https://doi.org/10.2352/ISSN.2470-1173.2017.9.IRIACV-264

Finding a Needle in a Haystack: Recognizing Surgical Instruments through Vision and Manipulation

Tian Zhou, Juan P. Wachs; Purdue University; West Lafayette, IN, USA

Abstract

This paper presents an accurate and robust surgical instrument recognition algorithm to be used as part of a Robotic Scrub Nurse (RSN). Surgical instruments are often cluttered, occluded and displaying specular light, which cause a challenge for conventional vision algorithms. A learning-through-interaction paradigm was proposed to tackle this challenge. The approach combines computer vision with robot manipulation to achieve active recognition. The unknown instrument is firstly segmented out as blobs and its poses estimated, then the RSN system picks it up and presents it to an optical sensor in an established pose. Lastly the unknown instrument is recognized with high confidence. Experiments were conducted to evaluate the performance of the proposed segmentation and recognition algorithms, respectively. It is found out that the proposed patch-based segmentation algorithm and the instrument recognition algorithm greatly outperform their benchmark comparisons. Such results indicate the applicability and effectiveness of our RSN system in performing accurate and robust surgical instrument recognition.

Introduction

US hospitals are facing critical problems of nurse shortage. One report predicts that there will be a shortage of 260,000 registered nurses by 2025 in the USA [1]. This could lead to an increment in mortality. It was found that patient mortality risk is 6% higher in hospitals understaffed with nurses compared to units fully staffed [2]. One solution to such nurse shortage problem is to create robotic solutions which can take over the highly mechanical, mundane and repetitive tasks from the nurses, so as to allow them to focus on more complicated tasks. Out of all the tasks that the nurses are responsible for, the surgical instrument preparation and delivery task is one of the most important and dominant cases, which the proposed RSN system is aiming to take over.

To build a functioning RSN system, accurate instrument recognition is a critical component. Although humans can deal with this challenge using both vision and tactile information, robots are not capable to achieve comparable level of performances currently. The reason for this is that surgical instruments are often clustered together on the mayo stand and have a reflective and uniform appearance in both shapes and colors, as shown in Figure 1. Such characteristics bring challenges to conventional object recognition algorithms, since the object-of-interest does not obtain distinctive visual features. Moreover, holistic images are not easy to be taken due to heavy occlusion. To enable a RSN system to serve surgeons effectively in the Operating Room (OR), it is critical for the robot to recognize the instrument accurately, localize them precisely and manipulate them robustly. The aim of this paper is to propose one such system which meets the aforementioned requirements.

Figure 1. *A realistic mayo tray configuration*

Related work

Robotic systems have been brought into the OR in mainly three different forms [3], 1) handheld robotic tools; 2) teleoperated surgical systems and 3) autonomous surgical assistants. The first two cases focus on designing robots to increase surgeons sensorimotor capabilities. Such systems include the minimally invasive telesurgery system da Vinci, the steady-hand robotic system for microsurgical augmentation [4] and the touch-less telesurgery systems [5]. The third category focuses on robots which assist surgeons and nurses without directly touching the patient. One specific type of robot surgical assistant is the Robotic Scrub Nurse (RSN) system, which is the focus of this work. Many previous efforts in building the RSN have been spent on the human robot interaction part, where gestures [6, 7], speech [8], haptics [9] and EEG/EMG sensors [10] were used to enable communications between surgeons and RSN systems. However, there is a lack of treatment to the problem of instrument recognition and manipulation in realistic/uncontrolled settings. One way to tackle this challenge has been to develop specially-designed mayo platforms to ease the process of detecting non-overlapping instruments [11]. In another case, the instrument locations were fixed and stored ahead of time and their categories recognized through infrared labels [9]. Post-attached barcode was also used to detect and locate surgical instruments [12]. This paper tries to bridge the gap of recognizing surgical instruments in an uncontrolled surgical setting without the usage of any additional aids.

The proposed recognition algorithm is based on a hybrid process of instrument segmentation, reactive grasping and finally instrument recognition. These problems are at the core of this paper and the relevant literature are discussed in the following.

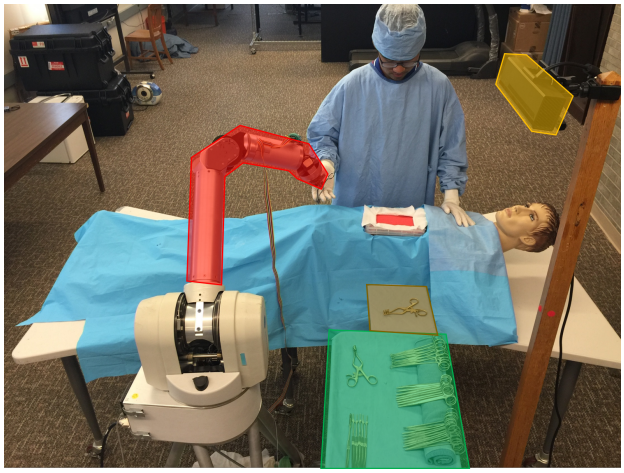

Figure 2. *Illustration of the system. The highlighted regions are mayo tray (green), recognition pad (brown), Kinect (yellow) and robotic arm (red).*

Traditional *segmentation* algorithms rely on intensity thresholding [13], edge detection [14] and region splitting [15]. Segmentation results show a split of the entire image into multiple unidentified sub-regions, nevertheless, the identity of the sub-regions was missing. Segmentation and recognition were performed at the same time using Convolutional Neural Network (CNN) [16, 17], however, significant memory and hardware requirement were necessary. In this paper, the proposed patch-based segmentation algorithm solves segmentation and recognition problem at the same time, while not requiring ambitious hardware.

