Joint Object Pose Estimation and Shape Reconstruction in Urban Street Scenes Using 3D Shape Priors

Francis Engelmann(✉), Jörg Stückler, and Bastian Leibe

Computer Vision Group, Visual Computing Institute,
RWTH Aachen University, Aachen, Germany
engelmann@vision.rwth-aachen.de

Abstract. Estimating the pose and 3D shape of a large variety of instances within an object class from stereo images is a challenging problem, especially in realistic conditions such as urban street scenes. We propose a novel approach for using compact shape manifolds of the shape within an object class for object segmentation, pose and shape estimation. Our method first detects objects and estimates their pose coarsely in the stereo images using a state-of-the-art 3D object detection method. An energy minimization method then aligns shape and pose concurrently with the stereo reconstruction of the object. In experiments, we evaluate our approach for detection, pose and shape estimation of cars in real stereo images of urban street scenes. We demonstrate that our shape manifold alignment method yields improved results over the initial stereo reconstruction and object detection method in depth and pose accuracy.

1 Introduction

Object-level shape priors are arguably important components for object pose estimation and 3D reconstruction from images. Shape priors provide strong regularization cues for these problems which are highly ill-posed in a purely data-driven manner. In this paper, we propose a novel approach that uses shape priors for segmenting objects and estimating their shape and pose from stereo images. This knowledge about objects is useful in applications such as autonomous driving or augmented/virtual reality. For autonomous driving, shape and pose can be vital for navigation in order to judge free-space or possible collisions. Accurate 6-DoF pose from single stereo pairs can also be useful for tracking applications.

It is often not feasible to match all possible instances of an object class to images directly using explicit CAD models. Thus, we take the approach to learn compact shape manifolds that represents the intra-class object variance. We find objects, e.g. cars, in the stereo images using a state-of-the-art 3D object detection method (3DOP, [3]) and align the reconstruction with our shape manifold. Our

Electronic supplementary material The online version of this chapter (doi:10.1007/978-3-319-45886-1_18) contains supplementary material, which is available to authorized users.

B. Rosenhahn and B. Andres (Eds.): GCPR 2016, LNCS 9796, pp. 219–230, 2016.
DOI: 10.1007/978-3-319-45886-1_18

method provides shape in occluded areas or regions where vision or laser-based methods often fail, such as textureless, reflective, or transparent surfaces. It also provides a segmentation of the object instance in the image.

In experiments, we assess the performance of our approach for detection, pose and shape estimation of cars in stereo images of urban street scenes using the popular KITTI benchmark [5]. We demonstrate superior pose estimation performance compared to the baseline 3D object detection approach (3DOP). Our method also provides shape reconstructions that improve on the initial stereo reconstruction.

In summary, we make the following contributions: (1) We propose a method that recovers shape and pose of objects from stereo images using class-specific 3D shape priors. Our registration method directly operates on the 3D stereo reconstruction from a single stereo pair. (2) We combine our shape matching method with a state-of-the-art 3D object detection approach to accurately detect objects, determine their 3D pose and shape, and segment them in stereo images. This approach improves the initial stereo reconstruction, especially at farther ranges or textureless and specular surfaces where purely stereo-based reconstructions are inherently limited in reconstruction quality. It also excels in pose accuracy compared to the initial 3D object detection approach.

2 Related Work

Traditionally, multi-view stereo approaches use local surface-based priors to regularize depth reconstruction [6,10,17,23]. SPS-Stereo [23], for instance, uses a piecewise planarity assumption within superpixels for joint stereo reconstruction, segmentation, and optical flow. Recently, several methods have been proposed that incorporate semantic and appearance-based priors that model surface properties of object classes. Saxena et al. [20] demonstrated that depth can be estimated from monocular images using a discriminatively-trained Markov random field model on appearance cues. Sun et al. [21] use random forests to cast votes for object detection, pose and 3D shape of objects in monocular images. A similar method has been proposed by Thomas et al. [22] which transfers depth from training image patches to a detected object using a patch-based implicit shape model detector. Haene et al. [9] propose a variational framework for joint 3D reconstruction and class segmentation in a multi-view stereo setup. The method uses trained object class-specific local surface priors for depth reconstruction. Joint semantic segmentation and 3D reconstruction has also been investigated by Kundu et al. [11] who pose this problem in a higher-order conditional random field to jointly estimate semantic segmentation and 3D occupancy in a volumetric map. Guney and Geiger [8] start from a sparse reconstruction and a semantic segmentation to perform CRF-based dense reconstruction with semantic priors. They impose local shape priors on the superpixels in a semantic segment.

