where $\mathbf{m}$ is the vertex mean and $\mathbf{e}_i$ denotes the i'th eigenvector weighted by the standard deviation of the corresponding eigenvalue. We define a triangular mesh for the vertices $\mathbf{v}(\gamma_k)$, transform it according to the object pose ξ_k and render a virtual disparity map[1] for the reference image in order to calculate the shape consistency term in Section 3.3.

Fig. 3 depicts the mean shape in the center and deformed versions of the model on the left and right, illustrating the range of different layouts covered by the first two principal components. While the first principal component accounts mostly for the size of the object, the second component determines its general shape. We limit our model to the first two principal components as we found this to be an appropriate tradeoff between model complexity and the quality of the approximation.

3.6 Inference

Due to the inherent combinatorial complexity and the mixed discrete-continuous variables, optimizing the CRF specified in Eq. 1 with respect to all superpixels and objects is an NP-hard problem. To minimize the energy, we iteratively and adaptively discretize the domains of the continuous variables in the outer loop of a max-product particle belief propagation (MP-PBP) framework (Trinh and McAllester, 2009, Pacheco et al., 2014). In the inner loop, we employ sequential tree-reweighted message passing (TRW-S) (Kolmogorov, 2006) to infer an approximate solution given the current set of particles.

To keep the computational burden tractable, we perform informed sampling of pose and shape parameters. In each iteration of the outer loop, we draw 50 particles, jointly sampling pose and shape from normal distributions centered at the preceding MAP solution. The respective standard deviations are iteratively reduced. To prune the proposals, the shape consistency term, Eq. 3, is evaluated for each particle with respect to the superpixels' MAP solution of the previous iteration. Only the best particle is kept and introduced into the optimization of Eq. 1.

In our implementation, we further use 10 shape particles for each superpixel, 5 particles for object motion, and 10 iterations of MP-PBP. All motion particles and half of the superpixel plane particles are drawn from a normal distribution centered at the MAP solution of the last iteration. The remaining plane particles are proposed using the plane parameters from spatially neighboring superpixels.

4. EXPERIMENTAL RESULTS

To demonstrate the value of our approach, we process challenging scenes from the scene flow dataset proposed by (Menze and Geiger, 2015). As we evaluate additional metrics regarding the quality of the estimated objects we use a set of representative training images for which ground truth information is publicly available. The observations evaluated in the data term comprise densely computed differences of Census features and additional sparse features. We use optical flow from feature point correspondences (Geiger et al., 2011) and precomputed disparity maps using semiglobal matching (SGM) (Hirschmüller, 2008). Sparse cross features, connecting the reference view with the right image at t_1, are computed by combining the optical flow matches with valid disparities from the SGM maps. We initialize all superpixel boundaries and their shape parameters using the StereoSLIC algorithm (Yamaguchi et al., 2013) with a parameter setting that yields approximately 1000 superpixels for the used input images. One typical oversegmentation of a car is depicted in Fig. 4. While most of the outline is faithfully recovered, shadows can lead to bleeding artifacts.

Figure 4. **Superpixels.** Typical oversegmentation of a car.

[1] `http://www.cvlibs.net/software/librender/`

	D1			D2			Fl			SF		
	bg	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*	*bg*	*fg*	*bg&fg*
(Menze and Geiger, 2015)	**4.71**	4.79	**4.72**	**5.44**	10.69	**6.14**	**5.95**	24.16	8.37	**7.39**	24.68	9.68
Ours	4.94	**4.24**	4.84	5.68	**9.32**	6.16	6.21	**19.70**	**8.00**	7.53	**19.99**	**9.18**

Table 1. **Scene Flow Error.** This table shows the benefits of integrating the proposed object model, evaluated for all results shown in the paper. We specify the percentage of outliers with respect to disparity estimates in the subsequent stereo pairs (D1,D2), optical flow in the reference frame (Fl) and the complete scene flow vectors (SF). See text for details.

