Single-Image 3D Pose Estimation for Texture-Less Object via Symmetric Prior

Xiaoyuan REN\textsuperscript{(*)}, Libing JIANG\textsuperscript{(**)}, Xiaoan TANG\textsuperscript{1}, Nonmembers, and Junda ZHANG\textsuperscript{1}, Student Member

SUMMARY Extracting 3D information from a single image is an interesting but ill-posed problem. Especially for those artificial objects with less texture such as smooth metal devices, the decrease of object detail makes the problem more challenging. Aiming at the texture-less object with symmetric structure, this paper proposes a novel method for 3D pose estimation from a single image by introducing implicit structural symmetry and context constraint as priori-knowledge. Firstly, by parameterized representation, the texture-less object is decomposed into a series of sub-objects with regular geometric primitives. Accordingly, the problem of 3D pose estimation is converted to a parameter estimation problem, which is implemented by primitive fitting algorithm. Then, the context prior among sub-objects is introduced for parameter refinement via the augmented Lagrange optimization. The effectiveness of the proposed method is verified by the experiments based on simulated and measured data.

key words: 3D, pose estimation, priori-knowledge, texture-less object

1. Introduction

To understand 3D objects from 2D images is a fundamental issue in machine vision, which is crucial for the machine to learn the visual scene. 3D object pose estimation is one of the important problems [1]. Numerous methods based on image feature matching have been proposed including SFM and stereo vision, which however are limited in some certain scenarios. On the one hand, plenty of image acquisitions from multiple views are often required, which is not preferred in many applications. On the other hand, it is often hard to deal with texture-less objects for these methods which depend on repeatable and stable image feature such as SIFT, SURF, etc. However, texture-less objects are common such as glass, box, or artificial metal space objects. In addition, image details of object always be lost in some degradation process. Note that many artificial objects can be represented as parts (sub-objects), which are often symmetric structures. On this basis, this paper aims to investigate the possibility of analyzing 3D configuration of these objects from a single 2D image.

Obviously, extracting 3D information from a single image is an ill-posed problem. Based on the prior information, recent studies usually work by a set of trained model-views taken in a range around the known model with different location and pose [2]. These methods succeed in estimating 3D information by limiting themselves to objects of specific category, which share a lot of common structure such as car and bicycle. However, for some cases, there are large topological changes even though belonging to the same category, such as satellites. This, actually, makes the problem more complex.

Many artificial objects can be seen as combination of sub-objects with regular geometric primitive, for example, plate, cube and cylinder. It’s more flexible to focus on sub-objects rather than the whole object. Based on this, a method is proposed to estimate 3D pose for artificial texture-less objects by using implicit prior information. This prior information consists of two parts: (1) Structural symmetry of sub-objects. Many artificial objects can be decomposed as symmetrical sub-objects, which can be represented as parameterized geometric primitive of 6-DOF; (2) Context constraints among sub-objects. For many artificial objects, there are constraint relationships among sub-objects such as parallelism, orthogonality, coplanarity, etc. These context constraints can be introduced as constraint terms in optimal equation for the refinement of the estimated parameters. Thus, the problem of 3D pose estimation is converted to the parameter estimation problem which becomes more accurate by introducing context constraint.

Recent works with similar ambitions as ours are modeling 3D objects from sketches, which generally require user to drag the primitives to the approximate intended position [3]. This is because they are weak to interpret the sketch which needs much cognitive endeavor. Manual interaction is not needed in this paper because the object is decomposed and matched with the corresponding primitive automatically.

It is to be noted that, although our method can obtain the explicit 3D pose for these objects with symmetric structure, obviously, there are innumerable possibilities for recovering 3D structures with actual size from a single image. Our method is devoted to showing how to better understand 3D information by using implicit priori-knowledge of object from a single image.

2. Preliminary

Although it is impossible to obtain the object’s actual 3D size from a single image, the pose of each sub-object can be determined given its symmetric characteristic. Taking rectangle as an example, it is proved as follows.