Several semantic SLAM methods have been proposed that include objects through rigid 3D shape templates into the mapping and localization process [7, 18]. Some recent depth reconstruction and SLAM methods also use 3D shape priors. The priors model the shape variation of an object class more generically as manifolds of 3D models. One key difference in such methods is how the

object shape manifold is modeled and how the objects can be detected and their pose recovered. Bao et al. [2] create shape and appearance-based shape priors that model the object shape as a deformable template whose shape depends on a set of anchoring 2D interest points. Implicit shape models based on a truncated signed distance function (TSDF) are used in [16]. This method applies GP-LVMs [12] for manifold learning on the TSDF and recovers segmentation, 3D pose and reconstruction in a level-set formulation. However, the level-set segmentation also requires a pre-trained random forest regressor for inside and outside probabilities on the specific target object, which is obtained through manual initialization on a video. Sandhu et al. [19] use kernel PCA and a region-based level-set formulation instead. Dame et al. [4] take the GP-LVM shape priors in [16] one step further into a monocular dense SLAM system. While we evaluate our method on stereo data from urban scenarios, the above methods are demonstrated in small-scale monocular settings. Closer to our data scenario, [25] detect objects similar to a few annotated examples in TSDF maps that are integrated from stereo depth obtained in urban street scenes. They propose a method to learn PCA shape priors on the detected objects and demonstrate the models to improve reconstruction quality. In contrast, we estimate the shape from a single stereo pair observation in order to be able to align shape priors also with dynamic objects. Furthermore, we use an object proposal method that is trained on a larger set of training examples and provides a coarse 3D pose estimate. Zia et al. [26] use a shape prior encoded by linear embedding of CAD wireframe models to estimate the 3D pose and shape of objects such as cars and bicycles in monocular images. They train a detector for the wireframe vertices and propose an optimization procedure that fits the wireframe shapes to the detected vertices. This shape prior is also used by Menze et al. [14] to jointly estimate object pose and shape together with scene flow from stereo images. Our shape priors represent surface implicitly using TSDFs and, hence, do not require correspondence of wireframe vertices and edges between example instances.

Related to our work are also 3D object detection methods. In 3DOP [3], objects are detected using a selective search approach in SPS-Stereo reconstructions. First, a large set of object proposals is generated in an energy minimization framework that finds potential objects according to a set of shape and appearance based features in a reduced volumetric search space. Each proposal is then classified as part of an object class using a convolutional neural network. The network also regresses the orientation of the object. We use 3DOP for initial 3D object detection and refine the detections using shape prior matching. The work of Zheng et al. [24] focuses on improving an object proposal method with shape priors. Different to our method, it uses a GP-LVM shape prior to sample shapes and render additional training images for a state-of-the-art object proposal approach.

3 Our Method

Our approach to 3D shape recovery and pose estimation aligns a 3D manifold of shapes with noisy stereo reconstructions of objects. We learn this 3D shape

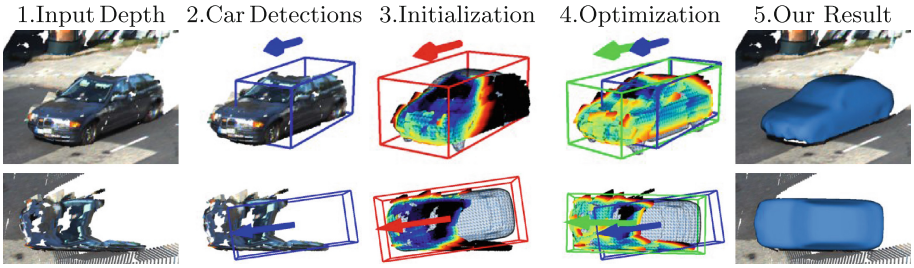

Fig. 1. Overview of our pipeline: from stereo images, we estimate depth using a stereo reconstruction method (e.g. SPS-Stereo [23]) and compute object detections (blue bounding boxes) using 3DOP [3]. We then optimize the shape and the pose of the detected object by solving an energy-minimization problem. The red bounding box shows the initialization that we determine from the 3DOP result and the object segment, the optimization result is shown in green. On the right, we display the optimized shape and pose superimposed on the initial stereo reconstruction. (Color figure online)