For *instrument recognition*, an interactive object recognition strategy was used [18]. In this scheme, the object-of-interest was grasped, manipulated and observed actively by a robot to determine its identity [19]. Such procedure mimics the process of human learning, where the interaction with the environment plays an important role in skill development [20]. Such approach requires the robot to be capable of grasping objects without prior knowledge of the environment nor the objects, which is challenging on its own. Most commonly, robot *grasping protocols* rely on properties of the target objects, such as their full 3D models [21], depth information [22] or physical properties [23]. Reactive grasping algorithms were proposed to compensate for the loss of full object model, using tactile feedback [24] and optical proximity sensors [25]. Our proposed force-based reactive grasping protocol can enable robots to grasp objects without full models.

This paper makes the following contributions: 1) proposes an innovative segmentation algorithm based on visual codebook generation and weighted histogram backprojection; 2) develops a force-based reactive grasping protocol to enable reliable instrument grasping; 3) outlines a surgical instrument recognition algorithm. All these components, when working together, can achieve robust and automated surgical instrument recognition in a cluttered setting without any additional labels. The design and integration of these components is key for successful introduction of RSN to the OR. An illustration of the developed RSN system is given in Figure 2. Different components are color-coded for better understanding.

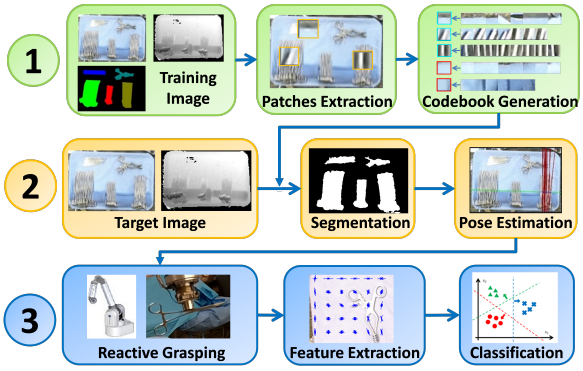

Figure 3. *System architecture*

Methodology

Recognizing and grasping surgical instruments from mayo trays in a clinical setting due to two reasons: (a) the instruments are packed together leading to occlusion, discontinuities and clutter; (b) salient points are difficult to find on metallic instruments due to their uniform and reflective composition. To tackle this challenge, we resorted to an active recognition process which involves hybrid manipulation and recognition. The system architecture is shown in Figure 3. First, a codebook of local appearances is created for segmentation purposes (step 1). In usage, the packs of instruments are segmented using the generated codebook, and their poses are estimated for grasping (step 2). Finally, a robotic arm grasps the surgical instrument and places it over a side tray where the recognition is performed (step 3). Once the instrument is recognized, clusters of instruments are assigned a specific label. The following subsections describe the instrument segmentation, reactive grasping and recognition algorithms respectively.

Instrument Pack Segmentation

To enable the robot to pick up and manipulate surgical instruments, each instrument needs to be localized on the mayo tray first. The instrument segmentation process consists of: (1) segmenting the mayo stand using adaptive threshold; (2) building color codebooks of distinctive instrument appearances; (3) segmenting instrument blobs out based on weighted histogram backprojection using the color codebook; (4) estimating poses for each cluster of instruments. Each step is discussed in the following.

Mayo Segmentation

The setup consists of a Kinect sensor placed directly on top of the mayo stand facing down to capture a complete view of the region of interest (as shown in Figure 2). A Kinect sensor delivers a color image I_C and depth image I_D, which serve as the input to the mayo and instrument segmentation process.

Otsu's threshold [13] and contour analysis are used together to get the mask of the mayo tray (denoted as M_{mayo}). Otsu's threshold was applied on the raw depth image I_D to generate a foreground mask M_{fg}. The *rectangularity* was then computed for every contours in M_{fg}. The *rectangularity* is defined as the ratio of the width to the height of the rotated bounding box of the contour. A standard mayo stand has a rectangular shape which can be identified by choosing the contour with the largest *rectangularity* among the candidates. This process is depicted in Figure 4.

Figure 4. *Mayo segmentation procedure. (left) raw input depth image I_D from Kinect; (middle) foreground mask M_{fg} generated from Otsu's threshold; (right) mayo mask M_{mayo} generated from contour examination*

Codebook Generation

From the mayo stand image, each pile of unknown instrument needs to be segmented and their poses must be determined. The first step is to segment the foreground (instruments) from the background (mayo surface). A classification-based segmentation technique was utilized to accomplish this task. Classification-based segmentation consists of building a class-specific codebook of local appearance (both color and texture) and then classifying each pixel in the image as belonging to one of those classes. Suppose there are R classes in the image, whose labels are denoted as $c_1,,c_R$. First, a class-specific codebook of local appearances, denoted as ϕ^r, was built for each object category c_r.The codebook ϕ^r consists of K characteristic local color models, represented by color histograms ($hist_k^r$), and a corresponding weight factor (w_k^r), as shown below:

$$\phi^r = \{\phi_k^r\} = \{(w_k^r, hist_k^r) | k = 1,...,K\} \quad (1)$$

where w_k^r and $hist_k^r$ represents the k_{th} weight and local color model for codebook ϕ^r. The histograms $hist_k^r$ are used later to generate a back-projected image for segmentation, and the weights w_k^r represent the relative importance of each color model. To build the codebook, we proposed a variant of the Bag of Visual Words (BoVW) algorithm [26]. The traditional BoVW method fails when directly used for instrument recognition, due to the similarity of the appearances when cluttered together. Therefore, we adopted an active recognition approach where instrument packs were segmented using BoVW as whole regions. After that, the pose for each instrument pack was estimated and a robotic arm was used to change the instrument pose for maximum visibility for further recognition.