First of all, the coordinate systems used in this paper...
is defined as: the camera coordinate system (CCS) and the image coordinate system (ICS) shown in Fig. 1, where \( f \) is the camera focal length.

The projection of the rectangle in the image is a quadrilateral, and \( P_i \) is the \( i \)th vertex of quadrilateral as \((X_i, Y_i)\). \( P_i \) denotes the \( i \)th vertex of rectangle as \((x_i, y_i, z_i)\) for \( i = 1, 2, 3, 4 \). Due to perspective projection, \((x_i, y_i, z_i)\) can be regarded as \( \left( \frac{z_i X_i}{f}, \frac{z_i Y_i}{f}, z_i \right) \). To simplify the calculation, only the rotation with \( X \) axis is considered. Hence two sides of rectangle, \( P_1P_2 \) and \( P_2P_3 \), are parallel to the \( ZY \) plane, where \( P_iP_j \) is the line connecting two points. The rotation angle \( \alpha \) between rectangle and \( X \) axis can be regarded as rotation angle of \( Y \) axis to \( P_1P_2 \) or \( P_2P_3 \). Due to the \( P_1P_3 \) as \( \frac{z_1 X_1 - z_3 X_3}{f}, \frac{z_1 Y_1 - z_3 Y_3}{f}, z_1 - z_3 \), \( \alpha \) is defined as follows:

\[
\alpha = \arctan \left( \frac{z_1 Y_1 - z_3 Y_3}{f \cdot (z_1 - z_3)} \right)
\]

(1)

As can be seen, the rotation angle is related to the depth \( z_1 \) and \( z_3 \) of point so that the pose information cannot be obtained (for arbitrary shaped polygon) if there is no prior information. Due to its symmetric characteristic, the constraint of rectangular is introduced below.

\[
\begin{align*}
& P_1P_2 \parallel P_3P_4 \\
& P_1P_2 \perp P_1P_3
\end{align*}
\]

(2)

Where \( \parallel \) denotes the parallelism, and \( \perp \) denote the orthogonality. Consequently, (1) can be represented as

\[
\alpha = \arctan \left( \frac{X_2 Y_3 + 2X_1 Y_1 - X_1 Y_3 - X_2 Y_1 - X_1 Y_1}{2f \cdot (X_3 - X_1)} \right)
\]

(3)

It can be seen that the 3D pose of the object is no longer related to the depth \( z \), but only relevant to the 2D projection in image after the symmetric prior introduced.

3. Problem Statement

The problem studied in this paper can be described as:

\[
T = \Pi S
\]

(4)

where \( S \) is the unknown 3D shape of object, and \( T \) denotes its projection in the 2D image. \( \Pi \) is the camera calibration matrix. It is impossible to reconstruct \( S \) from \( T \) only depending on image. This problem can be solved by introducing priori knowledge at two levels.

Firstly, object can be represented as a combination of sub-objects with symmetric characteristics shown as follows

\[
S' = \sum_{i=1}^{N} S_k = \sum_{i=1}^{N} \{P_k, \theta_k\}
\]

(5)

Where \( S_k \) for \( k \in [1, N] \) represents the \( k \)th primitive. \( P_k \) is the corresponding primitive of \( S_k \) and \( \theta_k \) denotes its parameters, which contains shape parameters as well as 3D pose parameters with 6-DOF. For example, \( \theta \) of rectangle includes its width, height, 3D position in space and rotation angles with three-axis. So the problem can be depicted by

\[
\min_{\theta_1, \theta_2, \ldots, \theta_N} D(T, \Pi S')
\]

(6)

Where \( D(\cdot) \) denotes the deviation between sub-object in image and primitive projection. Then \( T \subset \mathbb{R}^2 \) can be decomposed as \( \{T_1, T_2 \cdots T_N\} \) in image, and (6) is equal to

\[
\min_{\theta_1, \theta_2, \ldots, \theta_N} D \left( \sum_{k=1}^{N} T_k, \Pi \sum_{k=1}^{N} \{P_k, \theta_k\} \right)
\]