prior from a database of CAD models of an object class (for instance, cars). We use a local optimization method that determines shape and pose concurrently in an energy minimization framework. In order to align a shape model to each object in a scene (see Fig. 1), we first detect the objects in stereo data using the state-of-the-art 3D object detection approach from [3]. The method also provides us with a coarse pose initialization, which we subsequently refine with our shape matching method.

3.1 Shape Modelling and Manifold Learning

We learn 3D shape priors using linear subspace analysis (PCA) on a TSDF representation of objects in a class. We use TSDFs, as shapes of object instances can be homogeneously approximated in a shared voxel grid representation. Using a multi-view rendering pipeline, we transform a CAD model database of various object shapes into volumetric TSDF grids. The coordinate frame origin of each instance is placed at the center of gravity and at ground level height, while the axes are aligned with forward, sideward, and upward directions. Figure 2 illustrates example shapes on a learned manifold of cars.

More formally, a TSDF $\Phi(\mathbf{x}, \mathbf{z})$ yields the truncated signed distance at point $\mathbf{x} \in \mathbb{R}^3$ towards the object surface. Hence, the surface is implicitly represented as the zero-level set of the TSDF. The TSDF is approximated through trilinear interpolation of TSDF values $\widetilde{\phi}_i(\mathbf{z})$ at vertices $i \in \mathcal{N}(\mathbf{x})$ in a voxel grid. The vertex set $\mathcal{N}(\mathbf{x})$ corresponds to the corners of the voxel that point $\mathbf{x}$ falls into. The TSDF voxel grid values are embedded in the linear subspace through the mapping

$$\mathbf{z}\left(\widetilde{\boldsymbol{\phi}}\right) = \mathbf{V}^\top \left(\widetilde{\boldsymbol{\phi}} - \boldsymbol{\mu}_{\widetilde{\boldsymbol{\phi}}}\right), \tag{1}$$

Fig. 2. Example shapes in our learned shape manifold of cars. The center shape corresponds to the mean shape.

where $\widetilde{\boldsymbol{\phi}}$ is stacked from all vertex distances and $\boldsymbol{\mu}_{\widetilde{\boldsymbol{\phi}}}$ is its mean over all examples in the training set (i.e., the mean shape). The subspace projection matrix $\mathbf{V}^\top$ is obtained through eigen decomposition $\boldsymbol{\Sigma} = \mathbf{V}\mathbf{D}\mathbf{V}^\top$ of the covariance

$$\boldsymbol{\Sigma} = \frac{1}{M-1}\left(\widetilde{\boldsymbol{\Phi}} - \boldsymbol{\mu}_{\widetilde{\boldsymbol{\phi}}}\right)^\top \left(\widetilde{\boldsymbol{\Phi}} - \boldsymbol{\mu}_{\widetilde{\boldsymbol{\phi}}}\right) \tag{2}$$

where $\widetilde{\boldsymbol{\Phi}}$ is the design matrix stacked from the TSDF vertex distances $\widetilde{\boldsymbol{\phi}}^\top$ of the M examples. Given an encoding $\mathbf{z} \in \mathbb{R}^K$, the corresponding TSDF can be reconstructed using $\widetilde{\boldsymbol{\phi}}(\mathbf{z}) = \mathbf{V}\,\mathbf{z} + \boldsymbol{\mu}_{\widetilde{\boldsymbol{\phi}}}$.

