3D Recognition and Pose Estimation Using Model Reconstructed by Pairwise Correspondence

Lin Tian
Beihang University, Beijing, China
Email: alice.lintian@hotmail.com

***Abstract*—This paper focuses on pose estimation of 3D object from 2D images. The base system of Structure from Motion is applied. A 3D model with feature points and descriptors is prebuilt. 3D-2D matching is employed in recognition and pose estimation process. The method of pairwise correspondence is proposed. To reduce reprojection error of the model, a weighting step is added. Viewpoint change of camera is simulated by affine transformation in 3D model description. Simulated affine feature is modified to keep the model compact. In experiment section, the performance of the pose estimation system as well as the modified affine feature are tested, and the results are compared to related previous research.**

***Index Terms*—pose estimation, 3D model, 3D recognition, affine feature**

I. Introduction

The recognition of 3D object is an important and complex problem. Recognition is usually treated with hypothesis verification method depending on a reconstructed 3D model in literature [1]-[8]. Model based recognition systems build correspondence between features in testing images and those representing a well trained 3D model to find the category or pose of interest object in testing image. Features extracted for 3D-2D matching could be corners, interest points or image contour [9]-[11]. Related investigations are centred on two categories: geometric feature based recognition [7], [12]-[15] and local feature matching based recognition [5], [16], [17]. For the first one, shape features such as Fourier or wavelet descriptors [12], [13] are extracted and matched between training images and testing images. Dense sampling of view space is required and the pose of object can only be approximated by the closest training image. In [7], aspect-graph approach to 3D object recognition is employed. Images from arbitrary views are assigned different aspects according to the criteria of local monotonicity and aspect distinctiveness. Ref. [9] relies on the importance of contour in human visual attention. It identifies object by ordering constant Curvature Contour Primitives (CCP) generated by simple edge filter and linking methods. But difficulties are met when dealing with cluttered scene or inconspicuous contours. As other boundary-based recognition methods, CCP maps only solve very limited viewpoint changes.

The idea behind shape-based recognition is straightforward. But the lack of features representing 3D information restricts the performance and application of recognition. In addition, since shape is the only descriptor, foreground and background must be segmented accurately. The two reasons have made it inapplicable for robotic manipulation, industrial inspection and places where high precision and the ability of dealing with cluttered scene are required [3], [7], [11], [18]. A more efficient and precise method is to estimate pose by matching local features of an image to a prebuilt 3D model. Model of interest objects can be built by laser scanner [19], [20], as well as stereo system [1], [2], [16], [21], [22]. The former results in high precision that can be applied to defects detection of automobile surface [19], but is equipment demanding and time consuming. Stereo system method is more effective when a larger estimation error is acceptable, for example the manipulation of a household robot [3].

Local image region matching is widely used in recognition, 3D reconstruction, image registration and robotic manipulation. Matching procedure contains two steps: region detection and description. Regions as well as descriptors are represented by keypoints. Detection locates regions with distinctive texture or gradient distribution. Each region is assigned a descriptor invariant to some image transformations. Harris [10] uses second moment matrix to locate interest region. The DoG detector in SIFT [11] is image scaling, rotation and translation invariant. While rotation and translation can be normalized, no detector is fully scale invariant. The common solution is to simulate images with different scale parameters. For instance, SIFT creates octaves and levels of images blurred by Gaussian filter with different variance. In some cases, parts are chosen to represent local image feature. The 3D-2D matching becomes index matching between parts. Ref. [6] creates a codebook of local appearance surrounding Harris points and matches testing image patches to the codebook. Ref. [5] extends parts matching by introducing the concept of canonical view. Object model is defined by the canonical parts $\{P_{ch}, P_{ck}, ...\}$ and their geometrical relations $\{H_{chk}, ...\}$. Object

Manuscript received December 6, 2014; revised December 14, 2015.

doi: 10.18178/lnit.3.2.65-69

class label is determined by matching pair of candidate parts to pair of canonical views in each model. The matched model that yields the least residual error is assigned to the testing image. Pose estimation is achieved by finding the dominant canonical view by comparing residual error. Transformations, such as scale and rotation, of candidate parts are computed with respect to canonical views.