In the codebook generation process, N random patches $P_i (i = 1,...,N)$ were extracted from manually segmented image regions of each category c_r, from each training image. Each patch P_i is associated with local color and depth information and a patch label θ_i, all together represented as:

$$P_i = \{\theta_i, \{I_C, I_D\} \cap W(x_i, y_i)\}, i = 1,...,N \quad (2)$$

where i is the index of the patch and $\theta_i \in \{c_1,...,c_R\}$ is the label of the patch. $\{I_C, I_D\}$ represents the raw color and depth image, and $W(x_i, y_i)$ is a squared window of size $w \times w$ centered at (x_i, y_i), which is the location of the patch P_i. From each patch P_i, a stack of color and depth features were extracted, forming a feature representation F_i for patch P_i. The feature set F_i includes (with d as dimensions):

- Histogram of grayscale pixel values ($d = 16$).

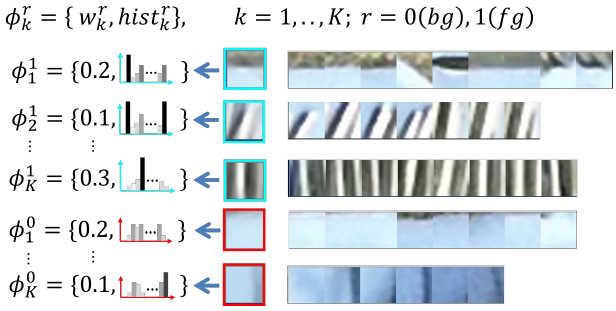

Figure 5. *Codebook clusters for foreground ϕ_k^1 (cyan) and background ϕ_k^0 (red). Each codebook entry ϕ_k^r is a pair consisting of its importance (weight w_k^r) and the local color model (histogram $hist_k^r$). The patch with colored border is the clusters average image.*

- Histogram of Hue and Saturation channel pixel values ($d = 32$).
- Histogram of oriented gradient (HOG) from the grayscale image ($d = 36$).
- Global descriptor of grayscale image: mean and standard deviation of pixel values, and mean of the Laplacian image ($d = 3$).
- Histogram of raw depth image pixels ($d = 16$).
- Global descriptor of raw depth image: mean and standard deviation of pixel values, and mean of the Laplacian image ($d = 3$).

In total, the feature representation F_i has a dimension of 106. The set of all the patches for category c_r were clustered using their feature representation F_i with Gaussian Mixture Models (GMM):

$$\{F_i | \theta_i = c_r\} = \sum_{k=1}^{K} \alpha_k^r \mathcal{N}(F | \mu_k^r, \Sigma_k^r) \quad (3)$$

where each cluster center (μ_k^r, Σ_k^r) represents the k_{th} prototypical local feature appearance for object type c_r, and the weight α_k^r represents the relative importance of that local appearance. From a given cluster k, a 2D Hue-Saturation histogram (256 bins for each dimension) was calculated from the local color images of all the patches that belong to that cluster. This histogram from cluster k, denoted as $hist_k^r$, forms one of the K local color models for codebook ϕ^r, as shown in Equation 1. The corresponding weight α_k^r of cluster k is used as the weight w_k^r for this local color model. An illustration of the generated codebook for the foreground and background categories is given in Figure 5.

Segmentation

Given a color image I_C, the probability of each pixel belonging to a category c_r is estimated to generate a per-pixel segmentation. Such probability for a given pixel (x, y), denoted as $P(c_r | I_C(x, y))$, can be estimated using the Bayes minimum error criterion [27]. It requires the estimation of a priori probability $P(c_r)$ and a data likelihood $p(I_C(x, y) | c_r)$.

The prior $P(c_r)$ for each category is estimated using the relative frequency of all the patches associated with each category: $P(c_r) = \frac{|\{P_i | \theta_i = c_r\}|}{\sum_{s=1}^{R} |\{P_i | \theta_i = c_s\}|}$. The data likelihood $p(I_C(x, y) | c_r)$ was es-

timated through the marginalization over all codebook entries of that category:

$$p(I_C(x,y)|c_r) = \sum_{k=1}^{K} p(I_C(x,y),\phi_k^r|c_r)$$
$$= \sum_{k=1}^{K} P(\phi_k^r|c_r)p(I_C(x,y)|\phi_k^r,c_r) \quad (4)$$

The generated codebook as of Equation 1 was used to estimate the parameters in the above equation. $P(\phi_k^r|c_r)$ is approximated using the weight w_k^r for the k_{th} codebook entry, indicating the relative importance of this codebook entry. $p(I_C(x,y)|\phi_k^r,c_r)$ is the likelihood of observing pixel $I_C(x,y)$ given category c_r and codebook entry $P(\phi_k^r)$. This likelihood was estimated using histogram back-projection (HBP). HBP algorithm can generate a probability of each pixel matching to a histogram from a given codebook entry. The segmentation probabilities calculated from HBP were accumulated using the corresponding weight w_k^r for the k_{th} codebook entry, according to as Equation 4. Finally, a likelihood ratio between the different categories was calculated for each pixel in the image I_C. This value was used to segment foreground from background. The likelihood of a pixel (x,y) belonging to the foreground (c_1) w.r.t. the background (c_0) is:

$$L(x,y) = \frac{P(c_1|I_C(x,y)}{P(c_0|I_C(x,y)} \propto \frac{P(c_1)p(I_C(x,y)|c_1)}{P(c_0)p(I_C(x,y)|c_0)}$$
$$= \frac{P(c_1)\sum_{k=1}^{K} w_k^1 p(I_C(x,y)|hist_k^1)}{P(c_0)\sum_{k=1}^{K} w_k^0 p(I_C(x,y)|hist_k^0)} \quad (5)$$

This value was compared with a threshold ρ to obtain a foreground mask (as $L(x,y) > \rho$). The hyper-parameter ρ was varied to generate a precision-recall curve in the later experiment.

