Robust 3D Car Shape Estimation from Landmarks in Monocular Image

Yanan Miao
miaoyn12@mails.tsinghua.edu.cn
Xiaoming Tao
taoxm@mail.tsinghua.edu.cn
Jianhua Lu
lhh-dee@mail.tsinghua.edu.cn

Tsinghua National Laboratory for Information Science and Technology (TNList)
Department of Electronic Engineering
Tsinghua University
Beijing, P. R. China

Abstract

The reconstruction of 3D object shape from monocular image is inherently an ill-posed problem. And it suffers significant performance degradation when large errors are present. In this paper, we propose a robust model to estimate 3D shape from 2D landmarks with unknown camera pose. The 3D shape of the object is assumed as a linear combination of predefined shape basis. To handle severely contaminated observations, we explicitly model the outliers as sparse noise. The objective function hence is non-convex and non-smooth constrained on Stiefel manifold, where the coupling of the unknown shape representation coefficients and camera pose makes it more difficult to solve. We then propose a numerical algorithm based on Alternative Direction Method of Multipliers to optimize it. We set the orthogonality constraints into the smooth sub-problem, which admits a closed-form solution. The proposed algorithm can achieve convergence rapidly. Experimental results both in controlled experiments and on real data show that, the proposed method outperforms the other methods.

1 Introduction

Recently, 3D object models have been received much attention for object recognition [15, 19], face model [4, 9, 22] and pose estimation [3, 11, 30]. The utilization of geometric structure can help to capture great intra-class variations of the object. An essential procedure in these works is to establish a 3D model and fit it into the 2D image plane. Geometrically, it is an inverse-problem to estimate the 3D geometry from their projections in a monocular 2D image. To resolve this kind of ambiguity, the common approach is to introduce priori information of the 3D shape. In this paper, we focus on a kind of widely used 3D global geometry model as shape prior. This model is defined as a collection of ordered vertices or landmarks, which is originated from “active shape model" (ASM) [8]. Followed the ASM, each shape is assumed to be represented as a linear combination of some predefined basis [20, 31, 32]. A typical framework for 3D car shape estimation is shown in Figure 1. In this paper, we focus on the modular in the dash block.

There are two main challenges when fitting such a kind of 3D model to 2D image. First, the 2D observations are usually acquired on cluttered “in-the-wild" images. To find the landmarks, contemporary methods train discriminative detectors for each local part [3, 18] or use

Pages 99.1-99.11
DOI: https://dx.doi.org/10.5244/C.30.99

kind of cascaded regression-based methods to encode the 2D geometry [28]. These works achieve successful results, however, this problem is still not fully solved. The landmarks are always inaccurately detected due to occlusions or clutters under complex environment and illumination conditions, which can affect the shape estimate. Second, the objective function for 3D shape estimating is usually highly non-linear [20, 32] and non-convex. It is strongly influenced by the initialization quality, and each of the initializations might get stuck a local optimal solution [13, 20]. In addition, if both camera pose and the object shape are unknown, the problem becomes much more difficult. To improve the initialization, some works have tried to use multiple start points [25, 32] or multi-scale scheme [17]. Yet, the time to achieve convergence may be increased, and the performance can still degrade due to bad initializations.

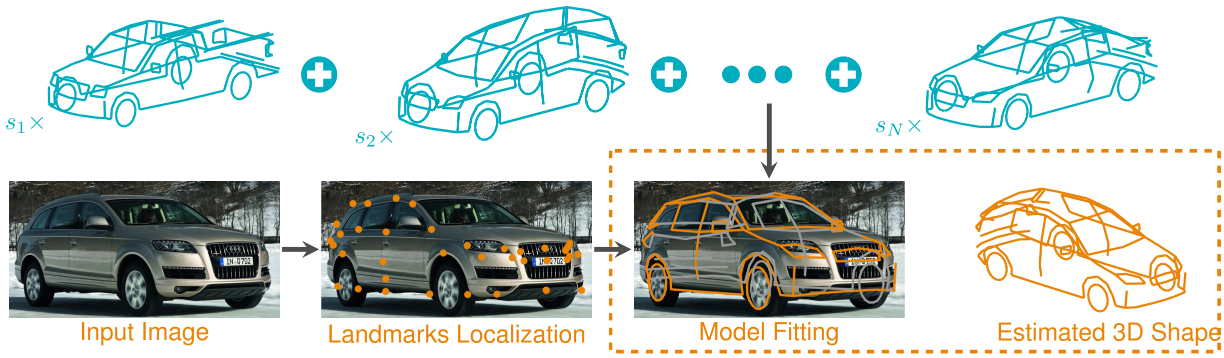

Figure 1: The framework for 3D shape estimation. **Top**: A series of prior 3D shape basis [19]. **Bottom**: The shape estimation procedure from 2D landmarks for a given input image.