In this paper, a 3D model consisting of weighted points is proposed. Instead of Structure from Motion (SfM) [1]-[3], the model is created by pairing 3D point sets. Each set is reconstructed from two images. For 2D to 3D matching, simulated affine features are added to the SIFT descriptors from original training images and demonstrate that our modified affine feature could represent viewpoint change by longitude. Pose estimation is regarded as a PnP problem [23] in this paper. Comparison experiment on pose estimation is performed with and without affine features. The paper is organized as follows. In Section II, a 3D model is prebuilt for interest object. Our method of weighted 3D model is detailed. In Section III, recognition and pose estimation is achieved using rigid and 3D-2D perspective transformations. Modified affine features are used to simulate features of images when viewpoints deviate by a large angle to training images. Section IV presents experimental results on recognition rate, impact of modified affine features and pose estimation. In Section V, potential applications of the recognition and pose estimation method is explored. Future work to improve recognition rate and precision is mentioned in brief.

II. 3D Weighted Model

A. 3D Model from Pairwise Merging

3D model building problem can be solved by SfM in previous works [1]-[3]. The problem of reconstruction is formulated as a non-linear least square minimization of the reprojection errors over all calibration parameters, camera positions and world coordinates. In this paper, 3D model of interest object is reconstructed through merging 3D point sets derived from neighboring pairs of images. The idea behind pairwise correspondence is similar to [5]. In [5], a set of paired parts across training images is derived by grouping algorithm. Canonical views $\{P_{ch}, P_{ck}, ...\}$ and their geometrical relations $\{H_{chk}, ...\}$ are used to represent 3D object. Instead of part matching, 3D point sets are merged from nearby pairs and the 3D structure of the object is reconstructed.

As the first step of reconstruction, SIFT feature is taken to find 2D-2D matching within a pair since it has advantages over many other local descriptors [24]. Training images in the same pair differ in viewpoint by no more than 20 degrees where the correct matching ratio is above 80% [11].

The camera frame, Oc, of the first image in training sequence is regarded as the world frame, Ow. 3D point sets are alighed from every image pairs to Ow. Each point set is constructed from a pair of images by epipolar constraint:

$$x_i^T F_{ij} x_j = 0 \tag{1}$$

Ransac [25] is used to guarantee that correspondences are correctly matched inliers. For neighboring pairs $\{I_{i-1}, I_i\}$, $\{I_i, I_{i+1}\}$ connected by the same image I_i, two sets of feature points $\{x^{(i-1)}_{i-1}, x^{(i-1)}_i\}$[1], $\{x^{(i)}_i, x^{(i)}_{i+1}\}$ are used for pairwise correspondence. The intersection of $\{x^{(i-1)}_i\}$ and $\{x^{(i)}_i\}$, denoted as $\{xc_i\} = \{x_i : x_i \in \{x^{(i-1)}_i\} \cup \{x^{(i)}_i\}\}$, merges the corresponding 3D points in both sets together, while non-intersection points are transformed by triangulation. The pairwise merging algorithm is detailed in Table I.

TABLE I. Pairwise Correspondence Algorithm

Input: xi, 2D coordinates of extracted feature points in image Ii;
n, the number of training images;
$\{x^{(i)}_i, x^{(i)}_{i+1}\}$, i=1, 2, ..., n-1, correct matches of image pair $\{I_i, I_{i+1}\}$;
$\{X^{(i)}\}$, i=1, 2, ..., n-1, 3D coordinates of point set reconstructed from pair $\{I_i, I_{i+1}\}$;
Output: $\{Xm\}$, 3D coordinates aligned to world frame Ow;

Begin

$i := 1; \{Xt^{(1)}\} := \{X^{(1)}\}$

while $i < n-1$

$i := i+1; \{xc_i\} = \{x_i : x_i \in \{x^{(i-1)}_i\} \cup \{x^{(i)}_i\}\}$

$\{Xtc^{(i-1)}\} := \{X : X \in \{Xt^{(i-1)}\}, X \leftrightarrow xc_i\}$[2]

$\{Xc^{(i)}\} := \{X : X \in \{X^{(i)}\}, X \leftrightarrow xc_i\}$

$\min_{s,R,T} norm\left(s \cdot (R \cdot Xc^{(i)} + T) - Xtc^{(i-1)}\right)$

$Xt^{(i)} := s \cdot (R \cdot X^{(i)} + T)$

end

$\{Xm\} := \{Xt^{(i)}\}, i = 1, 2, ..., n-1$

End

B. Weighting 3D Points

Each reconstructed 3D point is weighted by its reprojection error. Given that $\{X^{(i)}\}$ is reconstructed from $\{x^{(i)}_i, x^{(i)}_{i+1}\}$, the reprojection error as well as the weight is composed of two parts:

