Pose Estimation of GIS Pipeline Based on Spatial Transformation Network

Gong Zhang1,*, Boming Li2, Ti Liu1, Wenhan Chen1, Yichang Wang1
1State Grid Zhejiang Electric Power Co., Ltd Construction Company, 310000, Hangzhou, Zhejiang, China
2Zhejiang Electric Transmission and Transformation Co. Ltd, 310000, Hangzhou, Zhejiang, China

*Corresponding author e-mail: 5043289@sgzjep.org

Abstract. The manual flange structure alignment between GIS pipelines in the power system is inefficient and difficult to accurately align. To solve this problem, combined with the research results in the field of deep learning named spatial transformation network, a new pose estimation method based on single camera is proposed. In view of the high similarity between the moving flange and the static flange at the pixel level, the spatial transformation network is used to find the pixel mapping relationship of the two flange images. Thereby establishing the mapping relationship between the pixel coordinates of the two flange images and then using multiple points. In the perspective method, the pixel coordinates are mapped to the world coordinates to obtain the estimation of the position of the key point in the flange, and then the direction vector of the flange is calculated according to the position of the key point. Since there is a pixel coordinate transformation relationship between the static flange and the movable flange. Only the position of the key point in the static flange can be inversely solved by measuring the position of the key point in the static flange. Experiments show that, compared to the traditional method of measuring flange pose based on instrument measurement and linear regression, the method proposed in this paper can accurately measure the pose of the flange structure. And it can rely as little as possible on the measurement of the key points of the moving flange by the instrument.

Keywords: Pose Estimation, Flange Pose, Gis Pipeline, Spatial Transformation Network, Coordinate Transformation

1. Introduction
The estimation of the pose (position and direction) of the target object plays an important role in production practice. This article takes the docking of GIS pipelines in the power system as the background to discuss the estimation of the pose of the pipeline with flange structure. GIS (Gas Insulated Switchgear) is a gas-insulated metal-enclosed switchgear, which is a form of high-voltage power distribution equipment. When laying long-distance GIS pipelines, each sub-pipeline is generally connected by a flange structure. However, if the flange connection between large pipelines is
completely dependent on manual operation, it is not only inefficient, but also difficult to accurately connect. In order to improve the accuracy and efficiency of flange docking, it is required to measure the position of the flange pipe to be docked as accurately as possible before docking [1].

Spatial Transformer Network (STN) provides new ideas to solve this problem. The spatial transformation network was proposed by Google’s DeepMind team in 2015. They establish a visual attention mechanism by building three modules: local network, grid generator, and sampler. It can correct the input image changes caused by translation, scaling, rotation, etc.. Due to its powerful correction mechanism, the spatial transformation network has been applied to various computer vision tasks since it was proposed. In this paper, the spatial transformation network is used to learn the mapping relationship of the coordinate system from the moving flange to the static flange, that is, the affine transformation between the two. The main contributions of this article are as follows:

(1) A new idea of flange pose estimation is proposed: the position of the flange is reversed from the pose of the static flange;
(2) Based on the spatial transformation network, a new transformation method from the static flange coordinate system to the movable flange coordinate system is proposed;
(3) Compared with the traditional method of flange pose measurement based on instrument measurement and regression methods (such as the least square method and random Hough transform), the method proposed in this paper can accurately measure the pose of the flange structure. And it can rely as little as possible on the measurement of the key points of the moving flange by the instrument [2].

2. Definition and reverse solution of flange pose

![Figure 1. The definition of moving flange, static flange and their pose.](image)

