# Autonomous Removal of Perspective Distortion for Robotic Elevator Button Recognition

Delong Zhu, Jianbang Liu, Nachuan Ma, Zhe Min, and Max Q.-H. Meng

*Abstract*— Elevator button recognition is considered an indispensable function for enabling the autonomous elevator operation of mobile robots. However, due to unfavorable image conditions and various image distortions, the recognition accuracy remains to be improved. In this paper, we present a novel algorithm that can autonomously correct perspective distortions of elevator panel images. The algorithm first leverages the Gaussian Mixture Model (GMM) to conduct a grid fitting process based on button recognition results, then utilizes the estimated grid centers as reference features to estimate camera motions for correcting perspective distortions. The algorithm performs on a single image autonomously and does not need explicit feature detection or feature matching procedure, which is much more robust to noises and outliers than traditional feature-based geometric approaches. To verify the effectiveness of the algorithm, we collect an elevator panel dataset of 50 images captured from different angles of view. Experimental results show that the proposed algorithm can accurately estimate camera motions and effectively remove perspective distortions.

Fig. 1: A comparison between original image (left) and corrected image (right). The green diamonds indicate the centers of bounding box predicted by button detector. The red crosses represent the centers of a fitted grid on current image with 5 rows and 3 columns.

## I. INTRODUCTION

Autonomous elevator operation is regarded as a promising solution to inter-floor navigation of mobile robots, which is typically composed of three components: button recognition, motion planning, and robot control. As the most fundamental step, the performance of button recognition directly determines the success rate and robustness of the whole elevator operation system. In order to acquire robust recognition result, traditional approaches prefer leveraging fiducial markers to provide reference anchors, based on which, users have to manually calibrate elevator panel layout and figure out the geometric relationship of each button with respect to the markers. Obviously, such system configurations highly depend on human assistance and bring the users much inconvenience. For some applications, e.g., indoor robotic exploration [1]–[3], the elevators may not be accessible in advance and hence cannot be calibrated or operated.

To address the problems mentioned above, various button recognition algorithms [4]–[6] based on deep learning technique are proposed, which are typically performed on raw images that usually include severe perspective distortions. The key idea of this work is to autonomously remove such distortions to help improve button recognition accuracy. A typical button recognition algorithm usually contains two subtasks: button detection and character recognition [4], and the proposed distortion removal method can benefit both

The authors are with the Department of Electronic Engineering, The Chinese University of Hong Kong, Shatin, N.T., Hong Kong SAR, China. *email: {dlzhu, jbliu ncma, zmin, qhmeng}@ee.cuhk.edu.hk*

procedures. Firstly, the layout of elevator buttons usually follows certain patterns, e.g., a regular grid in most cases. Through fitting a grid for the elevator panel, some buttons that are not detected by the detector can be retrieved. Secondly, characters without perspective distortions can be easily identified by the recognizer, meanwhile, the mis-recognized characters can even be corrected by leveraging the rules of button layout [5].

The perspective distortion removal approach consists of two interdependent parts. The first part is a grid fitting algorithm, which models the grid centers as GMM centroids and leverages the Expectation-Maximization (EM) framework to find the optimal centroid positions regarding current detection results. This algorithm is developed based on [7], in which the images are assumed without distortions. Different from this work, we derive a closed-form solution in this paper and further extend it to handle perspectively distorted images. The second part is a pose estimation algorithm in the normalized image plane. This algorithm takes the optimized grid centers as reference points and calculates the rigid-motion that can align the centers of the predicted bounding boxes and the reference points. Through applying an inverse transformation, the perspective distortion can be removed, based on which, a new round of grid fitting and pose estimation can be performed to achieve more accurate distortion removal.

The contributions of this paper are summarized as follows:

- We propose a perspective distortion removal algorithm that can autonomously estimate distortion parameters using a single elevator panel image.

- We derive a closed-form solution for the maximization step of EM and successfully extend it to handle grid fitting problem on an image with perspective distortions.
- We collect and release an elevator panel dataset with ground-truth poses labeled on every image and initial experimental results are presented as baselines.