Pose Estimation

After the segmentation, the pose for each pack of unknown instrument was estimated for initial robot grasping, as shown in Figure 6. The dominant direction of each pile was found using the Hough transform [28] (Figure 6.d). The lines extracted correspond to the edges of the instruments from a top-view. A majority vote of line directions was casted to find the dominant one. After the dominant direction has been identified, the line perpendicular to the dominant direction through the centroid of the pack was found (see the blue line in Figure 6.d). The intersection of the line with the group contour was used as the initial grasping point (yellow dot in Figure 6.d). The grasping strategy relies on picking the instrument from the center of mass and in a perpendicular direction.

After the position of the initial grasping point has been identified, the rotation of the gripper was estimated. The depth image was used to determine whether the instrument is lying flat on the mayo tray, or vertically packed within a group of instruments. The depth along the perpendicular line (blue) was scanned and its depth fluctuation was used to determine which of those two poses the instrument is at (Figure 6.e). The standard deviation of the normalized height along the scanning line was calculated as: $\sigma_h = \sqrt{\frac{1}{N}\sum_{i=1} N(h_i - \mu)^2}$, where h_i is the depth value at point

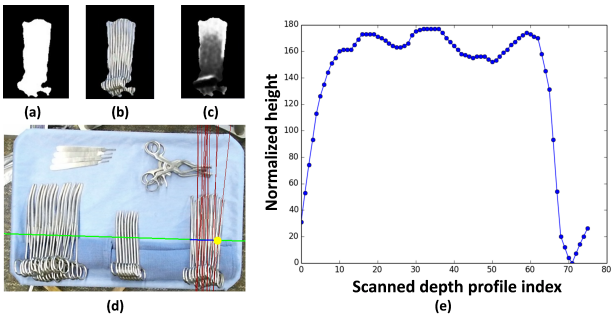

Figure 6. *Pose estimation process. (a) mask of the unknown instrument pile; (b) extracted color image; (c) extracted depth image; (d) red lines are the result of Houghline transform and blue line indicates the perpendicular depth profile scanning line, the orange point indicates the initial picking point. (e) the scanned depth profile. The fluctuation indicates a more likely vertical pose instead of flat pose*

i (in total N points) and $\mu = \frac{1}{N}\sum_{i=1}^{N} h_i$ is the average. σ_h indicates the depth fluctuation of the instrument and is compared with an empirical threshold σ_{ref} to determine whether the instruments state is flat or standing. The threshold σ_{ref} was chosen optimally based on a separate validation dataset. In the case where the instrument is lying flat, the robot would pick it from above; while a tilted instrument would require approach in a perpendicular direction to the instrument surface.

Reactive Grasping Protocol

Our approach to instrument recognition is done through manipulation of the instrument. Classical grasping approaches require a 3D model of the object that is to be grasped. Such approaches cannot be applied directly to our scenario since direct recognition of the instruments is not attainable. We first pick generic objects and only after manipulation will their category be determined. Assume that the instruments are packed in piles. The instrument orientation is determined through the value of the force feedback at the end-effector of the robot when in contact with the instrument. Once the orientation is determined, the instruments are picked by extending the end-effector in a perpendicular direction. A special end-effector was developed with an integrated electromagnet to attract metal instruments. Four force sensors were mounted between the electro-magnet gripper and the wrist of the robot, forming a cross shape in the xy-plane to acquire force information at the robots end-effector, as its diagram shown in Figure 7.

To improve the accuracy of force readings and compensate potential placement imbalances of the four force sensors, a calibration process was carried out before deployment. For raw force reading $\tilde{F}_x$ (where $x \in \{up, down, left, right\}$), the minimum force $\tilde{F}_x^{min}$ was recorded when the robotic arm is at a ready-to-pick position above the mayo stand, and the maximum force $\tilde{F}_x^{max}$ was recorded when the robotic arm is physically touching the instrument side ways to ensure that only force sensor x makes a point contact with the mayo stand. After recording the minimum and maximum readings for each sensor, the force value was

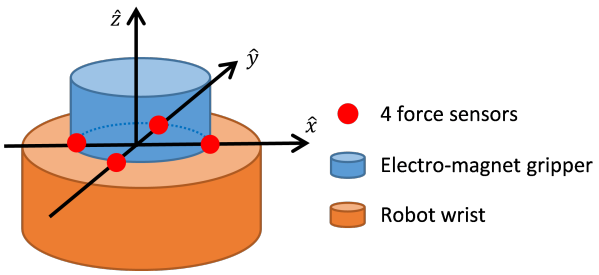

Figure 7. *Diagram of the force sensors and electromagnet. The four force sensors are placed on the x-y plane between the electromagnet and the robot wrist, forming a cross shape.*

normalized following equation:

$$F_x = \begin{cases} 0, & \text{if } \tilde{F}_x < \tilde{F}_x^{min} \\ 1, & \text{if } \tilde{F}_x > \tilde{F}_x^{max} \\ \frac{\tilde{F}_x - \tilde{F}_x^{min}}{\tilde{F}_x^{max} - \tilde{F}_x^{min}}, & \text{otherwise} \end{cases} \quad (6)$$