Rigid body motions are initialized by greedily extracting motion estimates from sparse scene flow vectors (Geiger et al., 2011) as follows: We iteratively estimate rigid body motions using the 3-point RANSAC algorithm on clusters of similar motion vectors and chose promising subsets with a large number of inliers using non-maxima suppression. The mean positions and the moving direction of the best hypotheses are used as initial values for the object pose parameters ξ. This leads to spurious object hypotheses, as evidenced by Fig. 5, which are pruned during inference because no superpixels are assigned to them. In our experiments, γ comprises two shape parameters controlling the two most significant principal components of the ASM. We initialize each object with the mean shape of the model by setting its shape parameters γ to zero. To compute the shape consistency term in Eq. 3, we use OpenGL to render all object proposals and compare the resulting disparity maps to those induced by the shape particles of each superpixel. In our non-optimized implementation, inference takes more than one minute on a single core, thus the method is not yet applicable to scenarios with real-time constraints.

Qualitative Results: Fig. 5 and Fig. 6 illustrate resulting disparity, optical flow and wire-frame renderings of the object models superimposed to the respective reference views of eight representative scenes. The top part of each sub-figure depicts the layout after initialization as described above. In most cases, the shapes do not match the observed cars and there are some significant positional offsets. In addition, there are spurious objects initialized due to wrong object hypotheses. The lower part shows our reconstruction results after optimizing Eq. 1. Objects which are not referred to by any of the superpixels are considered absent and thus not drawn. For all examples shown in Fig. 5, the model position is successfully aligned with the observed object and the shape of the model is faithfully adapted to the depicted cars. Spurious hypotheses are removed, demonstrating the intrinsic model selection capabilities of our approach. Sub-figures (b,c) of Fig. 6 contain successfully reconstructed cars in the foreground. Some of the spurious objects are removed while others remain in the final result. This is due to strong erroneous motion cues in the respective image patches contradicting the estimated background motion. Note that for visualization we only render fully visible faces of the CAD models. The last sub-figure (d) of Fig. 6 shows a failure case of the approach: Here, object hypotheses with many inliers occur in the very challenging regions next to the road. The numbers in the sub-captions specify the intersection-over-union (IOU) of the estimated object shape with respect to ground truth at initialization and after optimization as explained in the next paragraph.

Shape Adaption: To quantify the improvement gained by optimizing the model parameters, we evaluate the intersection-over-union criterion which is frequently used for evaluating segmentation and object detection in the literature. In particular, we compare the ground truth mask of the annotated objects to the mask of the projected 3D model as inferred by our method. We discard objects without successful initialization and report the intersection-over-union averaged over all detected cars. Table 2 compares the results after initialization to our final results. Although one car which has been correctly initialized is removed during optimization (cf. Fig. 5(b)), the averaged result is significantly improved.

Initialization	0.54
After optimization	**0.67**

Table 2. **Model Coverage.** Intersection-over-union (IOU), averaged over all foreground objects before and after optimization.

Scene Flow Error: The quantitative effect of incorporating the 3D object model is shown in Table 1 which specifies the mean percentage of outliers for all eight examples shown in Fig. 5 and Fig. 6 using the evaluation metrics proposed in (Menze and Geiger, 2015), i.e., a pixel is considered as outlier if the estimated disparity (D1,D2) or optical flow (Fl) exceeds 3 pixels as well as 5% of its true value. As a baseline, we optimize Eq. 1 without the shape consistency term κ and sample only motion particles for the objects instead, corresponding to the method of (Menze and Geiger, 2015). In contrast, our full model ("Ours") also optimizes shape and pose parameters of the 3D model as described in Section 3. Table 1 shows that the performance for background regions (bg) slightly decreases in all categories while there is a significant improvement of 5 percentage points for the foreground objects (fg) and moderately improved results for the combined scene flow metric (bg&fg).

5. CONCLUSIONS

We extended the scene flow algorithm of (Menze and Geiger, 2015) by a deformable 3D object model to jointly recover the 3D scene flow as well as the 3D geometry of all vehicles in the scene. Our results show that the estimation of only 5 model parameters yields accurate parametric reconstructions for a range of different cars. In the future, we plan to incorporate additional observations of a class-specific object detector as well as to estimate motion over multiple frames in order to improve completeness of the retained objects and to further increase robustness against spurious outliers.

