Stud Pose Detection Based on Photometric Stereo and Lightweight YOLOv4

Xuan Zhang, Guohui Wang

School of Opto-Electronic Engineering, Xi’an Technological University, Xi’an 710021, China
*Corresponding author: Xuan Zhang, spritexuan7@163.com

Received Month X, XXXX | Accepted Month X, XXXX | Posted Online Month X, XXXX

Abstract: There are hundreds of welded studs in a car. The posture of a welded stud determines the quality of the body assembly thus affecting the safety of cars. It is crucial to detect the posture of the welded studs. Considering the lack of accurate method in detecting the position of welded studs, this paper aims to detect the weld stud’s pose based on photometric stereo and neural network. Firstly, a machine vision-based stud dataset collection system is built to achieve the stud dataset labeling automatically. Secondly, photometric stereo algorithm is applied to estimate the stud normal map which as input is fed to neural network. Finally, we improve a lightweight YOLOv4 neural network which is applied to achieve the detection of stud position thus overcoming the shortcomings of traditional testing methods. The research and experimental results show that the stud pose detection system designed achieves rapid detection and high accuracy positioning of the stud. This research provides the foundation combining the photometric stereo and deep learning for object detection in industrial production.

Keywords: Stud pose, Photometric stereo, Neural Network, Machine vision

1 Introduction

Stud is widely used in the modern machine building industry because of its high interchangeability [1]. There are hundreds of welded studs in a car and these studs are used for interior assembly in the car body. Whether the position of the welded studs meets the design requirements not only determines the subsequent assembly, but also affects the performance of the vehicle directly. It is necessary to detect the poses of all studs in a car for quality control during the modern industrial automation production. Coordinate Measuring Machine (CMM)[2], it cannot be adapted accordingly to different objects, and its material in probe damages the surface of the measured target easily. What's more, the speed of CMM is far from meeting the demand of more efficient measurement in higher precision. Recently, with the continuously development of computer technology, machine vision is widely used for 3D measurement of objects[3], [4] due to its advantages of non-contact, fast speed and high accuracy, so that researchers prefer studying non-contact measurements for objects. There are three types of non-contact measurement methods:
acoustic[5], optical[6] and electromagnetic methods[7], of which the optical 3D measurement is the most widely applied. Conventional optical measurement systems are Laser scanner[8], Laser radar[9], structure light scanner[10], monocular vision[11-13], multi-view stereo vision [14-15] and so on. Recently, neural networks have shown to superior performance in many object-detection tasks due to its ability to learn from raw data automatically[16]. There are many kinds of networks in 3D object detection[17]-[20]. However, there are few studies in studs pose detection by machine vision and networks. In other words, the defects (e. g., large lens distortions, focal blur, heavy noise and extreme poses) of the stud images limited the stud pose detection using only neural networks. Wu et al.[21] developed a novel method based on monocular vision for measuring the weld studs pose. Liu et al.[22] proposed a stud measurement system based on photometric stereo vision and Histogram of Oriented Normal (HON) feature extractor. Studies above have been limited in detecting stud poses due to the fact that there has a highly variant reflection property in studs.

Photometric stereo[23], an emerging technology estimating normal maps under different illuminations, has been extensively applied for precision improvement in object measurement combined with deep learning[24]-[27]. Photometric stereo uses normal maps to evaluate the 3D measurement which contains more accurate information than 2d images and possesses lower cost. For this reason, more and more researchers dedicate to the combination with photometric stereo and deep learning for 3D reconstruction and 3D measurement, however, there are few studies for object detection. Liu et al[28] implemented optical measurements of studs through normal vector map estimation and heat map training.

On the basis of these studies, this paper proposes the method for stud pose detection based on photometric stereo and neural network. The main contributions in this work are threefold:

1. The monocular vision is applied to calculate the coordinate parameters of the camera for calibration, which can achieve the stud dataset labeling automatically.

2. Photometric stereo algorithm is applied to estimate the stud normal map which as input is fed to the neural network.

3. The lightweight YOLOv4 network is improved to locate the stud by analyzing the normal map images in studs, which directly process normal maps and output prediction results with multi prediction size.

The structure of the rest of this paper is as follows: Section 2 provides basic methods in automatic labeling of stud datasets, estimating normal maps and building neural network; Section 3 presents the detailed experiments; In Section 4, data and the experimental results are presented; and Section 5 draws the conclusions of this work.

2 Basic Method

Combining photometric stereo and deep learning, as shown in Fig.1, we first build a photometric stereo vision system and a machine vision measurement system to capture images of studs under 8 different light sources (LED lights). We calculated the closed solution in camera calibration to obtain the internal and external parameters of the camera. Then, we derive the image pixel coordinates of studs in the images by Harris corner point detection algorithm[29] for automatically labeling the studs. Secondly all stud images are processed by the light vector pseudo-inverse matrix to
obtain the normal maps of the studs, which as the training images are input to the neural network. Finally, all the training images and the corresponding labels (ground truth) are input to the neural localization network for iterative training and testing to achieve the pose detection of studs. As long as the nominal position of the stud is accurate, the pixel coordinates of the top and bottom center points of the studs agree with the nominal position of the studs.