The resultant force F_x was in the range $[0,1]$ and sampled at 5 Hz rate. The four normalized force readings are stacked as a vector $\vec{F} = [F_{up}, F_{down}, F_{left}, F_{right}]$. A sharp increase in the sensing force indicates contact between the end-effector and the instrument. Given the force readings $\vec{F}$ and the sensors known locations under the electromagnet, the norm direction $\vec{n} = [n_1, n_2, n_3]^T$ of the force plane is estimated. When proper contact is being made, all the force sensors acquire approximately the same force value (the force is evenly distributed in the surface when the magnet is perfectly perpendicular to the plane of the instrument), thus resulting in a force norm vector $\vec{n}$ to be aligned with the z-axis $\hat{z}$. If the force norm is not aligned, then the end-effectors orientation is corrected towards the direction to reduce discrepancy between the norm direction $\vec{n}$ and z-axis $\hat{z}$. Figure 8 shows a sample force reading during one instrument picking procedure.

The norm direction $\vec{n}$ was estimated using least square

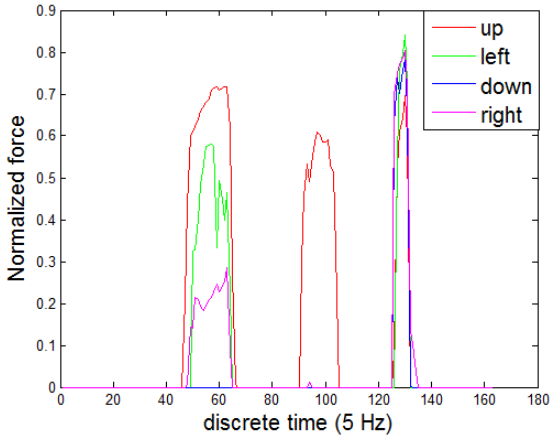

Figure 8. *Sample force readings for one instrument picking. The first peak is the first contact with instrument, while second peak indicates a second contact after adjusting orientation, and the last peak indicates a final stable contact (evenly distributed force).*

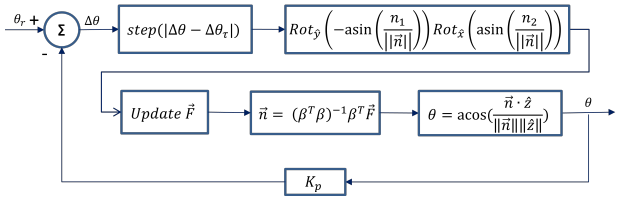

Figure 9. *System diagram for the reactive grasping protocol*

regression with force input $\vec{F}$ following the equation: $\vec{n} = (\beta^T\beta)^{-1}\beta^T\vec{F}$, where β is the coefficient incorporating locations of force sensors in the wrist coordinate, denoted as $\beta = [1,0,F_{up};-1,0,F_{down};0,1,F_{left};0,-1,F_{right}] \in \mathbb{R}^{4\times3}$. The angle difference θ between $\vec{n}$ and z-axis $\hat{z}$ was calculated and compared against a reference angle θ_r to form the control input $\Delta\theta$. Here the reference angle θ_r was set to 0 to ensure an evenly distributed force during contact. Different θ_r can be chosen in favor of other grasping poses. A tolerance $\Delta\theta_\tau$ was used to limit the orientation correction. The reactive grasping control algorithm is illustrated in Figure 9.

Instrument Recognition

Once the end-effector attracted an instrument, it was lifted and subsequently placed on a side surface (known as the recognition pad). The recognition pad has a uniform background and allows maximum exposure to enable instrument recognition. Then, an object recognition procedure was used to recognize the type of instrument. The instrument recognition framework is shown in Figure 10. It includes foreground extraction, feature encoding and instrument classification. Detailed steps of each module are presented next.

The pipeline of the image processing module involves first to record an image frame I_{t-1} when no instruments are present in the recognition pad. A color-based background model BG_{t-1} is built based on the pixel values of I_{t-1} using Gaussian Mixture Models [29]. Then, given the current frame I_t in which an instrument is present, the foreground mask M is extracted using I_t and BG_{t-1}. The object contour cnt is then extracted from foreground mask M. Certain region related properties are then extracted from the contour cnt to represent the foreground object.

Next, a straight line is fit to the contour pixels as the major axis for the object, then the foreground object is rotated based on

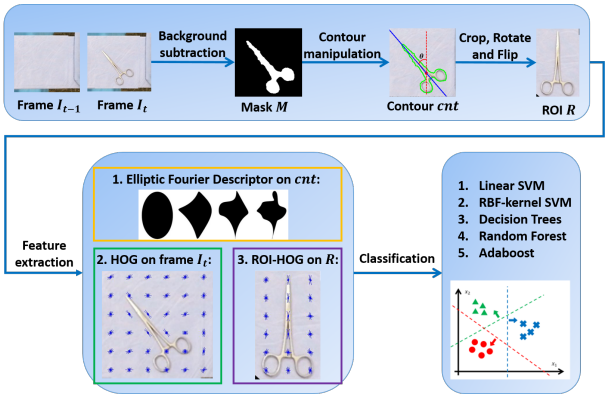

Figure 10. *Instrument recognition algorithm framework*

the angle difference between the major axis and the vertical line, so that the foreground object can be vertically oriented. Last, the instrument is rotated again (if necessary) so that the lower half part of the objects area is larger than the upper half. The pixel area was used as an approximation for mass. This is a reasonable assumption for our scenario, since most surgical instruments have a scissors-like shape where the larger area corresponding to the handle has a larger mass. After localization, rotation and clipping, the Region-Of-Interest (ROI) image R which contains the foreground object was generated. Certain features were then extracted from R for instrument classification.