For the remaining of this paper, we first introduce some related work in Sec. II, then present some preliminary knowledge and the distortion removal algorithm in Sec. III and Sec. IV, respectively. After that, the experimental results and discussions are presented in Sec. V. At the end of this paper, in Sec. VI, we conclude the paper and present the future work.

## II. RELATED WORK

Before the wide application of deep learning techniques in robotics [8], researchers have already attempted to develop button recognition algorithms. Klingbeil *et al.* [7] design a pipeline that can perform button detection and character recognition. With the assumption that all the images are captured without any distortion, this method achieves an accuracy of $86.2\%$ in a testing set with 50 images. Some other approaches [9] [10] are proposed afterwards, but their recognition accuracy and robustness are not satisfying due to the limited capacity of traditional algorithms and the small scale of related datasets. In more recent years, deep learning based large-scale object recognition for robotics becomes a more promising research area. Meanwhile, there emerge an increasing number of publications on the topic of elevator button recognition [4]–[6], [11]–[13]. For instance, Liu *et al.* [6] leverage a detection network, the SSD [14], to conduct button recognition, which essentially regards the recognition as a multi-object detection problem and hence cannot handle a lot of button categories. To address this problem, Zhu *et al.* [4] propose to integrate a character recognition branch into the Faster-RCNN [15] [16]. In this way, the previous multi-object detection problem is converted to a binary button detection task and a character recognition task, which enables the network to handle hundreds of button classes.

Although there exist various deep learning based button recognition methods, the removal of perspective distortions is seldom mentioned. Nevertheless, such an idea is actually a widely-adopted idea in practice. In [17], the authors utilize a competitive learning method to autonomously remove false positives of pedestrian detectors. In [18], an encoded pattern is leveraged to help estimate the rigid-body transformations between the target and robot, which is actually a similar idea with this work, except that the reference points are encoded in advance there, whereas we have to estimate these points in this work. From a technical view, the methodologies used in this work share many similarities with point set registration [19]–[22] and the main difference lies in the parameters that need to be optimized. Point set registration focuses on solving two key problems: point correspondences and parameter optimization [23], which are also the cores of grid fitting algorithm.

## III. PROBLEM FORMALIZATION

Throughout this paper, the following notations are used:
- $\mathbf{X} = [\mathbf{x}_1, \cdots, \mathbf{x}_N] \in \mathbb{R}^{2 \times N}$ - the detected button centers in image plane,
- $\hat{\mathbf{X}} = [\hat{\mathbf{x}}_1, \cdots, \hat{\mathbf{x}}_N] \in \mathbb{R}^{3 \times N}$ - the detected button centers in normalized image plane,
- $\mathbf{U} = [\mathbf{u}_1, \cdots, \mathbf{u}_N] \in \mathbb{R}^{2 \times N}$ - the centroids of GMM model, which is determined by model parameters $\mathbf{\Theta}$,
- $\mathbf{z}$ - the hidden variable that adopts one-hot encoding,
- $\boldsymbol{o} = [o_x, o_y]^T$ - the origin of the fitted grid, center of the most top-left cell of the grid,
- $\Delta = [\Delta_x, \Delta_y]^T$ - the interval distance between grid centers in $x$ and $y$ directions in image plane,
- $i, j \in \mathbb{Z}^*$ - the index of grid cells in $x$ and $y$ directions, respectively, and there are $K$ pairs of $(i, j)$,
- $\mathbf{\Sigma} = \text{diag}([\sigma_x, \sigma_y])$ - the diagonal covariance matrix shared by all the GMM components,
- $\mathbf{\Theta} = (\boldsymbol{o}, \Delta, \mathbf{\Sigma})$ - the parameters of GMM,
- $\boldsymbol{\xi} = [\boldsymbol{\theta}, \boldsymbol{t}]^T$ - the rigid-body motion between bounding box centers and fitted grid centers,
- $\boldsymbol{R}(\boldsymbol{\theta})$ - the matrix representation of angle-axis parameterized rotation $\boldsymbol{\theta}$.