In this paper, we investigate robust model to handle outliers and develop efficient algorithms. We first propose a robust model to estimate both 3D shape and the camera pose for rigid-body object, from the 2D observed landmarks. We assume the observations are contaminated by two kind of noise. One is dense noise in measurement, and the other is sparse noise. We encode the sparsity by use of ℓ_1-norm as a surrogate of cardinality, which can effectively model the large errors in 2D landmarks detection. In order to solve the non-convex and non-smooth problem, we then propose an efficient numerical method based on alternating direction method of multipliers (ADMM). The orthogonality constraints are set into the sub-problem, where we can get a closed-form solution. The proposed algorithm has also a fast convergence rate. Experimental results demonstrate that, the proposed model is robust to outliers and shows competitive results.

The remainder of this paper is structured as follows. Section 2 introduces some related works. In Section 3, we introduce the proposed robust model and formulate the problem. We then describe the proposed numerical algorithm in detail to solve the problem in Section 4. In Section 5, we do quantities of experiments to validate the effectiveness of the proposed model and algorithm. The last section draws a conclusion.

2 Related Works

Our method is related to recent works including the 3D shape model representation and estimation, and optimization on manifolds. To use this kind of model, part-based object representations [10] are widely used for modelling each landmark appearance under a number of discrete viewpoints. Some other methods such as discriminative [18, 22] or regression-based methods [21, 28] can also be utilized. In our work, we assume the 2D landmarks are given previously by any possible method.

The related works about 3D geometry reasoning can be roughly categorized into two classes: non-rigid (*e.g*. body, face) shape estimation and rigid object shape estimation (*e.g*. car). For non-rigid objects, the main challenge is the great structural variations of the shape. Some works [2, 24] customize model for 3D human pose recovery, which are not for shape reconstruction. Ramakrishna *et al*. [20] represent the 3D shape as sparse embedding in an over-complete dictionary and estimate the model using projected matching pursuit. Wang *et al*. [25] extend their work via ℓ_1-norm to measure the matching errors to handle inaccurate 2D joints. Recently, Zhou *et al*. [31] have proposed a convex-relaxation approach, where each shape basis is rotatable in their model and achieves appealing results with arbitrary initialization. This method, however, is not robust to outliers in observation. In contrast, we try to establish a model to cope with the interference from large errors. For rigid object, Leotta *et al*. [17] propose a deformable model combined mesh and points sampled from CAD models, and fit the parametrized-model to images by minimizing the error between hypothesised edges to observed edges. Both [32] and [19] use a ASM-like model which are more related to our work. Zia *et al*. [32] train discriminative detectors to localize 2D landmarks and estimate the 3D pose use a sample-based hill-climbing scheme. This method need to evaluate many hypotheses which lead to high time consumption. In contrast, we aim to develop more efficient algorithm to match model into 2D image.

The proposed optimization algorithm is related to some recent advances about optimization on orthogonality constraints. Orthogonal Procrustes [23] and other optimization problems on Stiefeld manifolds have been studied [1]. Recently, Wen *et al*. [26] propose a curvilinear descent path for generic manifold optimization. Inspired by Bregman iteration, Lai *et al*. [16] propose a splitting orthogonal constraints method. Boumal *et al*. [6] release a generic MATLAB toolbox to solve smooth optimization problems on various manifolds. All these methods focus on smooth optimization on different manifolds, however, in our problem, both the 3D shape and camera pose are unknown, and the shape representation term is non-smooth. The coupling of them makes the problem hard to be solved. Therefore, the above method can not be directly applied to our problem. Some works [14, 29] focus on the non-convex optimization based on ADMM [7]. The convergence properties of ADMM method for minimization on Stiefel Manifold have been studied [29]. In our work, inspired by their results, we investigate to use ADMM to solve the non-convex, non-smooth problem with orthogonality constrains.

3 Problem Formulation

Given the coordinates of p 2D landmarks $\mathbf{x} \in \mathbb{R}^{2\times p}$, and a series of 3D priori shapes $\{X_i\}_{i=1}^{N} \in \mathbb{R}^{3\times p}$ with mean shape $\mu \in \mathbb{R}^{3\times p}$, we pursue the real 3D shape $\mathbf{X}$ of the object. The 3D model consists of predefined points, each of which has the same semantic meaning. Assume $\mathbf{X}$ is represented as a linear combination of pre-defined shape basis $\mathbf{X} = \sum_{i=1}^{N} s_i X_i + \mu$, where $\mathbf{s} = [s_1, \ldots, s_N]^\mathsf{T} \in \mathbb{R}^N$ are the shape coefficients, and μ is the mean 3D shape. To simplify the problem, we use a weak perspective camera model, where the camera matrix is denoted by

$$\mathsf{P} = \begin{bmatrix} \alpha_x & 0 & 0 \\ 0 & \alpha_y & 0 \end{bmatrix} \mathbf{R}. \tag{1}$$

The rotation matrix $\mathbf{R}$ is in the set of spherical orthogonal group $SO(3) = \{\mathbf{R} \in \mathbb{R}^{3\times 3}, \mathbf{R}^\top \mathbf{R} = \mathbf{I}, \det(\mathbf{R}) = 1\}$ [12]. Constant α_x, α_y are the scale of the axis, which depend on the camera intrinsic parameters. In this paper, we assume the camera intrinsic parameters (e.g. the camera focal length) are known a priori and the scale $\alpha_x = \alpha_y = 1$ in P. Then the camera matrix is given by $\mathsf{M} = \mathbf{I}_{2\times 3}\mathbf{R}$.