$$error(i,k) = err1(i,k) + err2(i,k) \tag{2}$$

$$weight(i,k) = \exp(error(i,k)) \tag{3}$$

where $err1(i,k) = (P_1^{(i)} \cdot X^{(i)}(k) - x^{(i)}_i(k))^2$, $err2(i,k) = (P_2^{(i)} \cdot X^{(i)}(k) - x^{(i)}_{i+1}(k))^2$, $P_1^{(i)}$ and $P_2^{(i)}$ are 3×4 camera matrix, $k = 1, 2, \ldots, n^{(i)}$, $n^{(i)}$ is the number of correct matches in pair $\{I_i, I_{i+1}\}$.

For a merged point Xw(j), i.e. the weighted average of 3D points reconstructed repeatedly from more than one pair, the confidence of the point is in proportion to how many time it is reconstructed as well as the reprojection error of each reconstructed 3D coordinate. Correspondence between Xw(j), j = 1, 2, …, N, and

[1] For $x^{(i-1)}_i(k)$, the upper index (*i*-1) denotes the pair $\{I_{i-1}, I_i\}$, the lower index *i* means that the 2D point is extracted from the I_i and *k* means it is the kth point in set $\{x^{(i-1)}_i\}$.

[2] $X \leftrightarrow xc_i$ denotes 3D-2D point correspondence.

$\{Xt^{(i)}(k)\}$ is defined as $ind(i,k) = j$. N is the number of 3D model points after alignment and merging.

$$Xw(j) = \frac{\sum_{i,k} Xt^{(i)}(k) \cdot weight(i,k)}{\sum_{i,k} weight(i,k)} \quad (4)$$

$$W(j) = \max(weight(i,k)) \cdot (1 + e^{(-\lambda \cdot std(\{Xt^{(i)}(k)\}))}) \quad (5)$$

In (4) and (5), $Xt^{(i)}(k)$ satisfies the constraint of $ind(i,k) = j$, λ is taken as 2.5×10^3 empirically.

The idea of weighting 3D points is to reduce the confidence of points with relatively high reprojection error, and to regard points corresponding to multiple pairs as more reliable. It is demonstrated that the model error is reduced after the weighting process in Section IV.

III. 3 Recognition and Pose Estimation

A. 2D-3D Feature Matching

Image distortion introduced by viewpoint change can be approximated by affine transformation [26] where camera distortion is neglected. Affine map depicts tilt of a planar fairly well, but not tilt of 3D object with complex structure. However, in practice, affine invariant algorithms [2], [3] have derived results sufficient to deal with image matching problem where viewpoint change is within a certain range. With the same idea of scale invariant features, viewpoint change is simulated by image affine transformation. One thorny issue is that it would usually introduce tens of thousands of new descriptors to the original 3D model. Ref. [3] deals with the problem by quantization and [2] facilitates matching by BBF method.

In our implementation, training images are taken to be $\theta = 0$, at an interval of about 15 degrees by longitude. Tilts of $t = \{\sqrt{2}, 2\}$ corresponding to $\theta = \{45, 60\}$ are performed on each training image, to simulate the whole view space. To keep the 3D model compact, descriptors corresponding to 2D points not extracted in the original training images are discarded. The remaining affine transformed descriptors are added to 3D model. After the viewpoint change simulation, number of descriptors for one 3D point may increase, while the number of 3D points, *N*, does not change. 3D point cloud and camera positions are shown in Fig. 1.

For 3D-2D matching, the ratio test [11] of the nearest and second nearest matches between descriptors of 3D model and those of testing image is used.

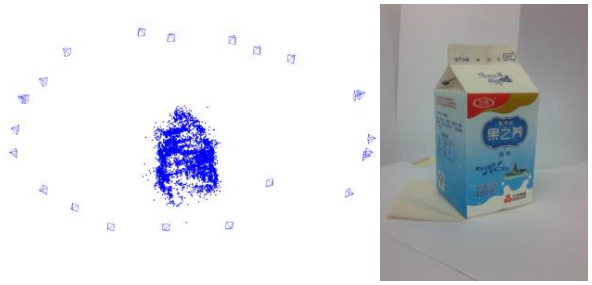


Figure 1. Reconstructed 3D milk bottle and one out of 24 training images.