In this work, the GMM is leveraged to learn the best grid fitting, in which the detected button centers $\mathbf{X}$ is considered to be sampled from a mixture Gaussian as follows:

$$\mathbf{X} \sim \sum_{k=1}^{K} \pi_k \mathcal{N}(\cdot | \mathbf{u}_k, \mathbf{\Sigma}), \tag{1}$$

where $\pi_k$ is the coefficient of GMM and $\sum_{k=1}^{K} \pi_k = 1$. The probability of a single sample $\mathbf{x}_n$ is calculated as follows:

$$p(\mathbf{x}_n | \mathbf{\Theta}) = \sum_{\mathbf{z}} p(\mathbf{z}) p(\mathbf{x}_n | \mathbf{z}) = \sum_{k=1}^{K} \pi_k \mathcal{N}(\mathbf{x}_n | \mathbf{u}_k, \mathbf{\Sigma}), \tag{2}$$

where $p(\mathbf{z}) \in \{\pi_1, \cdots, \pi_k\}$ represents a discrete distribution over the hidden variable $\mathbf{z}$, which is a prior that directs the mixture process of Gaussian distributions. Following the conventions in point set registration literature [19], $p(\mathbf{z})$ is assumed to be a uniform distribution, i.e. $\pi_k = 1/K$, in this work.

To account for noises and outliers, an additional uniform distribution is integrated to GMM,

$$p(\mathbf{x} | \mathbf{\Theta}) = \frac{(1 - \alpha)}{C} + \alpha \sum_{k=1}^{K} \frac{1}{K} \mathcal{N}(\mathbf{x} | \mathbf{u}_k, \mathbf{\Sigma}), \tag{3}$$

where $\alpha$ is a balance parameter and we omit the subscript $n$ for simplicity. In order to make $p(\mathbf{x}_n)$ a distribution in the image plane, the following constraint should be satisfied,

$$\sum_{\mathbf{x}[1]=1}^{w} \sum_{\mathbf{x}[2]=1}^{h} p(\mathbf{x} | \mathbf{\Theta}) = 1, \tag{4}$$

where the operator $[\cdot]$ is used to fetch the corresponding elements of vector $\mathbf{x}$. The solution of Eq. (3) and (4) gives us an approximation of $C$, i.e., $C \approx w \times h$, where $w, h$ are the width and height of the image plane, respectively.

For simplicity, we add an additional dimension to $\mathbf{z}$, thus $\mathbf{z} \in \mathbb{R}^{K+1}$, and then use a more compact formula to denote the marginal distribution presented in Eq. (3),

$$p(\mathbf{x}|\boldsymbol{\Theta}) = \sum_{k=1}^{K+1} \pi'_k p(\mathbf{x}|\mathbf{z}), \tag{5}$$

where $\pi'_k = \alpha \pi_k$ for $1 \leq k \leq K$ and $\pi'_{K+1} = 1 - \alpha$. The joint distribution of $\mathbf{x}$ and $\mathbf{z}$ also plays an important role in estimating grid fitting parameters. Based on Eq. (2) and (3), this distribution can be expressed as follows:

$$p(\mathbf{x}, \mathbf{z}) = \begin{cases} \alpha \mathcal{N}(\mathbf{x}|\mathbf{u}_k, \boldsymbol{\Sigma})/K, & \text{if } \mathbf{z}[k] = 1, \\ (1-\alpha)/C, & \text{if } \mathbf{z}[K+1] = 1. \end{cases} \tag{6}$$

Given the predicted button centers $\mathbf{X}$, the optimal $\boldsymbol{\Theta}$ can be achieved by optimizing the following negative log likelihood function:

$$\begin{aligned} E(\boldsymbol{\Theta}) &= -\sum_{n=1}^{N} \ln p(\mathbf{x}_n|\boldsymbol{\Theta}) \\ &= -\sum_{n=1}^{N} \ln \sum_{k=1}^{K+1} \pi'_k p(\mathbf{x}|\mathbf{z}) \\ &= -\sum_{n=1}^{N} \ln \left[ \frac{(1-\alpha)}{C} + \alpha \sum_{k=1}^{K} \frac{1}{K} \mathcal{N}(\mathbf{x}|\mathbf{u}_k, \boldsymbol{\Sigma}) \right] \end{aligned} \tag{7}$$

where $\mathbf{u}_k = [o_x + \Delta x \cdot i, o_y + \Delta y \cdot i]^T$. Eq. (7) is also called *complete log likelihood*, but it is difficult to be directly optimized, as the summation is inside the $\ln$ function. A general framework for solving Eq. (7) is the EM algorithm, which essentially utilizes the *expected log likelihood* to iteratively optimize the complete log likelihood.