The precise butt joint of the flange requires that the end faces of the two flanges are parallel, the circle centers are coincident, and all the bolt holes on the flange end faces are also accurately aligned. Therefore, the pose of the flange can be defined as a triplet: \( P = (O, n, l) \). \( O \in \mathbb{R}^3 \) represents the center position of the flange end surface, \( n \in \mathbb{R}^3 \) represents the normal vector of the flange end surface, and \( l = O - o \in \mathbb{R}^3 \) represents the connection vector between the center of the flange end face \( O \) and the center of a bolt hole \( o \) around the flange. Note that the \( O \), \( n \) and \( l \) are all measured in the world coordinate system \( \{w: O_w=X, Y, Z\} \), which we call "absolute pose". Let \( P_m = (O_m, n_m, l_m) \) and \( P_s = (O_s, n_s, l_s) \) respectively represent the absolute poses of the moving flange and the static flange (as shown in Figure 1). The process of flange butt joint assembly is a process in which the posture adjustment mechanism drives the moving flange to gradually realize the precise coincidence of the posture with the static flange. Therefore, for the posture adjustment mechanism, what needs to be known is the "relative posture" of the movable flange relative to the static flange. That is the
translation amount of the movable flange relative to the static flange \((T_x, T_y, T_z)\) in three directions \(X, Y, Z\), and the RPY angle \((a, b, g)\). Here, the RPY angle refers to the roll angle, pitch angle and yaw angle (Roll-Pitch-Yaw, RPY) of the moving frame relative to the reference frame. Therefore, we define the six-tuple \(R = (T_x, T_y, T_z, a, b, g)\) as the "relative pose" of the static flange. Obviously, it is easy to deduce the relative pose from the absolute pose of the static flange. This article focuses on the estimation of the absolute pose of the flange [3].

Since the position of the static flange is fixed, it can be considered that its pose is known a priori, and it provides a "target pose" for the moving flange. If the current pose of the moving flange can be solved inversely from the pose of the static flange, then an accurate measurement of the current pose of the moving flange can be obtained. Let \(A_1\) denote the transformation matrix from the static flange coordinate system to the moving flange coordinate system, the two key coordinate points \(O_m, o_m\) in the moving flange pose \(P_m\) and the sum of the corresponding static flange pose have the following relationship:

\[
\begin{bmatrix}
O_m \\
o_m
\end{bmatrix} = A_1 \begin{bmatrix}
O_s \\
o_s
\end{bmatrix}
\]

(1)

There is no direct mapping relationship between the normal vector of the moving flange \(n_m\) and the normal vector of the static flange \(n_s\), but it can be estimated by the position of the key points on the flange surface, which was discussed in detail by Hou et al. Therefore, the key to calculating the static flange pose from the dynamic flange pose is to calculate the transformation matrix \(A_s\). The next section will focus on how to learn the transformation matrix \(A_s\) through the network [4].

3. Coordinate transformation matrix learning based on space transformation network

3.1. Coordinate transformation matrix learning: from sampling principle to spatial transformation network

One of the most intuitive ways to learn the coordinate transformation matrix \(A_s\) is to mark several key points (number of key points is determined by the number of parameters in \(A_s\)) at the same position of the static flange and the movable flange (such as cylinder, end face, bolt hole, etc.). The positions of these key points are measured by optical instruments, and establish the parameters of the equation \(A_s\). However, this method requires that the key points measured on the two flanges must be completely consistent, which is very difficult in specific implementation. Hou provided another solution to solve the coordinate transformation matrix. First mark some key points on the two flanges, then measure the positions of all key points in the coordinate system of the static flange. Next, we use the dynamic method to measure the positions of all key points in the blue coordinate system. So the transformation equations of the two coordinate systems can be established, and the transformation matrix can be solved accordingly. However, the coordinate transformation matrix obtained by this scheme cannot directly map the coordinate position in the static flange to the coordinate position of the corresponding point in the moving flange, and can only map the position of the key point in the moving flange to the static flange. It is representation in flange coordinate system [5].

Different from the above two schemes (the above scheme attempts to find the mapping relationship between the key points of the moving flange and the static flange), we start from the perspective of computer vision and regard the static flange and the moving flange as two images, seeking the connection between the two at the pixel level. Obviously, regardless of the shape of the pipe connected by the static flange and the movable flange, the static flange and the movable flange themselves are completely the same in size, shape and structure. Therefore, the pixels in a static flange image must
correspond to the pixels in a dynamic flange image, that is, the static flange image can be regarded as a re-sampling of the pixels in the dynamic flange image. If this "sampling" relationship can be found, the position of the corresponding pixel in the dynamic flange image can be inferred from the position of the pixel in the static flange image [6].