3.2 Object Detection and Segmentation

We detect objects in a scene using 3DOP [3]. We observed that the 2D bounding boxes in the image domain are more accurate then the rather coarse bounding box estimates in 3D. Since we use stereo reconstructions (obtained with libELAS [6] or SPS-Stereo [23] in our experiments), the points $\mathcal{X}$ on an object may not fall inside the 3D bounding box due to disparity noise. Thus, we segment the points on the object that project into the 2D bounding boxes and find points that are close to the estimated 3D center position. As an additional segmentation cue in our urban street scene setting, we remove points on and below the road plane which is found by the same approach as used in 3DOP. Finally, we redetermine the 3D bounding box of the segmented points and determine a pose estimate $\boldsymbol{\xi}_0 := \left(\theta_0\ \mathbf{t}_0^\top\right)^\top$ from the rotation obtained by the 3D bounding box of 3DOP. For the translation we try both the center of the 3D bounding box of the segmented points and the center of the 3DOP bounding box, and pick the best scoring one in terms of energy. We additionally apply a verification step by pruning detections that have an unreasonably sized bounding box, i.e. where the bounding box extent in each dimension is unreasonably small.

3.3 Concurrent Shape and Pose Alignment

We optimize concurrently for shape and pose of the detected objects using the segmented object points $\mathcal{X}$. The pose estimate is initialized from the detected pose $\boldsymbol{\xi}_0$, while the shape estimation is started from the mean shape $\mathbf{z}_0 := \mathbf{0} \in \mathbb{R}^K$. Our energy function corresponds to the negative logarithm of the a-posteriori

probability of the stereo reconstruction given the reconstructed shape and pose estimate,

$$E(\mathcal{X}, \boldsymbol{\xi}, \mathbf{z}) = -\frac{1}{N}\left(\sum_{i=1}^{N} \log\left[p\left(\mathbf{x}_i \mid \boldsymbol{\xi}, \mathbf{z}\right)\right]\right) - \log p(\mathbf{z}) - \log p(\boldsymbol{\xi}), \tag{3}$$

where N is the number of object points. Using our TSDF shape representation, the observation likelihood depends on the distance from the surface,

$$\log p\left(\mathbf{x}_i \mid \boldsymbol{\xi}, \mathbf{z}\right) = \text{const.} - \frac{1}{2\sigma_d^2}\rho\left(\phi\left(R(\theta)\mathbf{x}_i + \mathbf{t}, \mathbf{z}\right)\right). \tag{4}$$

Instead of an outlier-sensitive quadratic norm corresponding to a Gaussian distribution ($\rho(y) = \|y\|_2^2$), we use the robust Huber-norm $\rho(y) = \|y\|_\epsilon$ on the residuals. The shape prior penalizes deviations from the mean shape through

$$\log p(\mathbf{z}) = \text{const.} - \frac{1}{2}\sum_{j=1}^{K}\left(\frac{z_j}{\sigma_j}\right)^2, \tag{5}$$

where σ_j^2 is the eigen value of the j-th principal component. For the pose prior, we can exploit domain knowledge. In the case of cars in urban street scenes, we model that the object should stand on the ground, i.e.,

$$\log p(\boldsymbol{\xi}) = \text{const.} - \frac{1}{2\sigma_y^2}\left(\boldsymbol{t}_y - g(\boldsymbol{t})\right)^2, \tag{6}$$

where $g(\boldsymbol{t})$ is the estimated road height at position $\boldsymbol{t}$. In this setting, we also only need to estimate the rotation of the car around the vertical direction. The noise parameters σ_d and σ_y implement a trade-off between observation likelihood $p\left(\mathbf{x}_i \mid \boldsymbol{\xi}, \mathbf{z}\right)$, shape prior and pose prior. We optimize for pose and shape alternatingly until convergence. The terms are optimized using gradient descent for which we employ the Ceres solver [1].

4 Experiments

We evaluate our method on the popular KITTI dataset [5]. Specifically, we use the KITTI Stereo 2015 training dataset [15], as it focuses mainly on cars in urban and rural scenes. This dataset consists of 200 stereo frames that contain semantic segmentations of 431 vehicles. The stereo images have a resolution of 1242×375 px each and a baseline of 0.54 m. The vehicle segments are also annotated with dense depth from manually fitted CAD models. We added manual ground truth pose annotations for the vehicles on the first 50 frames of the dataset in order to evaluate pose accuracy. For this, we fitted 3D models from Google Warehouse to the data. Throughout our experiments we use a bin size of 0.1 m for the TSDF voxel grids and a truncation distance of ±0.2 m. We empirically determine the noise parameters as $\sigma_d^2 = 0.03$ and $\sigma_y^2 = 3$. As stereo reconstruction inputs we use libELAS [6] and SPS-Stereo [23]. For libELAS, we remove noisy points whose normals are perpendicular to the viewing direction. For 3DOP we use an up-sampling factor of 2.415.