(7)

After the correspondence relationships are determined between \( \{T_1, T_2 \cdots T_N\} \) and \( \{S_1, S_2 \cdots S_N\} \), (7) is equivalent to the following

\[
\min_{\theta_1, \theta_2, \ldots, \theta_N} \sum_{k=1}^{N'} D(T_k, \Pi \{P_k, \theta_k\})
\]

(8)

Since the sub-objects are independent of each other, the parameters of each primitive can be obtained respectively by

\[
\left\{ \min_{\theta_k} D(T_k, \Pi \{P_k, \theta_k\}) \right\}_{N'}
\]

(9)

Finally, considering that the sub-objects which compose the whole object shares the same context information, the context constraint is introduced as the following

\[
\min_{\theta_1, \theta_2, \ldots, \theta_N} \left( \sum_{k=1}^{N'} D(T_k, \Pi \{P_k, \theta_k\}) + H(S_1, S_2 \cdots S_N) \right)
\]

(10)

Where \( H(S_1, S_2 \cdots S_N) \) contains the context information, which are the geometric relationships between different sub-objects. The initial value of the above optimization problem is produced from (9).

4. Method

4.1 Structural Symmetry of Sub-Objects

The primitives considered in this paper are plate, cuboid and cylinder. So \( S_k \) in (5) can be regarded as

\[
S_k = \{S_r, S_c, S_c\}
\]

(11)

which stands for plate, cuboid and cylinder respectively.
Firstly, shape of object \( T \subset \mathbb{R}^2 \) in image can be decomposed as sub-objects \( \{ T_1, T_2, \ldots, T_N \} \) given cut set \( C = \{ c_1, c_2, \ldots, c_{N-1} \} \) where \( T_i \cap T_j \in C \) or \( T_i \cap T_j = \emptyset \) for \( 1 \leq i, j \leq N \). We use method in [4] to conduct shape decomposing. During the fitting processing, the edge of sub-object is used because edge features always offer a good invariance to illumination changes and image noise and are particularly suitable with texture-less object. Edge of object is extracted using method of perceptual grouping in [5], then we obtain the edge of each sub-object as \( \{ T_1^c, T_2^c, \ldots, T_N^c \} \) on the basis of shape decomposing result.

Secondly, edge of each sub-object is fitted with corresponding primitive to obtain \( \theta_k \) shown as (9). The type of primitive depends on the feature of part edge. When the image features are not enough to judge, we try all the possibilities and take the best. To perform fitting, an objective function \( F_k \) for each part is constructed and minimized with the updating of \( \theta_k \). The fitting process is as follows.

Primitives are initially centered on the \( z = z_0 \) plane. In subsequent processing, the relative placement of primitives is determined by context constraints.

### 4.2 Context Constraints

There are always constraints among the primitives that belong to the same object, especially the symmetric object. Take the spacecraft as an example. Generally, the main body is coaxial, and solar panels are symmetrical on the sides of the main body. Under these context constraints, the relative placement of primitives can be refined by global optimization. Anchor points are defined that control the position and pose of primitives, and the direction vector of primitive is the unit vector between anchor points. We detect and impose the geometric relations as follows.

**Parallelism.** This is applied if the angle between two direction vectors is less than 15 degrees. The constraint is that the two normal vectors \( n_1, n_2 \) must satisfy \( n_1 \times n_2 = 0 \).

**Orthogonality.** This is applied if the angle between two direction vectors is between 75 and 105 degrees, and \( n_1, n_2 \) must satisfy \( n_1 \cdot n_2 = 0 \).

**Coplanarity.** Two pairs of anchors are defined as \( a_1^1, a_1^2, a_2^1 \) and \( a_2^2 \). This is applied if the distance between each point and the plane formed by the remaining three points is within 5% of the size of primitive. The constraint is that \( (n_1, n_2, (a_1^1 - a_1^2)) = 0 \).