B. Pose Estimation

Ransac and Gradient Descent Optimization [27], [28] are combined to estimate intrinsic and extrinsic parameters. Reprojection error of weighted model points is used as correct matching criterion. Optimization minimizes the reprojection error as:

$$\min_{K,R,T} \sum_j W(j) \cdot (K \cdot (R \cdot Xw(j) + T) - s \cdot xt(j)) \quad (6)$$

where *Xw* is 3D model point coordinate, *xt* is 2D testing image point coordinate, *K* is camera calibration matrix, *R* and *T* are the rotation and translation respectively.

IV. Experimental Results

The performance of 3D weighted model and modified affine feature are tested in this section. Recognition rate and pose estimation are evaluated. Algorithm codes are in Matlab and C++ and run on a PC with 2GHz CPU.

A. 3D Weighted Model

Reprojection error is evaluated on 24 training images. The error decreases by a large margin due to weighting process in Fig. 2. It is reduced by 50.32% at $\lambda = 2.5 \times 10^3$. The summation of reprojection error of $\{Xt^{(i)}(k)\}$ is the objective function in optimization. For training images, the error of reconstructed points is enlarged by merging $\{Xt^{(i)}(k)$:ind$(i,k) = j\}$ into one model point $Xw(j)$ and it decreases with λ exponentially as in (5).

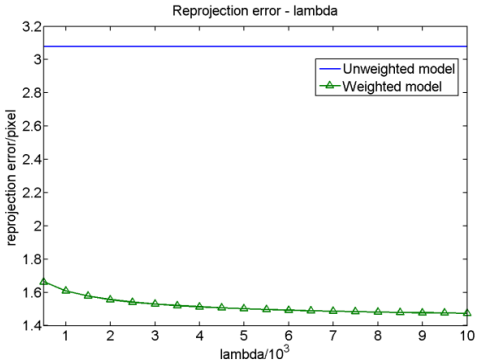


Figure 2. Average reprojection error of 3D model as a function of λ. Both weighted model and unweighted model are evaluated.

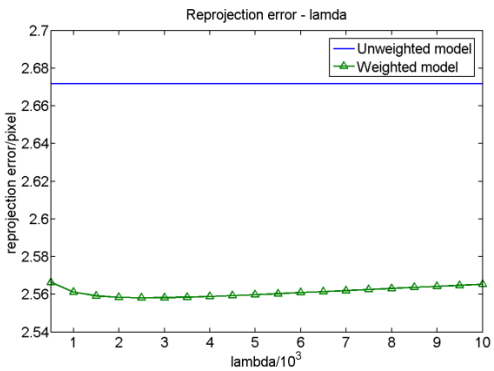


Figure 3. Average reprojection error of 12 testing images as a function of λ. Both weighted model and unweighted model are evaluated.

Reprojection error evaluation is applied to 14 testing images, 15 pose estimations are implemented for every

image. The average reprojection error of all 3D-2D correspondences with respect to λ is shown in Fig. 3. The introduction of weighted 3D model has greater effect than the variation of λ. In following experiments, weighted model are applied for pose estimation and choose λ to be 2.5×10^3, the value that yields the least error. Reprojection error of model point is reduced by 4.12% at $\lambda = 2.5\times10^3$.

B. Correct Recognition Rate, Pose Error and Translation Error

Table II compares the recognition rate, average rotation and translation error under different testing conditions. The threshold for correct detection is set as 5 cm and 22.5 degrees, the same with [3]. Rotation error is calculated as the quaternion angle of testing image to ground truth. Pose estimation is performed 195 times in total for testing images at large *t* and 180 times for small *t*. The recognition rate lies between correct detection rates of SA and SA+Q in [3].

TABLE II. RECOGNITION RATE, AVERAGE ROTATION AND TRANSLATION ERROR WITH RESPECT TO GROUND TRUTH

	recognition rate	R error (degree)	T error (cm)
not affine and *t*=0	0.8556	4.1407	1.5822
affine and *t*=0	0.8444	4.3976	1.4946
not affine and large *t*	0.7625	4.4745	1.5574
affine and large *t*	0.8328	3.7578	1.3220

The presence of affine feature has little impact on testing images at $t = 0$. However, at large *t*, the introduction of modified affine features boosts recognition rate. The rate is close to that at $t = 0$, which indicates that the simulation method in Section III.A is effective to describe 3D viewpoint change. It is also noticed that rotation error and translation error are reduced by 6.3% and 15.1% respectively. Pose and position accuracy is also increased.