We applied 3 types of feature extraction methods, as described below:

- Histogram of Oriented Histograms (*HOG*) [30] extracted from the raw image I_t
- Elliptic Fourier Descriptor (*EFD*) [31] extracted from contour *cnt*
- Region-Of-Interest HOG (*ROI-HOG*) is the proposed method which generates HOG features based on the ROI image R. This method combines the advantage of HOG features with the context of the task (the region of interest), aiming to achieve better performances.

The generated features were then classified by several discriminative classifiers, including Support Vector Machines (with linear and radial basis function kernel) [32], Decision Trees [33], Random Forest [34] and Adaboost [35]. The one-vs-all classification strategy was used to enable a multiclass recognition.

Experiments

Experiments were conducted to test the performance of the instrument segmentation, reactive grasping and instrument recognition algorithms, respectively. The image dataset used in our paper is publicly available at `https://github.com/tian-zhou/Surgical-Instrument-Dataset`

Instrument Pack Segmentation Experiment

This experiment aims to evaluate the proposed segmentation algorithm to discriminate the group of surgical instruments (foreground) from the mayo stand image (background). A dataset of surgical instrument images on the mayo stand was collected using Kinect, with various instrument layouts and lighting conditions. The instrument ground truth mask was manually annotated using the LabelMe toolbox [36]. Initially 60 image sets (color image, depth image and ground truth mask) were recorded and then each image was distorted by rotating in 8 angles (0 to 360 stepped by 45) to generate a larger data set (resulting in total 480 images).

The experiment follows a 10-fold cross-validation setup, where in each fold, 90% of the entire dataset was used for training and the remaining 10% for testing. From each image in the training data, 100 random patches ($N = 100$) of size $16 \times 16 (w = 16)$ were extracted for both foreground and background. The extracted patches from all the training images were then clustered by fitting a GMM model on their associated feature representations with 10 components ($K = 10$). Codebooks were then generated by saving the color histograms and corresponding weights for each cluster. For both foreground and background categories, a codebook was generated during the training process. For segmentation purposes, the weighted histogram backprojection algorithm was used to calculate the likelihood of a given pixel belonging to the foreground. The likelihood ratio was then compared to the threshold ρ to determine the final segmentation result. The segmentation result was compared against the ground truth mask. The F1 score was used for evaluation, which is a common metric to evaluate segmentation result [37]. The likelihood threshold was swept to obtain the precision-recall (PR) curve demonstrating the segmentation performance. The Area Under Curve (AUC) is used as a cumulative metric for evaluation. We compared the proposed segmentation algorithm against *grabCut* algorithm [38] and Morphological Geodesic Active Contours (*snake*) algorithm [39]. Both benchmark algorithms need an initial segmentation marker to start the full segmentation process. We generated the marker image using ground truth labels, where the centroid pixel of each instrument pile was labeled as positive. The segmentation was then carried out on the RGB image of the instrument. Figure 11 presents the PR curves for the three different algorithms. The maximum F1 score achieved by each algorithm was also marked.

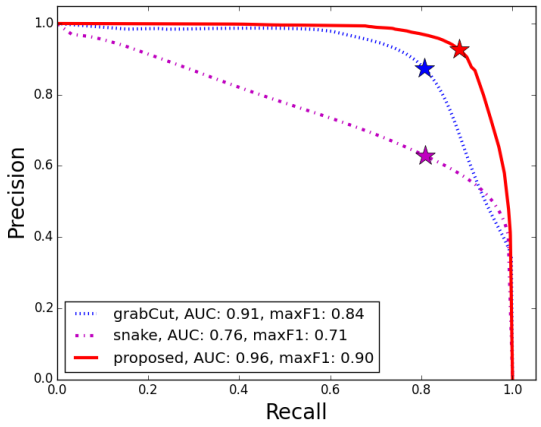

Figure 11. *Precision-recall curve for the proposed segmentation algorithm, compared against grabcut and active contour model (snake) algorithms. The AUC and max F1 score were shown in the legend. The max F1 scores were marked with star.*

It is found that the proposed segmentation algorithm achieves the highest AUC and maxF1 score, followed by *grabCut* and last *snake* algorithm. It is also worth noting that the proposed algorithm does not require an initial marker to start the segmentation process, while both of the benchmark algorithms do.

Reactive Grasping Experiment

To assess the effectiveness of the grasping method, two methods were compared: one using the reactive grasping protocol (RG), and the other using the open-loop grasping (OLG) protocol. Both protocols use the estimated pose from the optical sensor to build an initial picking strategy, while the RG protocol additionally used force sensors to correct the orientation of the tool-tip. This experiment compared the instrument picking success rate of these two protocols. Figure 13 shows the setup for the mayo stand in this experiment. It includes five sets of surgical instruments, named retractor (2 pieces), scalpel (5), Babcock forceps (6), scissors (8) and hemostat (10). They were all grouped together in a way similar to the surgical setting in the operating room. For both