Now, the shape estimation problem is equivalent to estimate the shape representation coefficients $\mathbf{s}$ and the unknown camera pose M. We try to minimize the error between the observations $\mathbf{x}$ and the projected points $\mathsf{M}\mathbf{X}$. Assume that there are Gaussian noises between the observations and the projected model points. Then, the objective function for 3D shape estimation can be formulated as

$$\begin{array}{ll} \min\limits_{\mathbf{s},\mathsf{M}} & \frac{1}{2}\left\|\mathbf{x} - \mathsf{M}\left(\sum_{i=1}^{N} s_i X_i + \mu\right)\right\|_F^2 + \lambda\|\mathbf{s}\|_1 \\ s.t. & \mathsf{M}\mathsf{M}^\top = \mathbf{I}_2, \end{array} \tag{2}$$

where $\mathbf{I}$ represents the identify matrix and λ is the regularization parameter.

There are always large errors when the landmarks are inaccurately detected, which result from the complex background and illumination conditions. To address this problem, we propose a robust 3D shape estimation model. We explicitly model the outliers by introducing an additional sparse error term $E \in \mathbb{R}^{2\times p}$. We encode the sparsity by use of ℓ_1-norm as a surrogate of cardinality. In this situation, the data cannot be pre-centralized, therefore, we must also estimate the translation $\mathbf{t} \in \mathbb{R}^{2\times p}$. Thus, the robust model is then formulated as

$$\begin{array}{ll} \min\limits_{\mathbf{s},\mathsf{M},E,\mathbf{t}} & \frac{1}{2}\left\|\mathbf{x} - \mathbf{t} - \mathsf{M}\left(\sum_{i=1}^{N} s_i X_i + \mu\right) - E\right\|_F^2 + \lambda\|\mathbf{s}\|_1 + \eta\|E\|_1 \\ s.t. & \mathsf{M}\mathsf{M}^\top = \mathbf{I}_2, \end{array} \tag{3}$$

where $\mathbf{t} = [t_x, t_y]^\top \cdot \mathbf{1}_{1\times p}$ and η is the regularization parameter. We will describe the estimation algorithm detail in the following section.

4 Method

Problem (3) is a non-convex optimization problem on the Stiefel manifolds with ℓ_1-norm regularized, which is non-smooth. Moreover, the coupling of the camera matrix and the shape representation coefficients make it difficult to solve problem (3). In this section, we propose an algorithm based on the alternating direction method of multipliers for problem (3). We first introduce a lemma which is useful for solving sub-problems in the subsequent algorithms.

Lemma 1. *For the following orthogonal constrained Least Square problem*

$$\begin{array}{ll} \min\limits_{X} & \frac{1}{2}\|X - Y\|_F^2 \\ s.t. & YY^\top = \mathbf{I}_2, \end{array} \tag{4}$$

where $X, Y \in \mathbb{R}^{2\times 3}$. The optimal solution is given by $X^* = U\mathbf{I}_{2\times 3}V^\top$, where U, V satisfying the SVD factorization $Y = U\mathbf{D}V^\top$ and $\mathbf{D}$ is the diagonal matrix. ■

4.1 Proposed Algorithm

A prevailing method is an iterative method called coordinate descent algorithm (CD) or alternative minimization. The basic idea is that, each iterate is obtained by fixing most components of the variable vector at their values from the current iteration, and approximately minimizing the objective with respect to the remaining components [27]. The coordinate descent algorithm shows good performance. Here, we introduce a novel numerical method based on ADMM which is more efficient. For problem (2) and (3), a key consideration is that how to deal with the orthogonality constraints. Some work estimate the camera matrix by solving a procrustes problem [20] or by splitting two orthogonal rows of the camera matrix and updating alternatively [25], which is not very efficient. In the proposed algorithm, we set the orthogonality constraints into the sub-problem, which can be simply solved in closed-form. The other sub-problems are well-known and can be solved easily. We present our method in detail in the followings.