C. Recognition Rate with Respect to Acceptable Rotation Angle Error

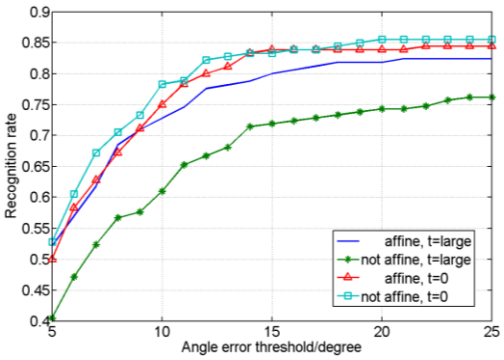


Figure 4. Recognition rate of 3D model with and without affine features, and testing images at different t. The angle error threshold varies from 5 to 25 degrees.

The recognition rate in Fig. 4 is higher at *t*=0 than at large t. One reason is that a lot more features are extracted from training images than from affine transformed images. More correct 3D-2D correspondences can be found for testing images at *t*=0. The addition of affine features is unnecessary and brings more mismatches which account for recognition rate decrease by a small margin (data points represented by squares and triangles).

For arbitrary t, the pose estimation problem in real world, the recognition rate is higher for 3D model with affine features. The augmented 3D model descriptors contain more correct correspondences to testing images that deviate from *t*=0 by a large angle.

V. CONCLUSION AND FUTURE WORK

The 3D modeling method proposed in this paper has two major differences from existing ones. First, the pairwise correspondence is robust since it does not depend on optimization. However, in the usual camera motion tracking problems, all camera parameters and world point coordinates need to be determined by minimization of non-linear least square problem. Second, the weighting process in Section II.B could describe the 3D model more precisely. It reduces reprojection error of model point by 50.32% at $\lambda = 2.5\times10^3$ in Fig. 2. Error of testing image points also decreases by 4.12% on average at the same λ.

To simulate viewpoint change, the model descriptors are augmented with SIFT features from affine transformed images. Only features corresponding to points that already exist in the 3D model are retained. Number of points in 3D model is not augmented. SA is modified by leaving out the dependence of epipolar constraint and searching to locate simulated affine features in nearby views [3]. The feasibility of this modification are verified by affine transformation and pose estimation experiments in Section IV.B, IV.C.

Compared to state-of-the-art 3D modeling algorithm [2], [3], our pairwise correspondence is more robust but not as accurate. The 3D point set reconstructed from a pair of images has negligible error. Most of the error comes from matching step. One of our future tasks is to improve pairwise matching and weighting, especially the former, to reduce model error.

In Section IV.D, special attention is paid to recognition rate with respect to angle error. Our plan is to apply the result of pose estimation to determine position of an interest region relative to the whole rigid body. Possible applications could be robotic manipulation and non-tactile component inspection.

REFERENCES

[1] R. Szeliski and S. B. Kang, "Recovering 3D shape and motion from image streams using non-linear least squares," in *Proc. CVPR*, 1993, pp. 752-753.

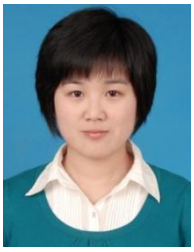
[2] I. Gordon and D. G. Lowe, *What and Where: 3D Object Recognition with Accurate Pose, Toward Category-Level Object Recognition*, Springer Berlin Heidelberg, 2006, pp. 67-82.

[3] E. Hsiao, A. Collet, and M. Hebert, "Making specific features less discriminative to improve point-based 3D object recognition," in *Proc. CVPR*, June 2010, pp. 2653-2660.

[4] M. Weber, "Unsupervised learning of models for object recognition," PhD dissertation, California Institute of Technology, Pasadena, California, 2000.

[5] S. Savarese and F. F. Li, "3D generic object categorization, localization and pose estimation," in *Proc. ICCV*, 2007, pp. 1-8.