4.1 Pose Estimation Accuracy

We evaluate pose estimation accuracy for 3DOP as our baseline method and variants of our own approach that use stereo reconstructions from libELAS, SPS-Stereo, and the ground truth depth annotation as 3D input for shape and pose alignment. Figure 3, top row, shows the pose accuracy results obtained with these approaches on the 50 pose-annotated frames of the KITTI Stereo 2015 training dataset. It can be seen that our approach is able to improve pose accuracy in translation as well as orientation. Our method even achieves better accuracy at larger distances where the input stereo depth is more noisy.

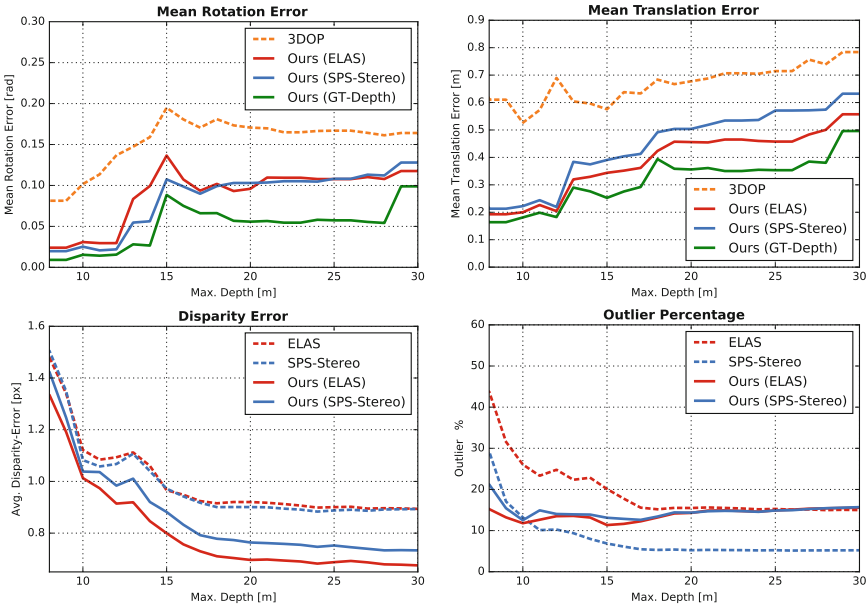

Fig. 3. Upper row: pose estimation results. Both the translation and rotation exhibit significant improvements compared to the baseline method (3DOP, [3]). Bottom row: depth reconstruction results. See text for details.

4.2 Shape Estimation Accuracy

We also assess the quality of the matched shapes using the ground-truth depth annotations on the KITTI Stereo 2015 dataset. To this end, we compare the initial stereo reconstructions by libELAS and SPS-Stereo with improved reconstructions obtained by our method using both initial stereo methods. To obtain the stereo depth maps, we backproject our estimated 3D shapes into the stereo images. To this end, we determine the zero-level set surface represented by the optimized TSDF using the marching cubes algorithm [13]. The results in the bottom row of Fig. 3 show that on average our method can achieve better accuracy in disparity, especially at larger distances. Note, that the outlier rate is

Table 1. Ablation study on shape reconstruction error averaged over all points. We show the effectiveness of each component in our method by enabling them step by step. Note the improvements in score with each step. Figure 4 visualizes the TSDF distance.

Pipeline components	TSDF distance (avg.±std.dev. [m])		
	libELAS [6]	SPS-Stereo [23]	Ground Truth
3DOP + init. pose + mean shape	0.136±0.044	0.134±0.044	0.105±0.032
3DOP + init. pose + optimized shape	0.133±0.045	0.131±0.045	0.086±0.035
3DOP + optimized pose + mean shape	0.127±0.050	0.130±0.048	0.079±0.044
3DOP + optimized pose + optim. shape	**0.124±0.051**	**0.127±0.049**	**0.063±0.040**