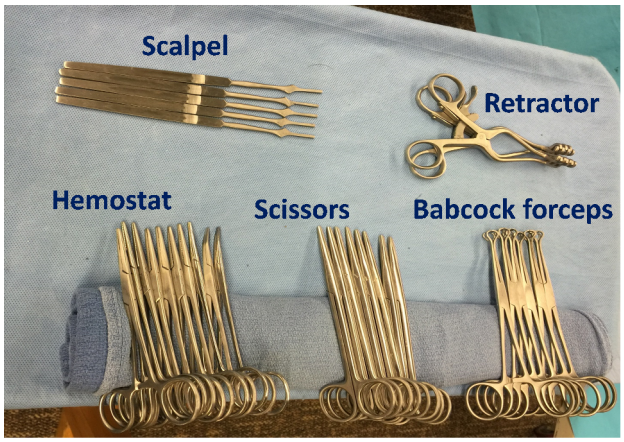

Figure 12. *Mayo setup for the grasping experiment. There are five sets of surgical instruments, with their name shown.*

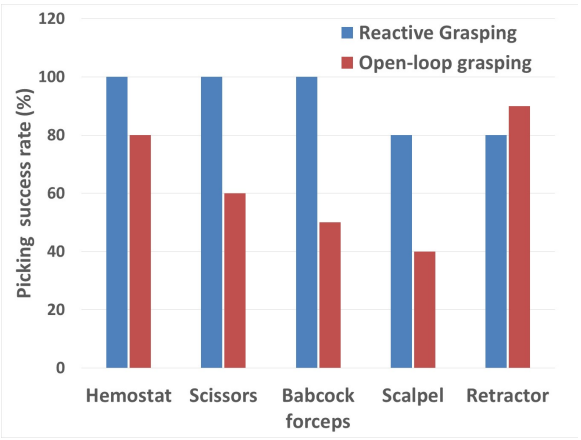

Figure 13. *Picking success rate for Reactive Grasping vs Open-loop Grasping.*

grasping methods, the task is to pick up each surgical instrument in sequence successfully. The grasping process was repeated 10 times for each type of instrument, resulting in a total of 50 picks. The average picking success rate was 64% for OLG protocol and 92% for RG protocol. The per-instrument average picking success rate was shown in Figure 14. A paired sample t-test between the per-instrument average success rates of the two protocols yield a one tail p-value of 0.029, which indicates the superiority of the RG policy compared to the OLG policy.

Since the instruments are packed tightly, the electromagnetic force was propagated to near-by instruments, resulting in potentially picking up multiple instrument at the same time. This was not a desired outcome and resulted in false alarms. Under the RG protocol, 6 out of the 46 successful picks include picking multiple instruments, resulting in a false alarm rate of 13%. The same number calculated for OLG protocol was 6.3%. The RG protocol has a higher false alarm rate, since the orientation correction algorithm leads to a better contact between the magnet and the instruments. Such better contact resulted in a stronger electromagnetic force propagated to the instrument pile. 62% of all the false alarms in both grasping methods occurred when picking the scalpel, since it is one of the lightest instruments

Instrument Recognition Experiment

A dataset of surgical instruments on the recognition pad was collected using our system setup. There are in total 5 types of instruments, named scalpel, retractor, scissors, hemostat and forceps. For each type of instrument, 20 images were taken initially with different illumination and layout conditions.

The dataset was increased by using label-preserving transformation [40] to produce variants of the original images. This helps to increase dataset size and also prevent over-fitting. Two distinct forms of data augmentation techniques were applied. The first consists of applying rotation and mirroring (horizontally) on the original image. Each input image was rotated in 8 angles (0 to 360 degrees stepped by 45), and then mirrored horizontally. The second consists of altering the RGB channel intensities. This technique follows an important property of natural images since an object recognition algorithm should be invariant to intensity and illumination changes. We followed the steps of [40]. A Principal Component Analysis (PCA) was first applied on the set of RGB pixel values of the entire training dataset. Then for each training image, multiples of the principal components were added. The magnitudes were proportional to the corresponding eigenvalues times a random variable drawn from a Gaussian with zero mean and standard deviation of 0.1. Therefore to each RGB image pixel $I_{xy} = [I^R_{xy}, I^G_{xy}, I^B_{xy}]^T$, we added the following brightness values:

$$[\mathbf{p}_1, \mathbf{p}_2, \mathbf{p}_3][\alpha_1\lambda_1, \alpha_2\lambda_2, \alpha_3\lambda_3]^T \tag{7}$$

Where $\mathbf{p}_i$ and λ_i are the i_{th} eigenvector and eigenvalue of the 3×3 covariance matrix of RGB pixel values, respectively, and α_i is the random variable mentioned before. Both the original version and the altered version were kept in the dataset. After applying both data augmentation techniques, 32 augmented images were generated out of 1 training seed, and the final dataset includes 3200 images with 640 for each instrument type. The learning setting follows a special 5-fold cross validation, where the original training instance and its augmented variants always reside in the same data split. This is to avoid the case where the original example was used for training and its variants were used for testing.

The hyper-parameters for each feature extraction method were chosen based on a grid search. The parameters yielding the highest recognition accuracy on a separated split were chosen as the optimal parameters. In the *HOG* feature method, image I_t was resized to dimension 72×56, and then 8×8 cell size, 2×2 cells per block and 9 orientation bins were used for *HOG* calculation. In the *EFD* case, the first 3 orders of Fourier coefficients were used as features. For the *ROI-HOG* case, the ROI image R was resized to 24×48, and then 8×8 cell size, 2×2 cells per block and 9 orientation bins were used. For all feature extraction methods, the generated features were normalized to have mean of zero and standard deviation of one for each channel.

The features were classified using five classifiers, including linear SVM, rbf-kernel SVM, Decision Tree, Random Forest and Adaboost. The implementation was based on scikit-learn library [41]. The default hyper-parameters were used to ensure a fair comparison. The resultant recognition accuracy on the testing split is shown in Figure 14.