<table>
<thead>
<tr>
<th>Table 1</th>
<th>Algorithm of primitive fitting</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Input:</strong></td>
<td>( T_i^c, \ S_i )</td>
</tr>
<tr>
<td><strong>Output:</strong></td>
<td>( \theta_i )</td>
</tr>
<tr>
<td>1 Initialize ( \theta_i ) ;</td>
<td></td>
</tr>
<tr>
<td>2 While not converged do</td>
<td></td>
</tr>
<tr>
<td>update ( \theta_i ) .</td>
<td></td>
</tr>
<tr>
<td>calculate ( F_k = \sum a(c_i, E_i)^t ), where ( c_i ) is point of ( T_i^c ), ( E_i ) is projection of ( S_i ), and ( a(c_i, E_i) ) denotes distance between ( c_i ) and ( E_i ).</td>
<td></td>
</tr>
<tr>
<td>End</td>
<td></td>
</tr>
</tbody>
</table>

Coaxial. From two pairs of anchors, this is applied if any three points can be formed with a triangle containing angle more than 170 degrees. The constraint is that \( n_1 \times n_2 = 0 \) and \( n_1 \times (a_1^1 - a_1^2) = 0 \).

The geometric relationships between each primitive and other primitive are detected after fitting. \( \psi \) includes the constraints met with conditions above. The problem to be solved is global optimization as follows.

\[
\begin{align*}
\min F &= \sum_{i=1}^{N} \omega_i F_i \\
\text{subject to} : \ &\psi_j(S_1, \ldots, S_N) = 0, \ j = 1, \ldots, L
\end{align*}
\]

where \( \omega_i \) is the weight of the primitive determined by the surface area of the primitive, and \( \psi_j(S_1, \ldots, S_N) \) includes geometric constraints. We use augmented Lagrange method [6] to solve this optimization problem.

### 5. Experiment

Experiment is conducted for measuring and simulated images. Firstly, we analyze the simulated image of spacecraft image with detail process. As a category of valuable target, spacecrafts are texture-less objects with image degeneration in ground-based observation. Figure 2 (a) shows the simulated image.

Figure 2 (b) denotes the shape decomposition. The edge of each sub-object shown in Fig. 2 (c) is fitted to the corresponding primitives respectively. The results of sub-object fitting and 3D model generation are as follows.

The Median Angular Error (MAE) is used to evaluate the error of pose estimation, which represents the mean deviation with three-axis. The result of pose estimation of spacecraft is shown as Table 2.

Then, the experimental results of some common objects are shown below.

The result of pose estimation is shown below, where \( \theta \),

![Fig. 2 Shape decomposition result of spacecraft](image-url)
The results of sub-object fitting and 3D model generation of spacecraft. Four images above are the fitting results, and two images below show the 3D model represented by primitives in different views.

Fig. 4 The 3D model generation results of some common objects. The left side are the original images, and the right sides are 3D models represented by primitives.

$\alpha$ and $\gamma$ are angles of each primitive with three-axis.

It can be seen that 3D model of object is generated and pose estimation is completed. The result shows that the method performs better for object consisted of multi primitives than single primitive. The factor of pose error of our method mainly lies in the edge extraction. The complete and smooth edge of object can make the result more accurate. It is also noteworthy that the introduction of primitive constraints can reduce the pose error.

Because method proposed depends on the accurate edge extraction of object, the performance suffers from the complex environment. In addition, pose estimation becomes difficult when the sub-objects overlap at certain viewpoints. In future, we will do further research on enhancing the robustness of sub-object detection.

6. Conclusion

This paper proposes a method to estimate 3D pose from a single image for the texture-less object with symmetric structure. To solve this problem, the object is represented as a combination of sub-objects, with two levels of priori-knowledge being introduced. One is structural symmetry of sub-objects, each of which is represented as parameterized regular geometric primitive. The other is context information which is introduced via constraint terms in augmented Lagrange optimization. In experiment, the method’s ability of estimating the object’s 3D pose in a certain range of error is proved. The future direction of this work is to increase the types of primitives so as to better describe the objects, and the robustness should be enhanced for sub-object detection.

References