First, we define an auxiliary variable $V \in \mathbb{R}^{2\times 3}$ and consider the equivalent optimization problem,

$$\begin{array}{ll} \min\limits_{\mathbf{s},\mathsf{M},V,E,\mathbf{t}} & \frac{1}{2}\left\|\mathbf{x}-\mathbf{t}-\mathsf{M}\left(\sum_{i=1}^{N} s_i X_i+\mu\right)-E\right\|_F^2+\lambda\|\mathbf{s}\|_1+\eta\|E\|_1 \\ s.t. & \mathsf{M}=V, \\ & VV^{\mathsf{T}}=\mathbf{I}_2. \end{array} \tag{5}$$

Then the augmented Lagrangian is

$$\begin{aligned} \mathcal{L}(\mathsf{M},V,\mathbf{s},E,\mathbf{t},\Lambda) = & \frac{1}{2}\left\|\mathbf{x}-\mathbf{t}-\mathsf{M}\left(\sum_{i=1}^{N} s_i X_i+\mu\right)-E\right\|_F^2 \\ & +\lambda\|\mathbf{s}\|_1+\eta\|E\|_1+<\Lambda,\mathsf{M}-V>+\frac{\tau}{2}\|\mathsf{M}-V\|_F^2, \end{aligned} \tag{6}$$

where Λ is the multiplier and the τ is the penalty parameter. We update each block with all the others are fixed. Superscript k indicates the iteration number.

Solve $\mathsf{M},E,\mathbf{t}$. For M-minimization step, we have

$$\mathsf{M}^{(k+1)} = \arg\min_{\mathsf{M}} \frac{1}{2}\left\|\mathsf{M}\mathbf{X}+E^{(k)}+\mathbf{t}^{(k)}-\mathbf{x}\right\|_F^2+<\Lambda^{(k)},\mathsf{M}-V^{(k)}>+\frac{\tau^{(k)}}{2}\left\|\mathsf{M}-V^{(k)}\right\|_F^2, \tag{7}$$

where $\mathbf{X}=\sum_{i=1}^{N} s_i^{(k)} X_i+\mu$. This step admits a closed-form solution. Let $\partial\mathcal{L}(\mathsf{M})/\partial\mathsf{M}=\mathbf{0}$, we get

$$\mathsf{M}^{(k+1)} = \left[(\mathbf{x}-\mathbf{t}^{(k)}-E^{(k)})\mathbf{X}^{\mathsf{T}}+\tau^{(k)}V^{(k)}-\Lambda^{(k)}\right]\cdot\left(\mathbf{X}\mathbf{X}^{\mathsf{T}}+\tau^{(k)}\mathbf{I}\right)^{-1}. \tag{8}$$

To extract the outlier pattern E, we update E using element-wise soft-thresholding [5] by simple calculus, which produces

$$E^{(k+1)} = \mathcal{T}_\eta\left[\mathbf{x}-\mathbf{t}^{(k)}-\mathsf{M}^{(k+1)}\mathbf{X}\right], \tag{9}$$

where $\mathcal{T}_\alpha(x)=\mathrm{sign}(x)\left(|x|-\alpha\right)_+$ is a shrinkage operator.

The translation $\mathbf{t}$ is easy to calculate, and $t=[t_x,t_y]^{\mathsf{T}}$ is given by the mean value of $\mathbf{x}-E^{(k+1)}-\mathsf{M}^{(k+1)}\mathbf{X}$ along the row. Therefore, we can get $\mathbf{t}^{(k+1)}=t\cdot\mathbf{1}_{1\times p}$.

Solve $\mathbf{s}, V$. To solve the shape representations $\mathbf{s}$, we first let $\mathbf{y} = \mathbf{vec}([\mathbf{x} - \mathbf{t} - E - \mathsf{M}\mu])$ and $\Phi = (\mathbf{I} \otimes \mathsf{M})B$, where $B = [\mathbf{vec}(X_1), \mathbf{vec}(X_2), \dots, \mathbf{vec}(X_N)]$, and $\otimes$ denotes the Kronecker production. Then, we can re-arrange the sub-problem as

$$\mathbf{s} = \arg\min_{\mathbf{s}} \frac{1}{2} \|\Phi \mathbf{s} - \mathbf{y}\|_2^2 + \lambda \|\mathbf{s}\|_1, \tag{10}$$

which is a standard *Lasso* problem. We solve it by use of FISTA [5] due to its efficiency. In the next, the V-minimization step is

$$V^{(k+1)} = \arg\min_{V} \left\{ \left\| V - \left(\mathsf{M}^{(k+1)} + \frac{\Lambda^{(k)}}{\tau^{(k)}} \right) \right\|_F^2 : VV^{\mathsf{T}} = \mathbf{I}_2 \right\}. \tag{11}$$

By Lemma 1, the closed-form solution is given by

$$V^{(k+1)} = U\mathbf{I}_{2\times 3}W^T, \tag{12}$$

where U and W satisfy $[U,S,W] = \mathcal{SVD}\left[\mathsf{M}^{(k+1)} + \Lambda^{(k)}/\tau^{(k)}\right]$.