REFERENCES

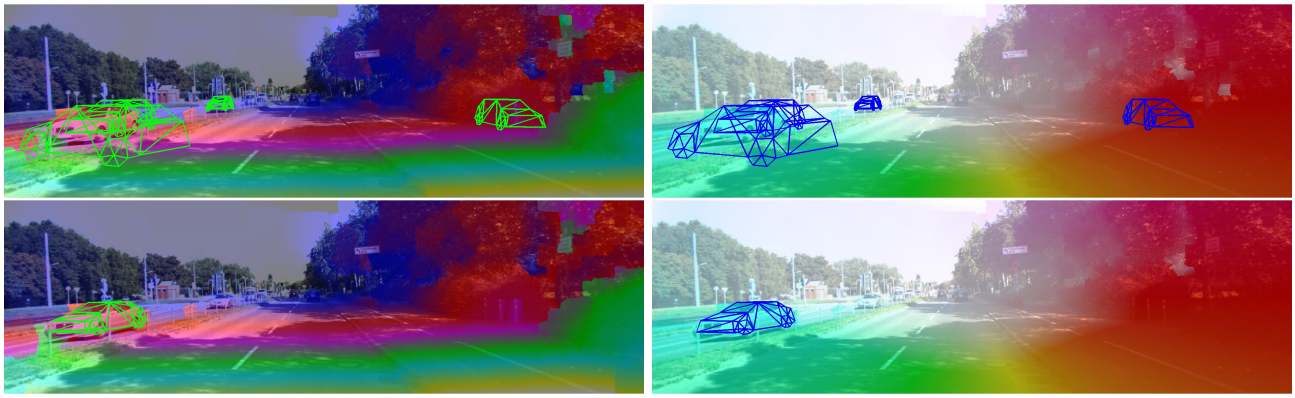
Bao, S., Chandraker, M., Lin, Y. and Savarese, S., 2013. Dense object reconstruction with semantic priors. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1264–1271.

Basha, T., Moses, Y. and Kiryati, N., 2013. Multi-view scene flow estimation: A view centered variational approach. International Journal of Computer Vision (IJCV) 101(1), pp. 6–21.

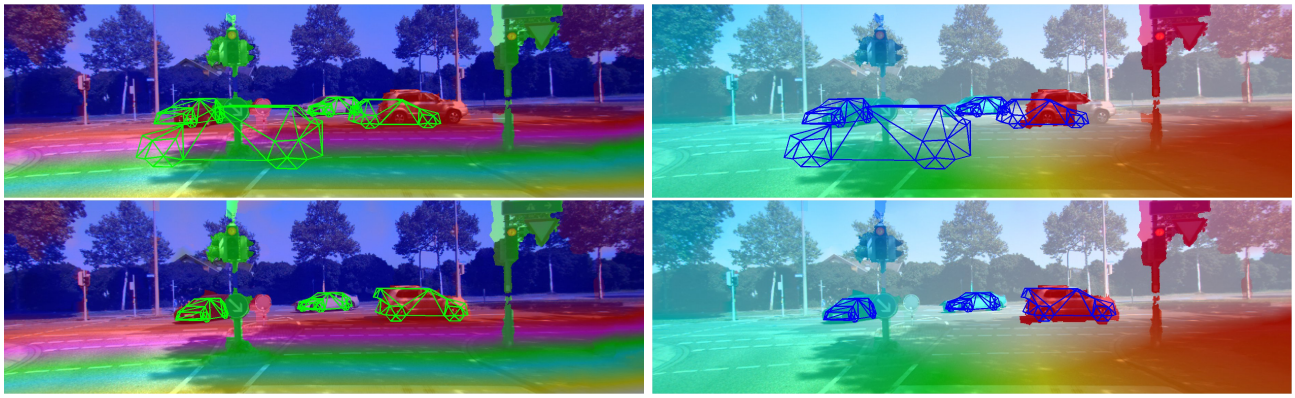
Bleyer, M., Rhemann, C. and Rother, C., 2012. Extracting 3D scene-consistent object proposals and depth from stereo images. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 467–481.

Bleyer, M., Rother, C., Kohli, P., Scharstein, D. and Sinha, S., 2011. Object stereo - joint stereo matching and object segmentation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3081–3088.