## IV. EM FOR DISTORTION REMOVAL

The EM framework contains two iteratively performed steps: E-step and M-step. In the E-step, we use current parameter values $\boldsymbol{\Theta}^{old}$ to calculate the posterior distribution of latent variable $p(\mathbf{z}|\mathbf{x})$,

$$p(\mathbf{z}|\mathbf{x}) = \frac{p(\mathbf{x}, \mathbf{z})}{p(\mathbf{x})} \propto p(\mathbf{x}, \mathbf{z}). \tag{8}$$

Denote a probability distribution table for $p(\mathbf{z}|\mathbf{x})$ and each element of this table records a probability value that is defined as follows:

$$\gamma(n, k) = p(\mathbf{z}[k] = 1|\mathbf{x}_n), \tag{9}$$

which actually specifies a type of soft correspondences between detected button centers and the grid centers.

In the M step, we use the posterior distribution to find the expectation of the complete log likelihood, i.e., the expected log likelihood, and maximize this log likelihood function to update the parameter $\boldsymbol{\Theta}$, which can be expressed as follows:

$$\begin{aligned} \boldsymbol{\Theta}^{\text{new}} &= \arg\max_{\boldsymbol{\Theta}} \mathcal{Q}\left(\boldsymbol{\Theta}, \boldsymbol{\Theta}^{\text{old}}\right) \\ &= \arg\max_{\boldsymbol{\Theta}} \sum_{n=1}^{N} \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}_n, \boldsymbol{\theta}^{\text{old}}) \ln p(\mathbf{x}_n, \mathbf{z}|\boldsymbol{\Theta}) \\ &= \arg\max_{\boldsymbol{\Theta}} \sum_{n=1}^{N} \sum_{\mathbf{z}} p(\mathbf{z}|\mathbf{x}_n, \boldsymbol{\Theta}^{\text{old}}) \ln p(\mathbf{x}_n|\mathbf{z}, \boldsymbol{\Theta}) \\ &= \arg\min_{\boldsymbol{\Theta}} \sum_{n=1}^{N} \sum_{k=1}^{K} \gamma_{nk}(\boldsymbol{e}_n^T \boldsymbol{\Sigma}^{-1} \boldsymbol{e}_n + \ln |\boldsymbol{\Sigma}|), \end{aligned} \tag{10}$$

where $\boldsymbol{e}_n$ is the distance between detected buttons center $\mathbf{x}_n = [x_n, y_n]^T$ and $k^{th}$ grid center indexed by $(i_k, j_k)$,

$$\boldsymbol{e}_n = \begin{bmatrix} x_n - (o_x + \Delta x \cdot i_k) \\ y_n - (o_y + \Delta y \cdot j_k) \end{bmatrix}. \tag{11}$$

The derivative of $\mathcal{Q}$ w.r.t. $\boldsymbol{\Theta}$ is as follows:

$$\begin{aligned} \frac{\partial \mathcal{Q}}{\partial \boldsymbol{\Theta}} &= (\frac{\partial \mathcal{Q}}{\partial \boldsymbol{e}})^T \frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\Theta}} \\ &= 2 \sum_{n,k} \gamma_{nk} (\boldsymbol{\Sigma}^{-1} \boldsymbol{e}_n)^T \frac{\partial \boldsymbol{e}_n}{\partial \boldsymbol{\Theta}}, \end{aligned} \tag{12}$$

$$\frac{\partial \boldsymbol{e}_n}{\partial \boldsymbol{\Theta}} = \begin{bmatrix} -1 & 0 & -i_k & 0 \\ \mathbf{0} & -1 & 0 & -j_k \end{bmatrix}. \tag{13}$$