Update Λ, τ. At last, the multipliers are updated by $\Lambda^{(k+1)} = \Lambda^{(k)} + \tau^{(k)}(\mathsf{M}^{(k+1)} - V^{k(+1)})$ and the penalty parameter τ is updated according to the suggestions in [7] by a varying manner. ■

The properties of convergence for non-convex problems by ADMM have been discussed in [29]. If there are more than two blocks to be updated in the procedure, the convergences can not be always guaranteed, which may be influenced by the update ordering. We find that the update ordering as specified in the Algorithm 1 can lead convergence. The proposed algorithm shows a fast convergence rate.

Algorithm 1: Robust Shape Estimation by ADMM

Input: 2D landmark positions $\mathbf{x}$, 3D basis $\{X\}_{i=1}^N$, regularization parameters λ,η
Output: M, $\mathbf{s}$, E, $\mathbf{t}$

Initialize $\mathbf{s}$, M, *E and* $\mathbf{t}$;
repeat
  Update M according to Eq. (8);
  Update $\mathbf{s}$ by solving Eq. (10);
  Update V according to Eq. (12);
  Update E according to Eq. (9);
  Update $\mathbf{t} \leftarrow \mathrm{mean}\,[\mathbf{x} - E^{(k+1)} - \mathsf{M}^{(k+1)}\mathbf{X}] \cdot \mathbf{1}_{1\times p}$;
  Update Λ, τ ;
  $k \leftarrow k+1$;
until *convergence*;

5 Experiments

To validate the capability of the proposed algorithm, we evaluate the proposed method in controlled experiments and on some dataset real image data. The only adjustable parameters in our proposed algorithm are the regularization parameters λ and η. By parameter selection, we set $\lambda = 0.1$ and $\eta = 0.01$. In purpose of comparison, we implement the widely-used

Coordinate Descent method (referred as CD) [20], and a convex relaxation method [31] (referred as CVXR), which shows the most state-of-the-art, where the corresponding parameters are set as suggested. In purpose of fairly comparison, the stop criterion and initializations are set the same.

5.1 Dataset

3D Data. The 3D car shape data we used is adopted from Zia introduced in [32]. Their wireframe exemplar is defined as a collection of ordered vertices residing in 3D space. Each exemplar is chosen from the set of vertices, which make up a 3D CAD model with topology pre-defined. There are total 36 vertices for each 3D exemplar. In this paper, we use 30 exemplars to form the 3D shape basis.

Test Image Data. We use the car dataset from MIT Street Database[1]. We annotate the images as suggested in [18]. All the labelled data are roughly divided into 5 different views, which are 900 frontal view, 1400 frontal-side view, 800 profile view, 1200 rear-side view, and 1160 rear view images. The images are labelled by 8,14,10,14,8 landmarks for each view respectively. We randomly select 50 percentage of each view for test.

5.2 Experiment in control

We first evaluate the performance on no-outlier case, where the ground-truth are regarded as inputs directly. The basis number is fixed to 18 in each case. To show the efficiency of the proposed method, we first compare the iteration times under different amount of observations. We repeat 10 times at a specified observation number, and each time we randomly select the points from all the available landmarks as inputs. In Figure 2, we show the average iteration number on *Crossover* and *Sedan*. The results demonstrate that our algorithms, need less iterations to achieve convergence under different amount of observations for different types of the car.

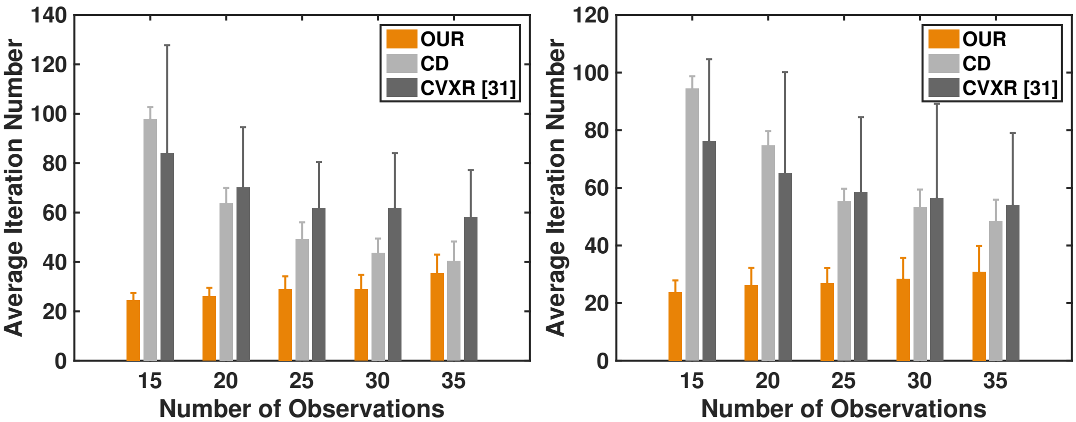

Figure 2: Average iteration numbers under different amount of observations for *Crossover* and *Sedan*. One standard deviation is preserved.