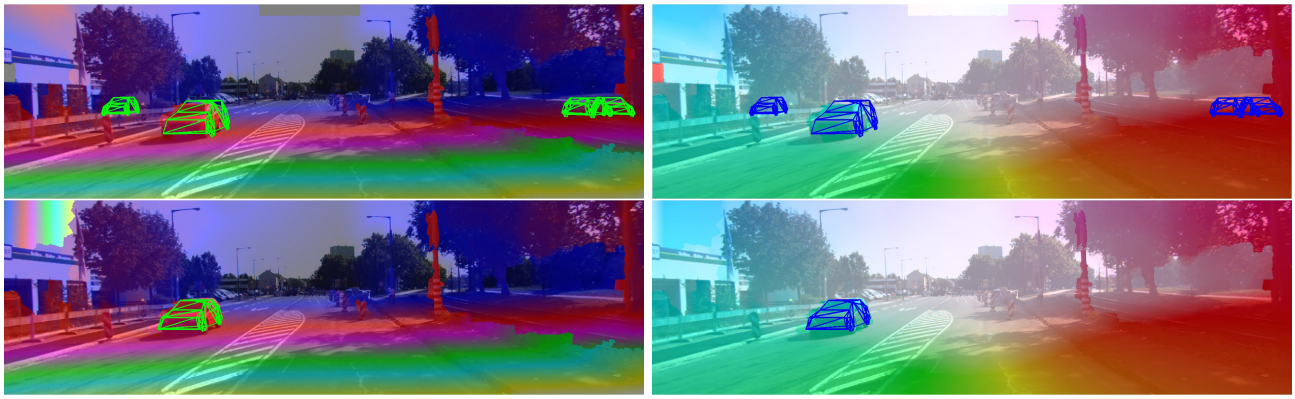
(a) 0.73 (init.) / 0.73 (opt.)



(b) 0.30 (init.) / 0.26 (opt.)



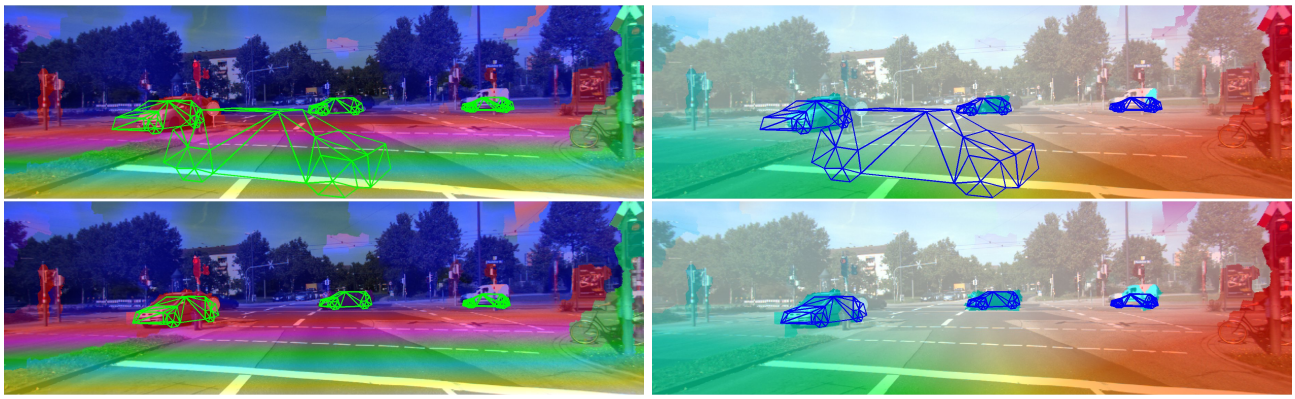
(c) 0.48 (init.) / 0.73 (opt.)