**Fig. 1** Stud pose detection system

**A. Monocular Vision-based dataset construction**

Fig. 2 illustrates the interrelationship between the point P in 3D space and its corresponding point p in the image, which contains coordinate transformation in four coordinate systems. These four coordinate systems are the world coordinate system, camera coordinate system, image coordinate system and pixel coordinate system respectively. As shown in Fig. 2, The 3D right-angle coordinate system $O_w - X_w Y_w Z_w$ is the world coordinate, which can be set artificially; The 3D right-angle coordinate system $O_c - X_c Y_c Z_c$ is the camera coordinate, $O_c$ is located at the projection center of the lens, the $Z_c$ axis is perpendicular to the image sensor and coincides with the optical axis of the lens, the $X_c$ axis is parallel to the long side of the image sensor image array, and the $Y_c$ axis is determined by the right-hand rule; The 2D right-angle coordinate $o - uv$ is the image coordinate, $o_c - uv$ is the pixel coordinate [30].

**Fig. 2** Dataset collection system based on monocular vision

The relationship of the transformance between the world coordinates and the pixel coordinates is expressed as
This article has been accepted for publication in a future issue of this journal, but it is not yet the definitive version. Content may undergo additional copyediting, typesetting and review before the final publication.

Citation information: Xuan Zhang, Guohui Wang, Stud Pose Detection Based on Photometric Stereo and Lightweight YOLOv4 (2021), DOI: https://doi.org/10.37965/jait.2021.12005

\[
\begin{bmatrix}
  u \\
  v \\
  1
\end{bmatrix} = \begin{bmatrix}
  f_x & 0 & u_0 & 0 \\
  0 & f_y & v_0 & 0 \\
  0 & 0 & 1 & 0
\end{bmatrix} \begin{bmatrix}
  R_{3 \times 3} & T_{3 \times 1} \\
  0_{1 \times 3} & I_{1 \times 1}
\end{bmatrix} \begin{bmatrix}
  X_w \\
  Y_w \\
  Z_w \\
  1
\end{bmatrix}
\]

(1)

where \( f_x = \frac{f}{dx}, f_y = \frac{f}{dy} \). \( f \) is the focal length of the lens, \( dx, dy \) are the physical dimensions of a pixel in \( x \)-axis and \( y \)-axis respectively. As the external parameters of the camera, \( R \) and \( T \) are the rotation matrix and the translation vector respectively. By equation (1), the parameters of the camera are obtained for camera calibration. On the basis of which, we construct the stud datasets. The details in dataset construction as follows:

a. Calculating the internal and external parameters of the camera for the camera calibration.

b. Calculating the pixel values of the top and bottom center points of the stud in the image coordinate from the 3D coordinate of the stud.

c. Labeling the stud by the image coordinate and defining the bottom center point of the stud as studb, the top center point as studt.

d. Feeding the pixel coordinates of the studs as ground truth to the neural network

**B. Photometric Stereo system**

Photometric stereo is a method to obtain local normal maps in several images under different illuminations. This paper applies 8 LED lights with different orientations for improving the accuracy and robustness of the result. The complexity of the threads on the stud surface and the soot from welding lead to a more pronounced diffuse reflection of the stud itself, so the photometric stereo vision system is established based on the Lambertian reflection.

According to the Lambertian reflection, the intensity of any pixel \( p(x,y) \) in the image can be expressed as

\[
I(x,y) = \rho(x,y)(S \cdot \hat{n}_{(x,y)})
\]

(2)

where \( I(x,y) \) is the pixel intensity under \( i \)-th illumination in row \( x \) and column \( y \), \( \rho \) is the albedo at the corresponding point of pixel \( p(x,y) \), \( S_i = [s_{x_i}, s_{y_i}, s_{z_i}]^T \) denotes the direction of light source projection. The equation (2) can be formulated as

\[
n(x,y) = \rho(x,y)\hat{n}(x,y) \cdot \hat{s} = g(x,y) \cdot \hat{s}
\]

(3)

\[
g(x,y) = \rho(x,y)\hat{n}(x,y)
\]

(4)

where \( \hat{n} \) is the surface unit normal vector, which is estimated by applying 8 LEDs and calculating the pseudo-inverse matrix of the light source vectors in this research. On the basis of which, the equation (3) can be described as:

\[
S^T \cdot I = \rho \cdot S^T \cdot S \cdot n
\]

(5)

The surface normal \( \hat{n} \) of pixel \( p(x,y) \) can be estimated:

\[
\hat{n}(x,y) = \frac{g(x,y)}{\rho(x,y) = \left\| g(x,y) \right\|}
\]

(6)
Finally, calculating every pixel through equation (6) repeatedly for normal map.