We use root-mean-square error (RMSE) of the landmark localisation to evaluate the estimation accuracy. Figure 3 shows the cumulative distribution of the landmarks localisation errors. From the results, we can see that these three method have similar performances without any outliers in the observations, where the proposed method still performs the best.

[1]http://cbcl.mit.edu/software-datasets/streetscenes/

Then, to investigate the ability to deal with outliers, by use of the proposed robust model, we randomly select some portions of the visible landmarks and add large shift to these points. For each test sample, we repeat this procedure 10 times. Figure 3 shows the average esti-

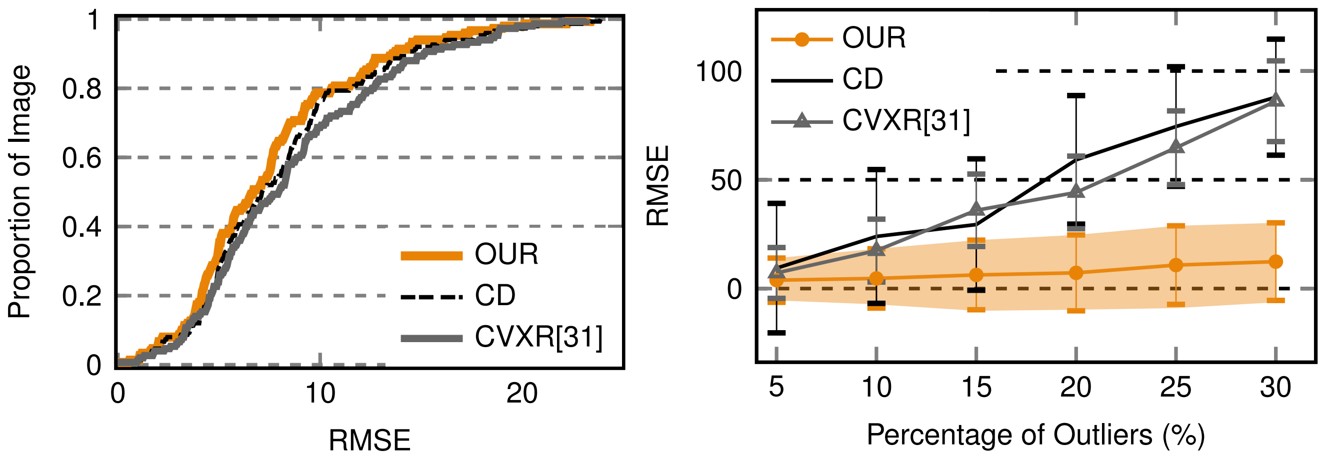

Figure 3: **Left:** The cumulative distribution for localization errors. The y-axis represents the ratio of the test samples, whose errors are no less than the corresponding value in x-axis; **Right:** Error versus different percentage outliers.

mation error under added large errors, with one standard deviation preserved. The proposed model is obviously robust for different percentage of added outliers. Especially, when there are more outliers, the proposed method achieves better performance.

5.3 Test on real data

To evaluate the applicability of the proposed robust model, we examine it on MIT street dataset. The selected test images are all resized into 700×700 pixels.

Landmark Detection. To detect the 2D landmarks, we first generate the local patches for feature extraction. For each landmark, we extract a 40×40 image patch as a positive patch, and 30 negative patches with their centred pixels apart from the true landmark at least 10 pixels. We then compute HoG descriptors on each extracted patch to describe the local appearance. These features are fed to SVM discriminative classifiers to independently learn a detector for each visible landmark. We then train a logistic regressor to map the output of the classifier to the range from 0 to 1. In the test stage, the position of each landmark is determined according to the response map from the corresponding classifier and regressor. The bounding-box of the car is precomputed by DPM [10] as assist information to accelerate the detection.

Figure 4 shows the results of landmark localisation and 3D shape estimation. Results show that the proposed model can handle the outliers more efficiently. Even if there are error detection points, the proposed method can recovery the 3D shape well. As a matter of fact, it is hard to just use these independent detectors to acquire good detection results. The performance can be improved with some other efficient landmarks detection method. From the results, we can see that the visualized 3D shapes may not seem very pleasant. The reason is that, the simple wireframe cannot fully describe the detail shape information of the car. In the future, we will define more points to represented the 3D shape, and investigate the estimation performance under different number of observed landmarks.