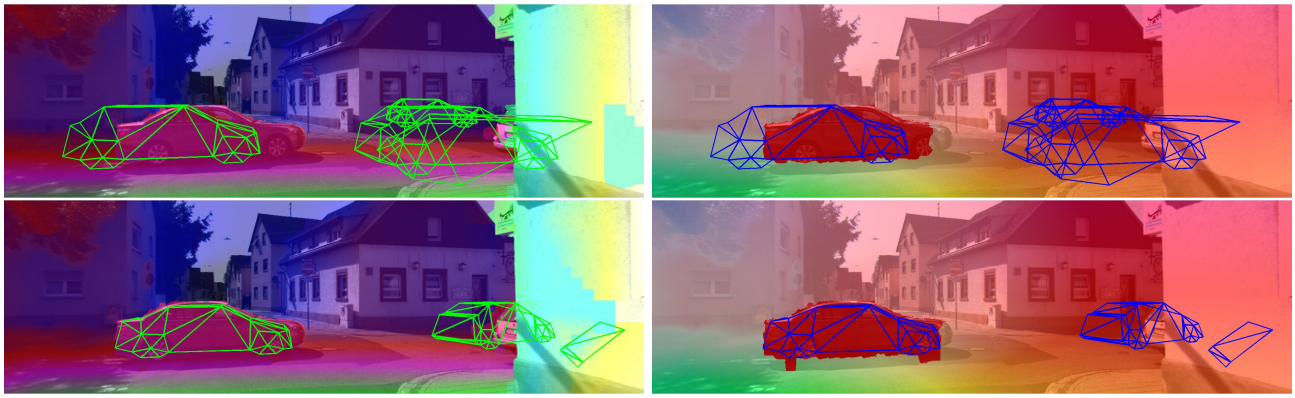
(d) 0.76 (init.) / 0.84 (opt.)

Figure 5. **Qualitative Results.** Each sub-figure shows our results at initialization (top) and after optimization (bottom). The reference view is superimposed with the color-coded disparity (left) and optical flow map (right). Object models are depicted as green and blue wire-frames. The numbers in the sub-captions specify the value of the intersection-over-union criterion at initialization and after optimization.

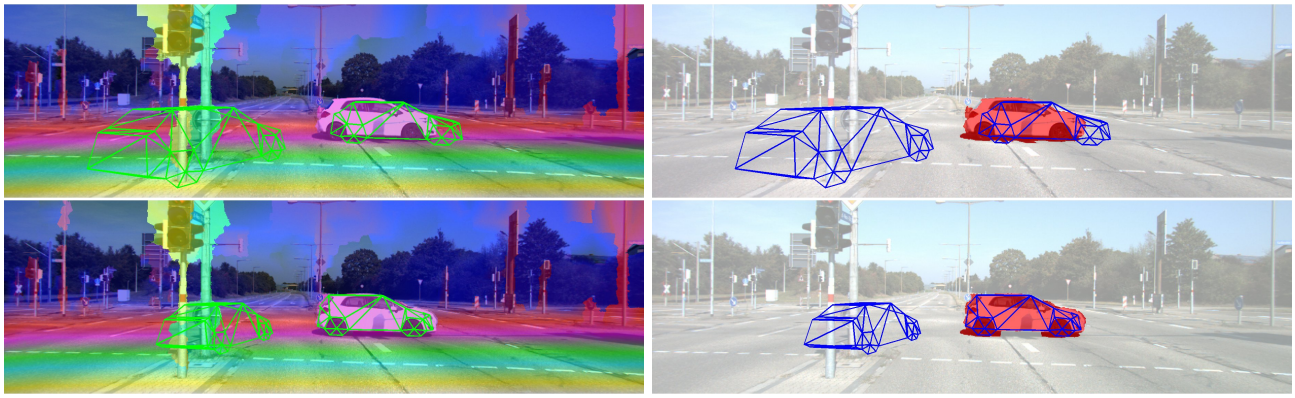
This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
Editors: A. Yilmaz, C. Mallet, M. Ying Yang, and Y. Vizilter
doi:10.5194/isprsannals-II-3-W5-427-2015