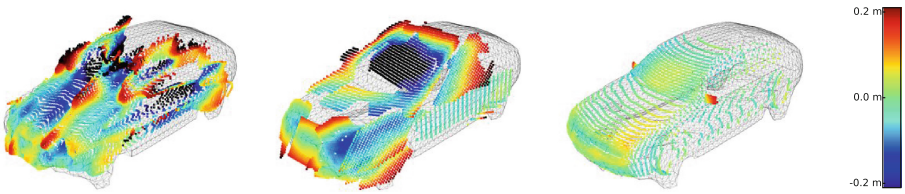

Fig. 4. Estimated pose and shape using different input depths. From left to right: libELAS [6], SPS-Stereo [23], ground truth depth. Points are color-coded by TSDF distance i.e. the Euclidean distance between a point and the zero-level set of the TSDF.

slightly larger for our method, but comparable with libELAS. One reason for this is that pose misalignments cause pixels in the background to be set onto the matched foreground shape.

Table 1 shows shape matching results in terms of mean and std. dev. of the TSDF distance of the points in the object segments. The results demonstrate how much the individual steps in our pipeline contribute to the improvements of the shape alignment. Shape as well as pose optimization improve the alignment. Note that when using stereo depth as input, the distances also include the noise of the stereo depth. Hence, we also give the distances for the ground truth as input in order to assess the shape reconstruction quality isolated from the stereo depth estimation algorithm. We show several qualitative examples of shape matching results in Fig. 4. The examples demonstrate that our method can well align TSDFs through shape and pose optimization to input stereo reconstructions. We also provide a result when using ground truth as input to demonstrate our method on clean inputs.

Figures 5 and 6 show qualitative shape matching results in whole image context using libELAS and SPS-Stereo inputs to our method. The results in Fig. 5 and the left column in Fig. 6 demonstrate that the aligned shapes in these images capture the shape of the objects well. In the upper right image of Fig. 6, it can be seen that for occlusions or far-distance measurements our method can yield misaligned results, but it still captures the coarse pose and shape of the objects in this example. In the middle right image, 3DOP cannot provide good object