Assume a diagonal covariance matrix and let Eq. (12) be zero, the following equations can be derived,

$$\begin{aligned} \sum_{n,k} \gamma_{nk} \sigma_x [x_n - (o_x + \Delta_x \cdot i_n)] &= 0, \\ \sum_{n,k} \gamma_{nk} \sigma_y [y_n - (o_y + \Delta_y \cdot j_n)] &= 0, \\ \sum_{n,k} \gamma_{nk} \sigma_x i_n [x_n - (o_x + \Delta_x \cdot i_n)] &= 0, \\ \sum_{n,k} \gamma_{nk} \sigma_y j_n [y_n - (o_y + \Delta_y \cdot j_n)] &= 0, \end{aligned} \tag{14}$$

Correspondingly, the closed-form solution of $\boldsymbol{\Theta}$ can be achieved,

$$\begin{aligned} o_x &= \frac{\sum_{n,k} \gamma_{nk}(x_n - \Delta_x \cdot i_k)}{\sum_{n,k} \gamma_{nk}}, \\ o_y &= \frac{\sum_{n,k} \gamma_{nk}(y_n - \Delta_y \cdot j_k)}{\sum_{n,k} \gamma_{nk}}, \\ \Delta_x &= \frac{\sum_{n,k} \gamma_{nk} i_k(x_n - o_x)}{\sum_{n,k} \gamma_{nk} i_k^2}, \\ \Delta_x &= \frac{\sum_{n,k} \gamma_{nk} j_k(y_n - o_y)}{\sum_{n,k} \gamma_{nk} j_k^2}, \\ \sigma_x &= \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk} \|e_n[1]\|_2^2, \\ \sigma_y &= \frac{1}{N} \sum_{n=1}^{N} \gamma_{nk} \|e_n[2]\|_2^2. \end{aligned} \tag{15}$$

Fig. 2: Demonstrations of perspective distortion removal. The first and second row present original and corrected images respectively.

Taking advantage of the estimated $\boldsymbol{\Theta}$, we can calculate the reference points $\mathbf{U}$, which is then used for estimation of rigid-motions that lead to perspective distortions. Here, we formalize the pose estimation into a least-square problem as follows:

$$\boldsymbol{\xi}^* = \arg\min_{\boldsymbol{\xi}} \frac{1}{2} \sum_{n=1}^{N} \boldsymbol{e}_n^T \boldsymbol{e}_n,$$
$$\boldsymbol{e}_n = \hat{\mathbf{x}}_n - \frac{1}{z_n}(\boldsymbol{R}(\boldsymbol{\theta})\hat{\mathbf{u}}_n + \boldsymbol{t}),$$
(16)

where $\boldsymbol{\xi} = [\boldsymbol{\theta}, \boldsymbol{t}]^T$, $z_n$ is the third dimension of $\boldsymbol{R}(\boldsymbol{\theta})\hat{\mathbf{u}}_n + \boldsymbol{t}$. The error $\boldsymbol{e}$ is defined in the normalized image plane, which is induced by rigid-body motion $\boldsymbol{\xi}$.

Eq. (16) can be efficiently solved using numerical methods, in which the Jacobian matrix is indispensable for performing gradient descent. According to [24], we directly present the Jacobian matrix of $\boldsymbol{e}$ w.r.t. $\boldsymbol{\xi}$ as follows:

$$\frac{\partial \boldsymbol{e}}{\partial \boldsymbol{\xi}} = \left[ \begin{array}{cccccc} \hat{x}\hat{y} & -\left(1+\hat{x}^2\right) & \hat{y} & \frac{-1}{z} & 0 & \frac{\hat{x}}{z} \\ 1+\hat{y} & -\hat{x}\hat{y} & -\hat{x} & 0 & \frac{-1}{z} & \frac{\hat{y}}{z} \end{array} \right]. \quad (17)$$

where $\hat{x}_n$ and $\hat{y}_n$ are the first two dimensions of $(\boldsymbol{R}(\boldsymbol{\theta})\hat{\mathbf{u}}_n + \boldsymbol{t})/z_n$. Leveraging the estimated pose $\boldsymbol{R}$ and $\boldsymbol{t}$, we are able to remove the perspective distortion by applying an inverse image warping operation.