Let $F \in \mathbb{R}^{H \times W}$ and $F_m \in \mathbb{R}^{H \times W}$ respectively represent the static flange image and the dynamic flange image, and $X[i,j]$ represent the pixel value of the image $X$ at the pixel point $(i,j)$. Then, $\forall x \in [1,2,\ldots,H], \forall y \in [1,2,\ldots,W]$, the sampling process from $F_m$ to $F$ can be expressed as:

$$F[x,y] = f_{\text{samp}}(F_m, (x_m, y_m); f_x, f_y) = \sum_{j_m \in [1,H]} \sum_{i_m \in [1,W]} F_m[i_m,j_m] \cdot k(x_m - i_m f_x) \cdot k(y_m - j_m f_y). \quad (2)$$

Among them, $k(\cdot)$ is the kernel function used for image interpolation, and $f_x, f_y$ are its parameter. And $(x_m, y_m)$ is the coordinate of the pixel in the moving flange image $F_m$ that is closest to the pixel value $F[x,y]$ in the static flange image $F$. That is, formula (2) indicates that $F[x,y]$ is obtained by "sampling" $F_m[x_m, y_m]$ and surrounding pixels, and the neighboring pixels $(x_m, y_m)$ are determined by the sum of the parameters $f_x, f_y$ of the interpolation kernel function $k(\cdot)$. Therefore, the key to simulating the sampling process from to by (2) is to establish the mapping relationship between $(x_m, y_m)$ and $(x, y)$. Since the static flange image can be obtained by rotating and translating the moving flange image, this mapping relationship can be established by the following affine transformation:

$$\begin{pmatrix} x_m \\ y_m \end{pmatrix} = T((x, y); q) = A \begin{pmatrix} x \\ y \\ 1 \end{pmatrix} = \begin{pmatrix} q_1 \\ q_2 \\ q_3 \end{pmatrix} \begin{pmatrix} x \\ y \\ 1 \end{pmatrix}. \quad (3)$$

The affine transformation parameters $q = \{q_i\}$ in the formula (3) can be learned through the Spatial Localization Network $f_{\text{Loc}}$:

$$q = f_{\text{Loc}}(F_m, f_{\text{Loc}}) \quad (4)$$

Among them, $f_{\text{Loc}}$ is the parameter of $f_{\text{Loc}}$.

Combining formulas (2), (3) and (4), the sampling process from the moving flange image $F_m$ to the static flange image $F$ can be further expressed as:

$$F[x,y] = f_{\text{samp}}(F_m, T((x, y); f_{\text{Loc}}(F_m, f_{\text{Loc}})); f_x, f_y) \quad (5)$$

This sampling process is actually the Spatial Transformer Network (STN) proposed by Jaderberg et al. in 2015, as shown in Figure 2(a). Under normal circumstances, the parameters of the interpolation kernel function $f_x, f_y$ can be preset, and the spatial transformation network needs to learn mainly "positioning parameters $f_{\text{Loc}}$". Therefore, the spatial transformation network can be defined as follows:

$$f_{\text{STN}}(F_m, (x, y); f_{\text{Loc}}) \triangleq f_{\text{samp}}(F_m, T((x, y); f_{\text{Loc}}(F_m, f_{\text{Loc}}))); f_x, f_y) \quad (6)$$


3.2. Constructing a spatial positioning network based on the expanded residual dense network

Since the spatial positioning network \( f_{Loc} \) is used to "locate" the position of the pixels \( F_s[x,y] \) in the static flange in the moving flange \( F_m \), it is very important for the spatial transformation network. However, it is difficult to achieve precise positioning using conventional convolutional neural networks or fully connected networks. In order to achieve precise positioning, we convert the precise positioning problem into a precise reconstruction problem, that is, if \( F_s \) can be accurately reconstructed by a neural network \( F_m \), then precise positioning can be achieved. Therefore, this article uses the cutting-edge method "Residual Dense Network (RDN)" in the field of image reconstruction to define \( f_{Loc} \), as shown in figure 2(b). Among them, the connection marked by the symbol "\( \oplus \)" is the residual connection, which means that the two connected feature maps are added and operated. The connection not marked with "\( \oplus \)" is the feature map splicing, which means that the splicing of several connected feature maps in the channel direction. Conv means the convolutional layer, the number marked in the parentheses of each convolutional layer indicates the size of the convolution kernel and the expansion factor (the number after the "\( \ast \)" symbol is the expansion factor, and the expansion factor of the convolutional layer without the "\( \ast \)" mark is 1 by default). After each convolutional layer, there is a ReLU activation layer. For simplicity, the ReLU layer is omitted in Figure 2(b). Since most of the image content contained in the static flange and the moving flange image are very similar, the use of residual connection can make the spatial positioning network \( f_{Loc} \) learn as much as possible the difference between the characteristics of the static flange image and the moving flange image; Dense connections can effectively reuse features, thereby greatly saving the scale of network parameters [7].