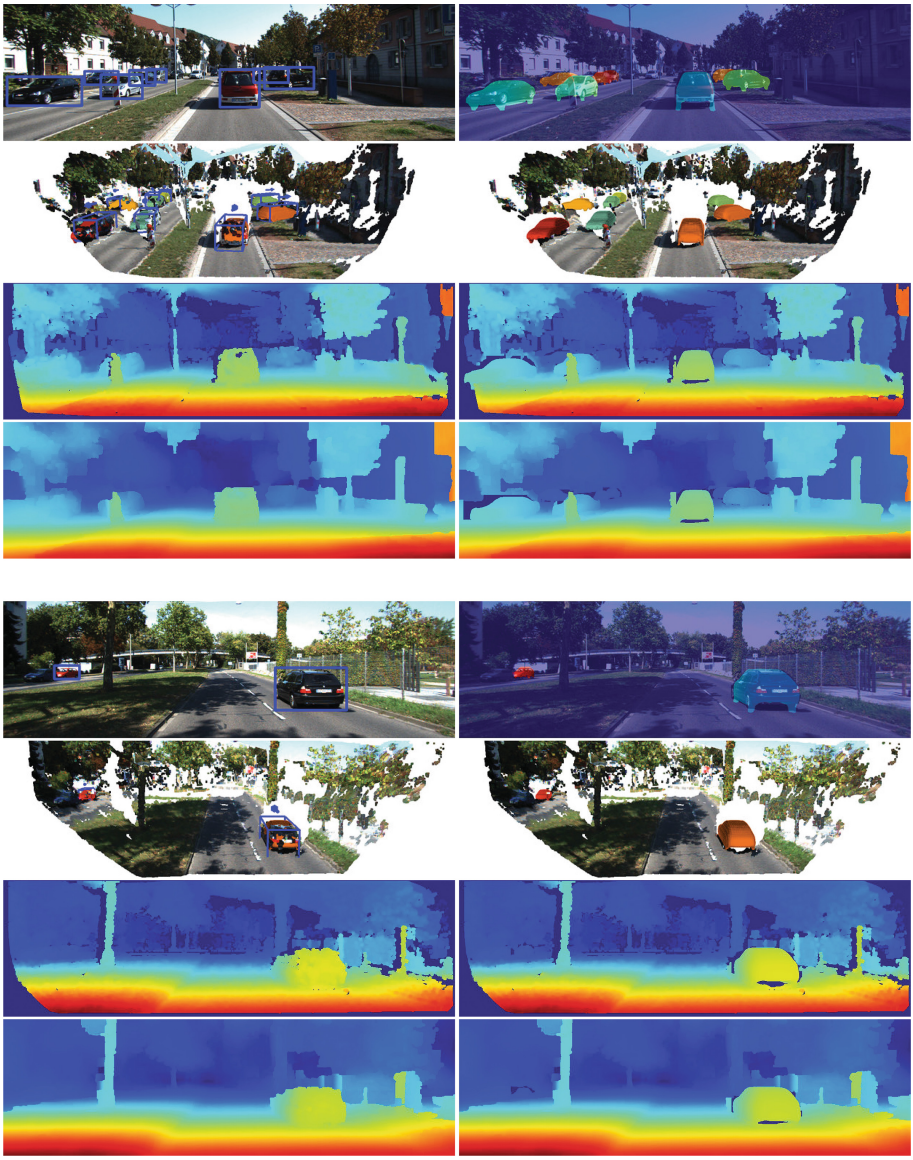

Fig. 5. Qualitative Results From top to bottom, every set of four rows shows: Input image with 2D 3DOP detections and back-projected inferred shapes (first row). Input image with 3D 3DOP detections and 3D view of inferred shapes (second row). libELAS depth map with applied normal filtering and our improved depth map (third row). SPS-Stereo depth map and our improved depth map (fourth row). Depth encoded from small (red) to large (blue) values. (Color figure online)

Fig. 6. Example depth reconstruction and segmentation results for libELAS depth input. Disparity encoded from small (red) to large (blue) values. Right column shows problematic examples for our method. See text for a detailed discussion. (Color figure online)

detections on the parked cars on the right street side. Finally, in the lower right image, a suboptimal matching result is obtained for a truck vehicle which is not represented in our shape manifold. Additional qualitative results are shown in the supplementary material.

5 Conclusions and Future Work

In this paper, we have presented a method for detecting objects of a given class and for concurrently estimating their shape and pose in stereo images from urban street scenes. Our method employs a state-of-the-art 3D object detection approach which coarsely estimates the pose of the objects. We learn a 3D TSDF shape manifold of instances of an object class using a model database. We propose an energy-minimization approach to align stereo reconstructions with object pose and shape within the manifold.

In experiments, we have demonstrated that our method is able to yield improved poses for cars on the KITTI Stereo 2015 dataset compared to [3]. The shape estimated by our method also yields improved depth compared to the input stereo reconstruction methods such as libELAS and SPS-Stereo.

We have demonstrated that our approach works well for cars in urban street scenes. Cars are rigid objects that can be well represented using the linear subspace embedding method. In future work, we want to further investigate the suitability of embedding methods to model deformable or articulated objects in our pipeline.

Acknowledgements. This work has been supported by ERC Starting Grant CV-SUPER (ERC-2012-StG-307432).