## V. EXPERIMENTS

To verify the proposed algorithm, we collect a dataset with 25 images from 5 different elevators. All the samples are captured from different angles of views and contain severe perspective distortions. In all the experiments, $\alpha$ is set to 0.8, the initial value of $(o_x, o_y)$ is set to the coordinate of the most top-left detected buttons, $(\Delta_x, \Delta_y)$ is set to (100, 100), and $\sigma_x = 640$, $\sigma_y = 640$. For pose estimation, the initialization value of $\boldsymbol{\xi}$ is set to (-1.0, -1.0, -1.0, 0.1, 0.1, 0.1). The intrinsic parameter of the camera is

$$\boldsymbol{K} = \left[ \begin{array}{ccc} 320 & 0 & 320 \\ 0 & 320 & 240 \\ 0 & 0 & 1 \end{array} \right]. \quad (18)$$

The detailed experimental results are listed in Table. I and some demonstrations are also presented in Fig. 2.

To measure the accuracy of distortion removal, the optimal value of Eq. (16) is utilized, which indicates an averaged square distance between the reference points and button centers in the corrected normalized image plane. As we can see from Table. I, the accuracy value of which is multiplied by 1000, the proposed algorithm achieves very high accuracy on all the five groups of images, which strongly varies the effectiveness of the proposed algorithm. According to Fig. 2, we can see that the algorithm is also robust to outliers. Even for panel images that contain irregular button layout, the grid fitting algorithm can also give accurate estimation.

## VI. CONCLUSIONS

In this paper, we present a novel algorithm for autonomously removing perspective distortions on a single image. The algorithm takes as input the outcomes of button recognizer, then conducts grid fitting and pose estimation procedures, and finally outputs the corrected images, which can help improve the accuracy of character recognition. Currently, the algorithm can only handle internal panel images that contain a number of buttons. For external elevator panel images, as there are not enough button samples, the algorithm may easily fail. In the next step, we will integrate more feature points, e.g., button corners, into the framework to make it capable of handling external elevator panels. We will also collect a larger dataset to verify the robustness of the algorithm, based on which further improvement will be made.

TABLE I: Accuracy of distortion removal. In each group, the fist and second row present the image number and accuracy, respectively.

| No. | I-55 | I-660 | I-1180 | I-1610 | I-2190 | Average |
|---|---|---|---|---|---|---|
| 1 | 0.3026 | 0.0086 | 0.8447 | 2.1007 | 1.1680 | 0.8849 |
| No. | I-20 | I-450 | I-820 | I-1210 | I-1590 | Average |
| 2 | 0.0001 | 1.8100 | 0.0003 | 0.0003 | 0.2041 | 0.4030 |
| No. | I-20 | I-440 | I-910 | I-1090 | I-1500 | Average |
| 3 | 0.0003 | 0.1854 | 0.7971 | 0.2339 | 1.9590 | 0.6351 |
| No. | I-20 | I-440 | I-840 | I-1000 | I-1320 | Average |
| 4 | 0.0000 | 0.0000 | 0.0000 | 0.0332 | 0.3392 | 0.0745 |
| No. | I-20 | I-270 | I-410 | I-610 | I-760 | Average |
| 5 | 0.5846 | 1.8926 | 0.3684 | 0.8250 | 0.2323 | 0.7806 |

## REFERENCES

[1] D. Zhu, T. Li, D. Ho, C. Wang, and M. Q.-H. Meng, "Deep reinforcement learning supervised autonomous exploration in office environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7548–7555.

[2] D. Zhu, Y. Du, Y. Lin, H. Li, C. Wang, X. Xu, and M. Q.-H. Meng, "Hawkeye: Open source framework for field surveillance," in *2017 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2017, pp. 6083–6090.

[3] C. Wang, D. Zhu, T. Li, M. Q. . Meng, and C. W. de Silva, "Efficient autonomous robotic exploration with semantic road map in indoor environments," *IEEE Robotics and Automation Letters*, vol. 4, no. 3, pp. 2989–2996, July 2019.

[4] D. Zhu, T. Li, D. Ho, T. Zhou, and M. Q. Meng, "A novel ocr-rcnn for elevator button recognition," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 3626–3631.

[5] Z. Dong, D. Zhu, and M. Q.-H. Meng, "An autonomous elevator button recognition system based on convolutional neural networks," in *Robotics and Biomimetics (ROBIO), 2017 IEEE International Conference on*. IEEE, 2017, pp. 2533–2539.