[6] B. Leibe, A. Leonardis, and B. Schiele, "Combined object categorization and segmentation with an implicit shape model," in *Proc. Workshop on Statistical Learning in Computer Vision*, May 2004.
[7] C. M. Cyr and B. B. Kimia, "3D object recognition using shape similiarity-based aspect graph," in *Proc. ICCV*, 2001, pp. 254-261.
[8] A. Thomas, V. Ferrari, B. Leibe, T. Tuytelaars, B. Schiele, and L. V. Gool, "Towards multi-view object class detection," in *Proc. CVPR*, 2006, pp. 1589-1596.
[9] R. Bergevin and J. F. Bernier, "Detection of unexpected multi-part objects from segmented contour maps," *Pattern Recognition*, vol. 42, no. 11, pp. 2403-2420, 2009.
[10] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. Alvey Vision Conference*, 1988, pp. 15-50.
[11] D. G. Lowe, "Distinctive image features from scale-invariant keypoints," *IJCV*, vol. 60, no. 2, pp. 91-110, November 2004.
[12] G. C. H. Chuang and C. C. J. Kuo, "Wavelet descriptor of planar curves: Theory and applications," *IEEE Trans. on Image Processing*, vol. 5, no. 1, pp. 56-70, Jan. 1996.
[13] P. Wunsch and A. F. Laine, "Wavelet descriptors for multiresolution recognition of handprinted characters," *Pattern Recognition*, vol. 28, no. 8, pp. 1237-1249, 1995.
[14] D. W. Leng and W. D. Sun, "Contour-Based iterative pose estimation of 3D rigid object," *IET Computer Vision*, vol. 5, no. 5, pp. 291-300, 2011.
[15] S. Belongie, J. Malik, and J. Puzich, "Shape matching and object recognition using shape contexts," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 24, no. 4, pp. 509-522, April 2002.
[16] D. G. Lowe, "Three-Dimensional object recognition from single two-dimensional images," *Artificial Intelligence*, vol. 31, no. 5, pp. 355-395, 1987.
[17] D. G. Lowe, "Local feature view clustering for 3D object recognition," in *Proc. CVPR*, 2001, pp. 682-688.
[18] A. Mian, M. Bennamoun, and R. Owens, "On the repeatability and quality of keypoints for local feature-based 3D object retrieval from cluttered scenes," *IJCV*, vol. 89, no. 2-3, pp. 348-361, 2010.
[19] F. Chen, G. M. Brown, and M. Song, "Overview of three-dimensional shape measurement using optical methods," *Optical Engineering*, vol. 39, no. 1, pp. 10-22, 2000.
[20] B. Drost, M. Ulrich, N. Navab, and S. Ilic, "Model globally, match locally: Efficient and robust 3D object recognition," in *Proc. CVPR*, June 2010, pp. 998-1005.
[21] M. Brown and D. G. Lowe, "Unsupervised 3D object recognition and reconstruction in unordered datasets," presented at Int. Conf. on 3-D Digital Imaging and Modeling, June 2005, pp. 56-63.
[22] R. V. Tsai and T. S. Huang, "Uniqueness and estimation of three-dimensional motion parameters of rigid objects with curved surfaces," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 6, no. 1, pp. 13-27, Jan. 1984.
[23] V. Lepetit, F. M. Noguer, and P. Fua, "EPnP: An accurate O(n) solution to the PnP problem," *IJCV*, vol. 81, no. 2, pp. 155-166, 2009.
[24] K. Mikolajczyk and C. Schmid, "A performance evaluation of local descriptors," *IEEE Trans. on Pattern Analysis and Machine Intelligence*, vol. 27, no. 10, pp. 1615-1630, 2005.
[25] M. Fenzi, R. Dragon, L. Leal-Taixe, B. Rosenhahn, and J. Ostermann, "3D object recognition and pose estimation for multiple objects using multi-prioritized RANSAC and model updating," in *Pattern Recognition*, Springer Berlin Heidelberg, 2012, pp. 123-133.
[26] J. M. Morel and G. Yu, "ASIFT: A new framework for fully affine invariant image comparison," *SIAM Journal on Imaging Sciences*, vol. 2, no. 2, pp. 438-469.
[27] M. I. A. Lourakis, "A brief description of the Levenberg-Marquardt algorithm implemented by levmar," Institute of Computer Science, Foundation for Research and Technology, vol. 11, 2005.
[28] K. Madsen, H. B. Nielsen, and O. Tingleff, "Methods for non-linear least squares problems," Informatics and Mathematical Modelling, Richard Petersens Plads, 1999.



Lin Tian was born in Hebei Province, P.R. China, 1988. She received the BS degree in applied physics from East China University of Science and Technology, Shanghai, China in 2011 and the MS degree in instrumentation science and technology from Beihang University, Beijing, China in 2014.
She is an Algorithm Engineer in Novatek research center in Shanghai. She had been working on image segmentation, camera calibration and human face detection. Her interest is in pattern recognition, machine learning and intelligent robot.