References

1. Agarwal, S., Mierle, K.: Ceres solver. http://ceres-solver.org
2. Bao, S.Y., Chandraker, M., Lin, Y., Savarese, S.: Dense object reconstruction with semantic priors. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
3. Chen, X., Kundu, K., Zhu, Y., Berneshawi, A., Ma, H., Fidler, S., Urtasun, R.: 3D object proposals for accurate object class detection. In: Proceedings of Neural Information Processing Systems (NIPS) (2015)
4. Dame, A., Prisacariu, V.A., Ren, C.Y., Reid, I.D.: Dense reconstruction using 3D object shape priors. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
5. Geiger, A., Lenz, P., Urtasun, R.: Are we ready for autonomous driving? the KITTI vision benchmark suite. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2012)
6. Geiger, A., Roser, M., Urtasun, R.: Efficient large-scale stereo matching. In: Kimmel, R., Klette, R., Sugimoto, A. (eds.) ACCV 2010, Part I. LNCS, vol. 6492, pp. 25–38. Springer, Heidelberg (2011)
7. Geiger, A., Wang, C.: Joint 3D object and layout inference from a single RGB-D image. In: Gall, J., et al. (eds.) GCPR 2015. LNCS, vol. 9358, pp. 183–195. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24947-6_15
8. Güney, F., Geiger, A.: Displets: resolving stereo ambiguities using object knowledge. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2015)
9. Häne, C., Zach, C., Cohen, A., Angst, R., Pollefeys, M.: Joint 3D scene reconstruction and class segmentation. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR). pp. 97–104 (2013)
10. Hirschmüller, H.: Stereo processing by semiglobal matching and mutual information. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **30**(2), 328–341 (2008)
11. Kundu, A., Li, Y., Dellaert, F., Li, F., Rehg, J.M.: Joint semantic segmentation and 3D reconstruction from monocular video. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part VI. LNCS, vol. 8694, pp. 703–718. Springer, Heidelberg (2014)
12. Lawrence, N.: Probabilistic non-linear principal component analysis with Gaussian process latent variable models. J. Mach. Learn. Res. (JMLR) **6**, 1783–1816 (2005)
13. Lorensen, W.E., Cline, H.E.: Marching cubes: a high resolution 3D surface construction algorithm. In: Proceedings of SIGGRAPH (1987)
14. Menze, M., Heipke, C., Geiger, A.: Joint 3D estimation of vehicles and scene flow. In: Annals of the Photogrammetry, Remote Sensing and Spatial Information Sciences (2015)
15. Menze, M., Geiger, A.: Object scene flow for autonomous vehicles. In: CVPR (2015)
16. Prisacariu, V.A., Segal, A.V., Reid, I.: Simultaneous monocular 2D segmentation, 3D pose recovery and 3D reconstruction. In: Lee, K.M., Matsushita, Y., Rehg, J.M., Hu, Z. (eds.) ACCV 2012, Part I. LNCS, vol. 7724, pp. 593–606. Springer, Heidelberg (2013)
17. Ranftl, R., Gehrig, S., Pock, T., Bischof, H.: Pushing the limits of stereo using variational stereo estimation. In: Proceedings of the Intelligent Vehicles Symposium (2012)

18. Salas-Moreno, R.F., Newcombe, R.A., Strasdat, H., Kelly, P.H., Davison, A.J.: SLAM++: simultaneous localisation and mapping at the level of objects. In: Proceedings of the IEEE International Conference on Computer Vision and Pattern Recognition (CVPR) (2013)
19. Sandhu, R., Dambreville, S., Yezzi, A., Tannenbaum, A.: A nonrigid kernel-based framework for 2D–3D pose estimation and 2D image segmentation. IEEE Trans. Pattern Anal. Mach. Intell. **33**(6), 1098–1115 (2011)
20. Saxena, A., Chung, S.H., Ng, A.Y.: Learning depth from single monocular images. In: Proceedings of Neural Information Processing Systems (NIPS) (2005)
21. Sun, M., Bradski, G., Xu, B.-X., Savarese, S.: Depth-encoded hough voting for joint object detection and shape recovery. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010, Part V. LNCS, vol. 6315, pp. 658–671. Springer, Heidelberg (2010)
22. Thomas, A., Ferrari, V., Leibe, B., Tuytelaars, T., Van Gool, L.: Depth-from-recognition: inferring meta-data by cognitive feedback. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2007)
23. Yamaguchi, K., McAllester, D., Urtasun, R.: Efficient joint segmentation, occlusion labeling, stereo and flow estimation. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014, Part V. LNCS, vol. 8693, pp. 756–771. Springer, Heidelberg (2014)
24. Zheng, S., Prisacariu, V.A., Averkiou, M., Cheng, M.-M., Mitra, N.J., Shotton, J., Torr, P.H.S., Rother, C.: Object proposals estimation in depth image using compact 3D shape manifolds. In: Gall, J., et al. (eds.) GCPR 2015. LNCS, vol. 9358, pp. 196–208. Springer, Heidelberg (2015). doi:10.1007/978-3-319-24947-6_16
25. Zhou, C., Güney, F., Wang, Y., Geiger, A.: Exploiting object similarity in 3D reconstruction. In: Proceedings of the IEEE International Conference on Computer Vision (ICCV) (2015)
26. Zia, M., Stark, M., Schiele, B., Schindler, K.: Detailed 3D representations for object recognition and modeling. IEEE Trans. Pattern Anal. Mach. Intell. (PAMI) **35**, 2608–2623 (2013)