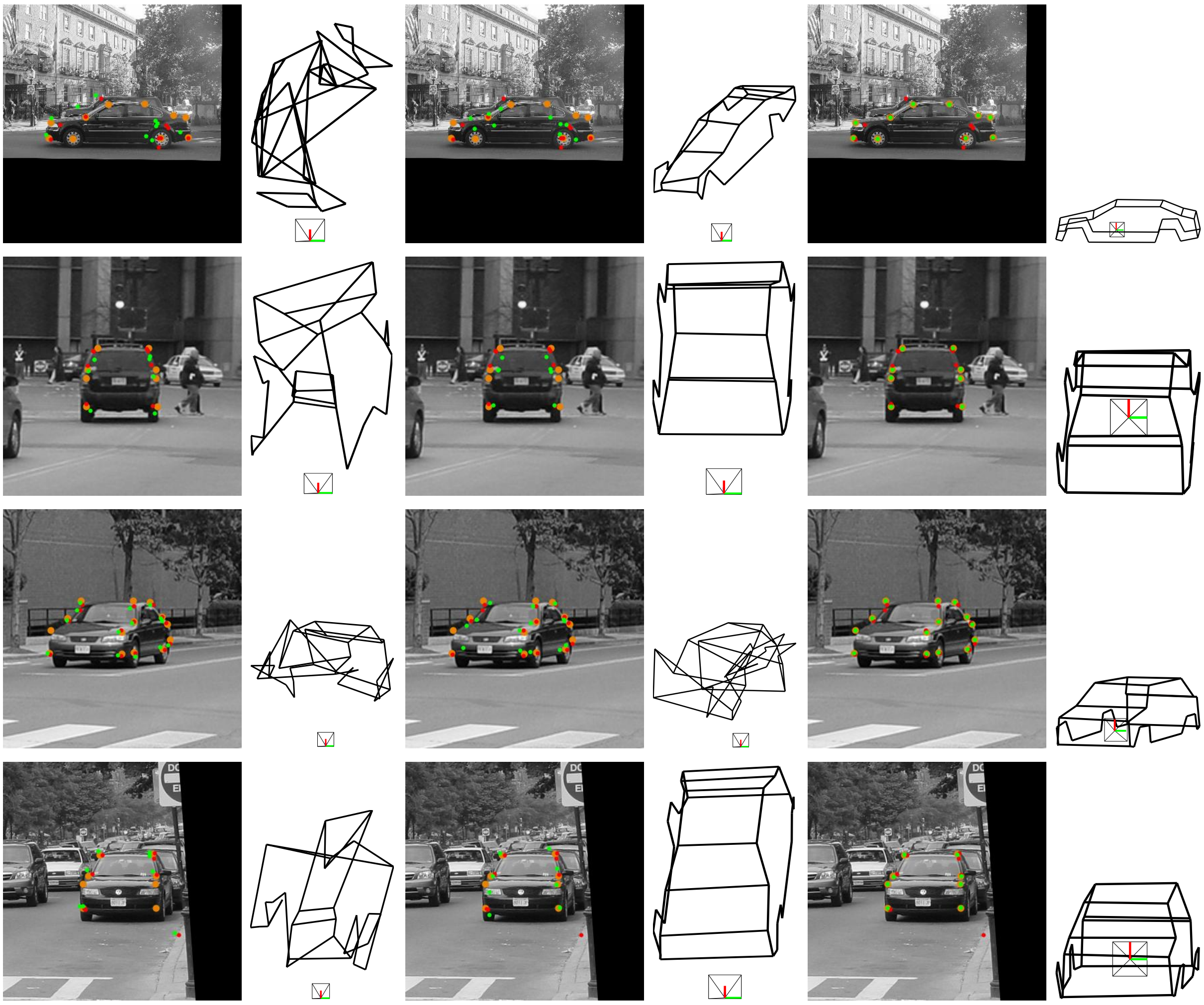

Figure 4: Estimates results on MIT Street Dataset. From **left** to **right**, the results are given by CD, CVXR, and the proposed method, where ● denotes the detected landmarks, and the estimated and ground-truth landmarks are marked as ● and ● respectively.

6 Conclusions

In this paper, we have proposed a robust model to handle the outliers in observations, when estimating 3D car shape from 2D landmarks in monocular image. We model the outliers as sparse noise and encode them explicitly by use of ℓ_1-norm as a surrogate of cardinality. To solve the non-convex and non-smooth problem effectively, we propose a numerical method based on ADMM. The orthogonality constraints are set into sub-problem, which admits a closed-form solution. The proposed algorithm achieves a very fast convergence rate. Experimental results have shown better performances in controlled experiments and on real data, compared with the-state-of-the-art.

Acknowledgement

This work was supported by the National Basic Research Project of China (973) (2013CB-329006) and National Natural Science Foundation of China (NSFC, 61471220, 91538107, 61021001).

References

[1] P-A Absil, Robert Mahony, and Rodolphe Sepulchre. *Optimization algorithms on matrix manifolds*. Princeton University Press, 2009.

[2] Ankur Agarwal and Bill Triggs. Recovering 3d human pose from monocular images. *IEEE Trans. Pattern Anal. Mach. Intell.*, 28(1):44–58, 2006.

[3] Mykhaylo Andriluka, Stefan Roth, and Bernt Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, pages 1014–1021, 2009.

[4] Akshay Asthana, Stefanos Zafeiriou, Shiyang Cheng, and Maja Pantic. Robust discriminative response map fitting with constrained local models. In *CVPR*, pages 3444–3451, 2013.