[6] J. Liu and Y. Tian, "Recognizing elevator buttons and labels for blind navigation," in *2017 IEEE 7th Annual International Conference on CYBER Technology in Automation, Control, and Intelligent Systems (CYBER)*. IEEE, 2017, pp. 1236–1240.

[7] E. Klingbeil, B. Carpenter, O. Russakovsky, and A. Y. Ng, "Autonomous operation of novel elevators for robot navigation," in *IEEE International Conference on Robotics and Automation*, 2010, pp. 751–758.

[8] T. Li, J. Pan, D. Zhu, and M. Q. . Meng, "Learning to interrupt: A hierarchical deep reinforcement learning framework for efficient exploration," in *2018 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2018, pp. 648–653.

[9] H. H. Kim, D. J. Kim, and K. H. Park, "Robust elevator button recognition in the presence of partial occlusion and clutter by specular reflections," *IEEE Transactions on Industrial Electronics*, vol. 59, no. 3, pp. 1597–1611, 2011.

[10] N. F. W. Z. Wan, M. R. Daud, S. Razali, and M. F. Abas, "Elevators external button recognition and detection for vision-based system," *Proceeding of the Electrical Engineering Computer Science & Informatics*, 2014.

[11] P.-Y. Yang, T.-H. Chang, Y.-H. Chang, and B.-F. Wu, "Intelligent mobile robot controller design for hotel room service with deep learning arm-based elevator manipulator," in *2018 International Conference on System Science and Engineering (ICSSE)*. IEEE, 2018, pp. 1–6.

[12] Y.-C. Hsu and Y.-S. Chen, "Button segmentation from an elevator inside door image," in *Proceedings of the International MultiConference of Engineers and Computer Scientists*, vol. 1, 2018.

[13] F. Wang, G. Chen, and K. Hauser, "Robot button pressing in human environments," in *2018 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2018, pp. 7173–7180.

[14] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C. Y. Fu, and A. C. Berg, "Ssd: Single shot multibox detector," in *European Conference on Computer Vision*, 2016, pp. 21–37.

[15] Z. Wojna, A. Gorban, D. S. Lee, K. Murphy, Q. Yu, Y. Li, and J. Ibarz, "Attention-based extraction of structured information from street view imagery," 2017.

[16] S. Ren, K. He, R. Girshick, and J. Sun, "Faster r-cnn: towards real-time object detection with region proposal networks," in *International Conference on Neural Information Processing Systems*, 2015, pp. 91–99.

[17] D. Zhu and M. Q. . Meng, "Automatic false positive canceling for indoor human detection," in *2016 IEEE International Conference on Information and Automation (ICIA)*, Aug 2016, pp. 381–384.

[18] T. Li, D. Zhu, and M. Q. . Meng, "A hybrid 3dof pose estimation method based on camera and lidar data," in *2017 IEEE International Conference on Robotics and Biomimetics (ROBIO)*, Dec 2017, pp. 361–366.

[19] Z. Min, J. Wang, S. Song, and M. Q.-H. Meng, "Robust generalized point cloud registration with expectation maximization considering anisotropic positional uncertainties," in *2018 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2018, pp. 1290–1297.

[20] Z. Min, J. Wang, and M. Q. . Meng, "Robust generalized point cloud registration with orientational data based on expectation maximization," *IEEE Transactions on Automation Science and Engineering*, pp. 1–15, 2019.

[21] Z. Min, J. Wang, and M. Q.-H. Meng, "Joint rigid registration of multiple generalized point sets with hybrid mixture models," *IEEE Transactions on Automation Science and Engineering*, 2019.

[22] Z. Min, D. Zhu, and M. Q. Meng, "Accuracy assessment of an n-ocular motion capture system for surgical tool tip tracking using pivot calibration," in *2016 IEEE International Conference on Information and Automation (ICIA)*. IEEE, 2016, pp. 1630–1634.

[23] R. Horaud, F. Forbes, M. Yguel, G. Dewaele, and J. Zhang, "Rigid and articulated point registration with expectation conditional maximization," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 33, no. 3, pp. 587–602, 2010.

[24] T. D. Barfoot, "State estimation for robotics," 2019.