C. Deep Network-based Normal Map of Studs in Localization

In this paper, Lightweight YOLOv4 network based on YOLOv4[31] is proposed to locate the stud by analyzing the normal map images in studs. As shown in Fig.3, the size fed to the network is 608×608×3, where 3 indicates the three channels. The lightweight YOLOv4 network applies convolution layers, up sampling, down sampling and deep concatenation layers to directly process normal maps and output prediction results with multi prediction size. The multi-size output contains three kind of sizes: 76×76×24, 38×38×24, 19×19×24, which can get better network performance in extracting important features from the training data. Root Mean Squared Error (RMSE) is used as the regression loss function during the training process.

\[ \text{RMSE} = \sqrt{\frac{1}{N} \sum_{i=1}^{N} (y_i - \hat{y}_i)^2} \]  

(7)

where \( y_i \) is the ground-truth value, \( \hat{y}_i \) is the predicted value, \( N \) is the number of the testing samples of stud normal maps.

3 Experiments

a. Dataset and Experimental Platform

In this study, a total of 5000 groups of samples for studs are constructed. We apply the software MTLAB to program the microcontroller program Arduino to ensure that the LEDs are lit in the clockwise from the number 1 in Fig.2 for capturing stud images. Every group of the stud sample contains 8 stud images from different illuminations. These images are calculated by photometric stereo to obtain the normal maps of studs, which as the input is fed to neural network for training.

The hardware server configuration for the experiments is: Intel(R) Core (TM) i5-9600KF processor, NVIDIA GeForce GTX 2080Ti graphics card. The software environment is Ubuntu 18.04, python 3.7.7, TensorFlow-gpu-2.1.0, PyCharm 2020.1. Proposed method in this paper utilizes several libraries such as NumPy, Pillow, OpenCV and MATLAB.

b. Normal Maps of Studs

We estimate the vector maps of studs by the least square algorithm based on photometric stereo. Eight stud images of the same stud pose with different illuminations are integrated into a stud vector map. The normal maps of studs are obtained by converting the channels of the stud vector map. This paper displays the normal map of one stud pose in Fig.4.
Fig. 4 Normal map diagram for a stud pose

### c. Evaluation Metrics

In this paper, the RMSE and mAP (mean Average Precision) are used to evaluate the model.

\[
\text{precision} = \frac{TP}{TP + FP} \quad (8)
\]

RMSE suggests the precision of the measurement, which indicates the overall difference between the predictions and the ground truth for all testing samples. TP stands for true-positive, FP for false-positive, mAP as an important evaluation metrics is used to evaluate the accuracy of object detection.

### d. Network Training

The network is trained on Adaptive moment estimation (Adam) method, which possesses a very fast convergence rate and powerful generalization ability with optimization. Mosaic and Image augmentation (Imgaug) are applied to expand the stud dataset with a total of 30,000 data samples. All the labeled data (corresponded with ground truth) are randomly divided into the training and testing datasets with the ratio of 4:1. During the network training, the epoch and batch size of the training data are set to 60 and 4 respectively. The weights of the Pascal VOC (Pascal Visual Object Classes) are used as the initial weight input. The learning rate given an initial value with 0.001 is updated every 2500 iterations.

### 4 Results

The training loss curves are shown in Fig. 5. The numbers on x-axis and y-axis represent the training epochs and the loss values respectively. Fig. 5 (a) shows the trend of loss values, (b) and (c) show the distribution of loss values in the locally enlarged region of figure (a) respectively. Fig. 5 indicates that the loss function converges rapidly at the beginning of the training with oscillating decrease in the followed training. After 60 epochs of the network training, the loss value is 13.9392 (unnormalized) and the prediction result performs best.

The weights perform best trained in network are used to predict the stud pose. Fig. 6(a), (b) illustrates the prediction result in stud normal map images and stud images captured by the camera directly. It is obvious that the neural network provides a good performance in detecting stud normal map with its key points on the top and bottom of the stud. However, the raw image of the stud is detected incorrectly under the complex background shown in Fig. 6(c), (d). Fig. 6 (c) and (d) show that the top and bottom key points of the stud are not accurately recognized, or are arbitrarily recognized as other key points, or are not recognized. RMSE and mAP in proposed network is 0.074% and 99.65% respectively, a low error and high precision. In terms of detect speed for every stud image, our method requires less average computation time of 0.002584s, which indicates that the proposed method can be applied in a real production environment for stud real-time detection.
5 Conclusions

In this paper, a dataset system for automatically collecting and labelling studs was built. The photometric stereo in 8 light sources was applied to estimate stud normal map as input to improved neural network with good experimental results in stud poisoning. After the prediction, RMSE and mAP were used as the evaluation metrics to validate the prediction performance. A comparison of stud normal maps with stud raw images fed in network was made and suggested that proposed method indicated superior prediction performance. The conclusions in this paper are also applicable to multi-stud identification and detection. This research provides the foundation combining the photometric stereo and deep learning for object detection in industrial production. In future, the combination of deep learning and photometric stereo will be studied more intensively to improve the accuracy and speed of object detection.

![Fig.5 Loss trend during training](image1)

![Fig.6 Prediction of stud normal map image and stud raw images](image2)

(a) (b) (c) (d)

(a, b), prediction of stud normal map images; (c, d), prediction of stud raw images. The blue box, red box and green box indicate the position, the top point and the bottom point of stud respectively.
Acknowledgments
The work is partly supported by the Natural Science Basic Research Plan in Shaanxi Province of China (No. 2016JM6041).

References