[5] Amir Beck and Marc Teboulle. A fast iterative shrinkage-thresholding algorithm for linear inverse problems. *SIAM J. Imaging Sci.*, 2(1):183–202, 2009.

[6] Nicolas Boumal, Bamdev Mishra, P-A Absil, and Rodolphe Sepulchre. Manopt, a matlab toolbox for optimization on manifolds. *J. Mach. Learn. Res.*, 15(1):1455–1459, 2014.

[7] Stephen Boyd and Neal etc. Parikh. Distributed optimization and statistical learning via the alternating direction method of multipliers. *Foundations and Trends® in Machine Learning*, 3(1):1–122, 2011.

[8] Timothy F Cootes, Gareth J Edwards, and Christopher J Taylor. Active appearance models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 23(6):681–685, 2001.

[9] David Cristinacce and Timothy F Cootes. Feature detection and tracking with constrained local models. In *BMVC*, volume 2, page 6, 2006.

[10] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1627–1645, 2010.

[11] Sanja Fidler, Sven Dickinson, and Raquel Urtasun. 3d object detection and viewpoint estimation with a deformable 3d cuboid model. In *NIPS*, pages 611–619, 2012.

[12] Richard Hartley and Andrew Zisserman. *Multiple view geometry in computer vision*. Cambridge university press, 2003.

[13] Mohsen Hejrati and Deva Ramanan. Analyzing 3d objects in cluttered images. In *NIPS*, pages 593–601, 2012.

[14] Takafumi Kanamori and Akiko Takeda. Non-convex optimization on stiefel manifold and applications to machine learning. In *NIPS*, pages 109–116. Springer, 2012.

[15] Jan Krause, Michael Stark, Jia Deng, and Li Fei-Fei. 3d object representations for fine-grained categorization. In *ICCVW*, pages 554–561, 2013.

[16] Rongjie Lai and Stanley Osher. A splitting method for orthogonality constrained problems. *J. Sci. Comput.*, 58(2):431–449, 2014.

[17] Matthew J Leotta and Joseph L Mundy. Vehicle surveillance with a generic, adaptive, 3d vehicle model. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(7):1457–1469, 2011.

[18] Yan Li, Leon Gu, and Takeo Kanade. Robustly aligning a shape model and its application to car alignment of unknown pose. *IEEE Trans. Pattern Anal. Mach. Intell.*, 33(9): 1860–1876, 2011.

[19] Yen-Liang Lin, Vlad I Morariu, Winston Hsu, and Larry S Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*, pages 466–480. Springer, 2014.

[20] Varun Ramakrishna, Takeo Kanade, and Yaser Sheikh. Reconstructing 3d human pose from 2d image landmarks. In *ECCV*, pages 573–586, 2012.

[21] Jason Saragih. Principal regression analysis. In *CVPR*, pages 2881–2888. IEEE, 2011.

[22] Jason M Saragih, Simon Lucey, and Jeffrey F Cohn. Deformable model fitting by regularized landmark mean-shift. *Int. J. Comput. Vis.*, 91(2):200–215, 2011.

[23] Peter H Schönemann. A generalized solution of the orthogonal procrustes problem. *Psychometrika*, 31(1):1–10, 1966.

[24] Leonid Sigal and Michael J Black, Predicting 3d people from 2d pictures. In *Articulated Motion and Deformable Objects*, pages 185–195. Springer, 2006.

[25] Chunyu Wang, Yizhou Wang, Zhouchen Lin, Alan L Yuille, and Wen Gao. Robust estimation of 3d human poses from a single image. In *CVPR*, pages 2369–2376, 2014.

[26] Zaiwen Wen and Wotao Yin. A feasible method for optimization with orthogonality constraints. *Math. Program.*, 142(1-2):397–434, 2013.

[27] Stephen J Wright. Coordinate descent algorithms. *Math. Program.*, 151(1):3–34, 2015.

[28] Xuehan Xiong and Fernando De la Torre. Supervised descent method and its applications to face alignment. In *CVPR*, pages 532–539, 2013.

[29] Yin Zhang. Recent advances in alternating direction methods: Practice and theory. In *IPAM workshop*, 2010.

[30] Zhaoxiang Zhang, Tieniu Tan, Kaiqi Huang, and Yunhong Wang. Three-dimensional deformable-model-based localization and recognition of road vehicles. *IEEE Trans. Image Process.*, 21(1):1–13, 2012.

[31] Xiaowei Zhou, Spyridon Leonardos, Xiaoyan Hu, and Kostas Daniilidis. 3d shape reconstruction from 2d landmarks: A convex formulation. In *CVPR*, pages 4447–4455, 2015.

[32] M Zeeshan Zia, Michael Stark, Bernt Schiele, and Konrad Schindler. Detailed 3d representations for object recognition and modeling. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2608–2623, 2013.