Saving the scale of network parameters is particularly important for solving the problem of spatial positioning from static flange image to moving flange image in this article. This is because the spatial transformation network in this article only involves two end-to-end images: static flange image and dynamic flange image. Flange image is a typical small sample problem. Using a network that is too deep or the parameter scale is too large can easily lead to overfitting. Therefore, in addition to using dense connections, the spatial positioning network \( f_{Loc} \) in this paper (Figure 2(b)) uses a shallower fully convolutional network with only 7 layers. Research in the field of image reconstruction has shown that the size of the receptive field of the network model has a greater impact on the reconstruction performance of the network, and a shallower network will greatly reduce the receptive field of the network. In order to increase the receptive field without increasing the network parameters, this paper further introduces the hole convolution in the spatial positioning network \( f_{Loc} \), as shown in Figure 2(b), where the dilated convolution is only used in the middle three layers. This is because the extracted features need to be refined in the first 2 layers and the end 2 layers of the network. The use of dilated convolutions with dilation factors greater than 1 in the four convolutional layers will lead to the iterative process of the network unstable [8].

3.3. Loss function

It is worth noting that the original spatial transformation network is usually embedded in other network models (such as classification networks) as an attention mechanism. So there is no independent loss function for it. In order for the spatial transformation network to be suitable for the problem in this article, a new loss function needs to be established for it. Using the spatial transformation network \( f_{STN} \) constructed in this article, for each pixel in a given static flange \( F_s[x,y] \), it can be sampled from the moving flange:

\[
\hat{F}_s[x,y] = f_{STN}(F_m(x,y); f_{Loc})
\]
Compared $\hat{F}[x,y]$ and true $F[x,y]$, the following optimization formulas $f_{\text{loc}}$ for learning positioning parameters can be established:

$$f_{\text{loc}} = \arg \min_{\omega} \sum_{i=1}^{n} \sum_{x,y} \ell(F[x,y], f_{\text{local}}(F_{\omega}(x,y)))$$

(8)

Among them, $\ell(\cdot)$ is the loss function, which $\ell_1$ can be norm, norm $\ell_2$ or Charbonnier penalty function. Among them, the Charbonnier penalty function is defined as:

$$\ell(x_i, x_j) = \sqrt{(x_i - x_j)^2 + \epsilon^2}.$$ 

Because the Charbonnier penalty function is differentiable and robust to noise, this paper uses it as the loss function.

3.4. Transformation from image plane coordinates to world coordinates based on PnP method

After training to obtain the optimal parameters $f_{\text{loc}}$ of the network, the optimal affine transformation parameters $q^*$ can be obtained from equation (4). And then substituting $q^*$ into equation (3) to calculate the position of any pixel in the static flange in the moving flange. It is worth noting that the position obtained at this time is the position in the image plane (the pixel coordinates), which is two-dimensional and the key point that needs to be measured is the position in the world coordinate system, which is three-dimensional. Therefore, it is also necessary to consider the mapping from a two-dimensional image plane to a three-dimensional world coordinate system. To solve this problem, you can use the PnP method. The PnP method considers the positions of points $n \geq 3$ in the target object in the world coordinate system $\{w:O_w - X_w, Y_w, Z_w\}$, camera coordinate system $\{c:O_c - X_c, Y_c, Z_c\}$ and pixel coordinate system $\{p:O_p - X_p, Y_p\}$. Let $\{P_w, P_{e,i}, P_{v,i}\}_{i=1}^{n}$ denote the position of a point in the target object $n$ in the three coordinate systems. For any point in the flange, its coordinates in the world coordinate system $P_{w,i}$ and the camera coordinate system have the following mapping relationship:

$$\begin{bmatrix} \hat{P}_{e,i}^T \\ \hat{P}_{v,i}^T \end{bmatrix} = \begin{bmatrix} R & T \\ 0 & 1 \end{bmatrix} \begin{bmatrix} P_{w,i}^T \\ 1 \end{bmatrix}$$

(9)

Using the PnP algorithm, the sum of equation $R$, $T$ in $\{P_{w,i}, P_{v,i}\}_{i=1}^{n}$ can be obtained from the three-dimensional world coordinates and two-dimensional pixel coordinate pairs. After $R$ and $T$ are obtained, the equation (6) can be inversely transformed, and calculate the coordinates in the world coordinate system from the coordinates in the camera coordinate system.