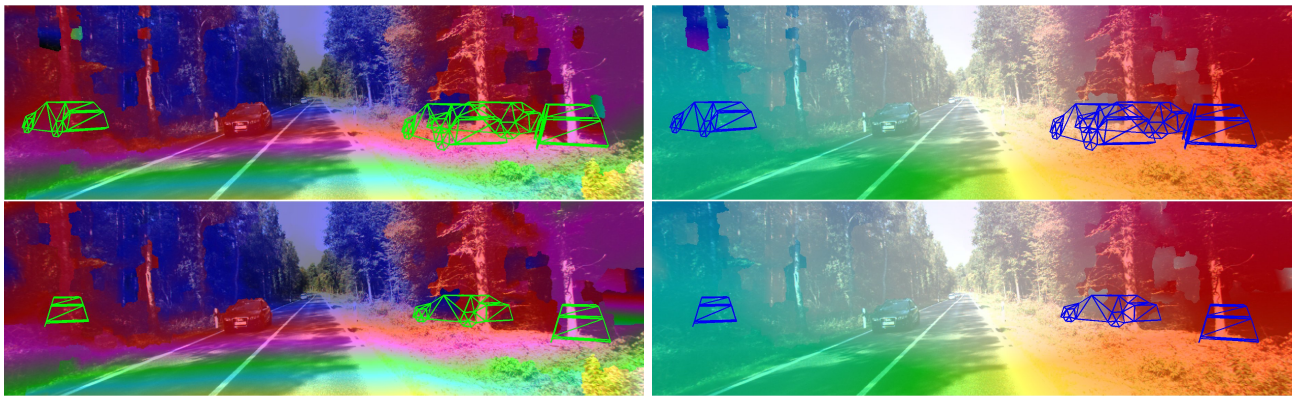
(a) 0.44 (init.) / 0.56 (opt.)



(b) 0.45 (init.) / 0.72 (opt.)



(c) 0.60 (init.) / 0.87 (opt.)



(d) 0.00 (init.) / 0.00 (opt.)

Figure 6. **Qualitative Results.** Each sub-figure shows our results at initialization (top) and after optimization (bottom). The reference view is superimposed with the color-coded disparity (left) and optical flow map (right). Object models are depicted as green and blue wire-frames. The numbers in the sub-captions specify the value of the intersection-over-union criterion at initialization and after optimization.

This contribution has been peer-reviewed. The double-blind peer-review was conducted on the basis of the full paper.
Editors: A. Yilmaz, C. Mallet, M. Ying Yang, and Y. Vizilter
doi:10.5194/isprsannals-II-3-W5-427-2015

Cootes, T. F., Taylor, C. J., Cooper, D. H. and Graham, J., 1995. Active shape models-their training and application. Computer Vision and Image Understanding (CVIU) 61(1), pp. 38–59.

Dame, A., Prisacariu, V., Ren, C. and Reid, I., 2013. Dense reconstruction using 3D object shape priors. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1288–1295.

Debevec, P. E., Taylor, C. J. and Malik, J., 1996. Modeling and rendering architecture from photographs: A hybrid geometry- and image-based approach. In: ACM Trans. on Graphics (SIGGRAPH), pp. 11–20.

Geiger, A., Lenz, P. and Urtasun, R., 2012. Are we ready for autonomous driving? The KITTI vision benchmark suite. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3354–3361.

Geiger, A., Ziegler, J. and Stiller, C., 2011. StereoScan: Dense 3D reconstruction in real-time. In: Proc. IEEE Intelligent Vehicles Symposium (IV), pp. 963–968.

Güney, F. and Geiger, A., 2015. Displets: Resolving stereo ambiguities using object knowledge. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 4165–4175.

Hirschmüller, H., 2008. Stereo processing by semiglobal matching and mutual information. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 30(2), pp. 328–341.

Horn, B. K. P. and Schunck, B. G., 1980. Determining optical flow. Artificial Intelligence (AI) 17(1-3), pp. 185–203.

Huguet, F. and Devernay, F., 2007. A variational method for scene flow estimation from stereo sequences. In: Proc. IEEE International Conf. on Computer Vision (ICCV), pp. 1–7.

Kolmogorov, V., 2006. Convergent tree-reweighted message passing for energy minimization. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 28(10), pp. 1568–1583.

Menze, M. and Geiger, A., 2015. Object scene flow for autonomous vehicles. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3061–3070.

Nir, T., Bruckstein, A. M. and Kimmel, R., 2008. Over-parameterized variational optical flow. International Journal of Computer Vision (IJCV) 76(2), pp. 205–216.

Pacheco, J., Zuffi, S., Black, M. J. and Sudderth, E., 2014. Preserving modes and messages via diverse particle selection. In: Proc. of the International Conf. on Machine learning (ICML), pp. 1152–1160.

Pons, J.-P., Keriven, R. and Faugeras, O., 2007. Multi-view stereo reconstruction and scene flow estimation with a global image-based matching score. International Journal of Computer Vision (IJCV) 72(2), pp. 179–193.

Prisacariu, V., Segal, A. and Reid, I., 2013. Simultaneous monocular 2D segmentation, 3D pose recovery and 3D reconstruction. In: Proc. of the Asian Conf. on Computer Vision (ACCV), pp. 593–606.

Rother, C., Kolmogorov, V., Lempitsky, V. and Szummer, M., 2007. Optimizing binary MRFs via extended roof duality. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1–8.

Sun, D., Wulff, J., Sudderth, E., Pfister, H. and Black, M., 2013. A fully-connected layered model of foreground and background flow. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 2451–2458.

Szeliski, R., 2011. Computer Vision - Algorithms and Applications. Texts in Computer Science, Springer.

Trinh, H. and McAllester, D., 2009. Unsupervised learning of stereo vision with monocular cues. In: Proc. of the British Machine Vision Conf. (BMVC), pp. 1–11.

Valgaerts, L., Bruhn, A., Zimmer, H., Weickert, J., Stoll, C. and Theobalt, C., 2010. Joint estimation of motion, structure and geometry from stereo sequences. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 568–581.

Vedula, S., Baker, S., Rander, P., Collins, R. and Kanade, T., 1999. Three-dimensional scene flow. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 722–729.

Vedula, S., Rander, P., Collins, R. and Kanade, T., 2005. Three-dimensional scene flow. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 27(3), pp. 475–480.

Vogel, C., Roth, S. and Schindler, K., 2014. View-consistent 3D scene flow estimation over multiple frames. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 263–278.

Vogel, C., Schindler, K. and Roth, S., 2011. 3D scene flow estimation with a rigid motion prior. In: Proc. IEEE International Conf. on Computer Vision (ICCV), pp. 1291–1298.

Vogel, C., Schindler, K. and Roth, S., 2013. Piecewise rigid scene flow. In: Proc. IEEE International Conf. on Computer Vision (ICCV), pp. 1377–1384.

Wedel, A., Brox, T., Vaudrey, T., Rabe, C., Franke, U. and Cremers, D., 2011. Stereoscopic scene flow computation for 3D motion understanding. International Journal of Computer Vision (IJCV) 95(1), pp. 29–51.

Wulff, J. and Black, M. J., 2014. Modeling blurred video with layers. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 236–252.

Yamaguchi, K., McAllester, D. and Urtasun, R., 2013. Robust monocular epipolar flow estimation. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 1862–1869.

Yamaguchi, K., McAllester, D. and Urtasun, R., 2014. Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 756–771.

Zabih, R. and Woodfill, J., 1994. Non-parametric local transforms for computing visual correspondence. In: Proc. of the European Conf. on Computer Vision (ECCV), pp. 151–158.

Zia, M., Stark, M. and Schindler, K., 2013a. Explicit occlusion modeling for 3d object class representations. In: Proc. IEEE Conf. on Computer Vision and Pattern Recognition (CVPR), pp. 3326–3333.

Zia, M., Stark, M. and Schindler, K., 2015. Towards scene understanding with detailed 3d object representations. International Journal of Computer Vision (IJCV) 112(2), pp. 188–203.

Zia, M., Stark, M., Schiele, B. and Schindler, K., 2011. Revisiting 3d geometric models for accurate object shape and pose. In: Proc. IEEE International Conf. on Computer Vision (ICCV) Workshops, pp. 569–576.

Zia, M., Stark, M., Schiele, B. and Schindler, K., 2013b. Detailed 3D representations for object recognition and modeling. IEEE Trans. on Pattern Analysis and Machine Intelligence (PAMI) 35(11), pp. 2608–2623.