When using the PnP algorithm to solve equation (6), real-time and fault tolerance need to be considered. The existing PnP algorithms are mainly divided into iterative methods and non-iterative methods. The iterative method is slower and more sensitive to parameter initialization, so it is easy to fall into the local optimum. The non-iterative method has a small amount of calculation, but it is sensitive to noise and easily interfered by noise. In order to balance real-time and fault tolerance, this paper uses the simple, robust and fast Perspective-n-Point (SRPnP) method recently proposed by Wang to solve equation (6). SRPnP first simplifies the PnP problem to the solution of a 7th-order element polynomial and a 4th-order element polynomial, using the least square residual method for rough solving. And then using the Gauss-Newton iterative method for fine tuning. Because SRPnP uses non-iterative methods, it is effectively combined with the iterative method. So it can take into account real-time and fault tolerance.

4. Conclusion

This paper discusses the pose estimation problem of GIS pipeline with flange structure. Flange structure has a wide range of applications in the butt joint of GIS pipes in power systems. It is not only
inefficient to install flanges manually but also difficult to accurately align. In order to solve this problem, accurate measurement of flange pose is needed. Since the static flange is fixed and its pose is easy to obtain, this paper proposes a new idea of flange pose estimation which is the pose of the static flange is reversely resolved from the pose of the static flange. For this reason, this article uses the similarity between the static flange and the moving flange, and uses the spatial transformation network in the deep learning field to establish the mapping relationship from the moving flange to the static flange coordinate system, that is: between the two Affine transformation between. Experiments show that, compared to the traditional method of measuring flange pose based on instrument measurement and regression methods (such as the least square method and random Hough transform), the method proposed in this paper can not only accurately measure the pose of the flange structure, but also can rely as little as possible on the measurement of the key points of the moving flange by the instrument. How to measure the position of the key points that are not visible to the camera (such as the key points on the inner wall of the bolt hole of the flange) more accurately, and how to accelerate the model optimization speed of the spatial transformation network are topics that need further research.

Acknowledgments
Fund projects: In this paper, the State Grid Zhejiang Electric Power Co., LTD. "Intelligent manipulator in substation engineering research and application" (5211JJ18000K) subsidizes.

References
[1] Liu S, Xu D, Zhang D, et al(2014). High precision automatic assembly based on microscopic vision and force information. IEEE Transactions on Automation Science and Engineering, vol.13, no.1, pp: 382-93.
[2] Yamamoto H, Sano T(2002). Study of micromanipulation using stereoscopic microscope [J]. IEEE Transactions on Instrumentation and Measurement, vo.51, no.2, pp: 182-7.
[3] Yu YK, Wong KH, Or SH, et al(2008). Robust 3-D motion tracking from stereo images: A model-less method [J]. IEEE Transactions on Instrumentation and Measurement, vol.57, no.3, pp: 622-30.
[4] Peng J, Xu W, Yuan H(2017). An efficient pose measurement method of a space non-cooperative target based on stereo vision. IEEE Access, vol.5, no.22, pp: 344-62.
[5] Meng C, Li Z, Sun H, et al.(2018). Satellite pose estimation via single perspective circle and line [J]. IEEE Transactions on Aerospace and Electronic Systems, vol.54, no.6, pp: 3084-95.
[6] Chen Fei(2007). Research on the Development and Application of GIS Equipment; Zhejiang University.
[7] Guan Yonggang, Cai Yuanji, Liu Weidong, et al.(2015). The relationship between the rise of the shell potential caused by the operation of the GIS isolation switch and the VFTO. High Voltage Technology, vol.41, no.12, pp: 4097-104.
[8] Zhang P C, Xu D, Zou W, et al.(2013). Positioning cylindrical target based on three-microscope vision system [J]. IEEE/ASME Transactions on Mechatronics, vol.19, no.5, pp: 1612-24.