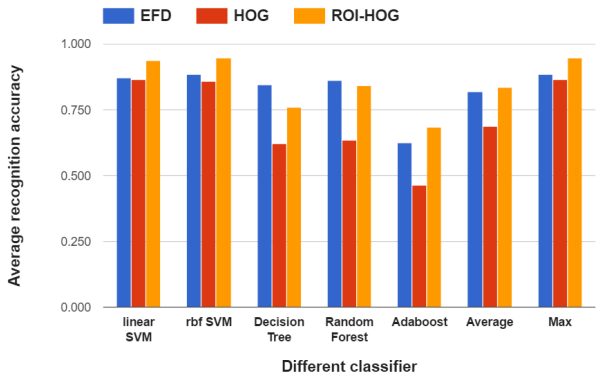

Figure 14. *The recognition accuracy for different features using different classifiers. The average and maximum performance of 5 classifiers for each features set was summarized in the two rightmost columns. The best performance was achieved by ROI-HOG with rbf-kernel SVM, with a recognition accuracy of 94.8%.*

In all except Decision Tree cases, the *ROI-HOG* features outperformed the *EFD* method and *HOG* features, which indicates the superiority of the proposed feature extraction method. The best recognition accuracy was achieved by the proposed *ROI-HOG* algorithm with rbf-kernel SVM, with an average accuracy of 94.8%. Recognition results showed that the strategy consisting of grasping and moving the instruments allows for reliable instrument recognition.

Conclusion

In this paper we presented a solution to the challenge of accurate surgical instrument recognition by a Robotic Scrub Nurse in the Operating Room. Recognizing surgical instruments from mayo trays in a clinical setting is challenging due to the instrument clustered in groups, leading to self-occlusion, discontinuities and clutter. A hybrid computer vision and robotic manipulation strategy was adopted to tackle this challenge. Initially, the instruments were segmented and their poses estimated for robotic manipulation. Later, the RSN system picked up the unknown instrument and presented it in an optimal view for robust recognition. Experiments were conducted to evaluate the performance of each critical component of the system. The proposed segmentation algorithm achieved an F-score of 0.90, which outperforms the baseline grabCut algorithm (0.84) and snake algorithm (0.71). The proposed force-based grasping protocol achieved an average picking success rate of 92% with various instrument layouts, compared to a success rate of 64% for open-loop grasping protocol. The proposed based instrument recognition module can reach a recognition accuracy of 94.8%, which is the highest among several benchmark methods. The experimental results proved the feasibility and effectiveness of the proposed RSN system.

Future work includes the design of an adaptive force controller to ensure one instrument being picked up every time. We also plan to deploy the developed system in the OR. To that end, it is necessary to design a safe instrument delivery protocol for a reliable human-machine collaboration in the surgical setting.

Acknowledgments

Research supported by the NPRP award (NPRP 6-449-2-181) from the Qatar National Research Fund (a member of The Qatar Foundation). The statements made herein are solely the responsibility of the authors.

References

[1] P. I. Buerhaus, D. I. Auerbach, and D. O. Staiger, "The recent surge in nurse employment: Causes and implications," *Health Affairs*, vol. 28, no. 4, pp. w657–w668, 2009.

[2] J. Needleman, P. Buerhaus, V. S. Pankratz, C. L. Leibson, S. R. Stevens, and M. Harris, "Nurse staffing and inpatient hospital mortality," *New England Journal of Medicine*, vol. 364, no. 11, pp. 1037–1045, 2011.

[3] P. Dario, B. Hannaford, and A. Menciassi, "Smart surgical tools and augmenting devices," *IEEE Transactions on Robotics and Automation*, vol. 19, pp. 782–792, Oct. 2003.

[4] R. Taylor, P. Jensen, L. Whitcomb, A. Barnes, R. Kumar, D. Stoianovici, P. Gupta, Z. Wang, E. Dejuan, and L. Kavoussi, "A steady-hand robotic system for microsurgical augmentation," *The International Journal of Robotics Research*, vol. 18, no. 12, pp. 1201–1210, 1999.

[5] T. Zhou, M. E. Cabrera, T. Low, C. Sundaram, and J. Wachs, "A comparative study for telerobotic surgery using free hand gestures," *Journal of Human-Robot Interaction*, vol. 5, no. 2, pp. 1–28, 2016.

[6] M. G. Jacob, Y.-T. Li, and J. P. Wachs, "Gestonurse: a multimodal robotic scrub nurse," in *Proceedings of the seventh annual ACM/IEEE international conference on Human-Robot Interaction*, pp. 153–154, ACM, 2012.

[7] A. Agovic, J. Levine, A. Agovic, and N. Papanikolopoulos, "Computer vision issues in the design of a scrub nurse robot," in *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pp. 2921–2926, IEEE, 2011.

[8] C. Prez-Vidal, E. Carpintero, N. Garcia-Aracil, J. M. Sabater-Navarro, J. M. Azorn, A. Candela, and E. Fernandez, "Steps in the development of a robotic scrub nurse," *Robotics and Autonomous Systems*, vol. 60, no. 6, pp. 901–911, 2012.

[9] A. Agovic, S. Levine, N. Papanikolopoulos, and A. Tewfik, "Haptic interface design considerations for scrub nurse robots in microsurgery," in *Control & Automation (MED), 2010 18th Mediterranean Conference on*, pp. 1573–1578, IEEE, 2010.

[10] T. Zhou and J. Wachs, "Early turn-taking prediction in the operating room," 2016.

[11] Anna Kochan, "Scalpel please, robot: Penelope's debut in the operating room," *Industrial Robot: An International Journal*, vol. 32, pp. 449–451, Dec. 2005.

[12] Y. Xu, Y. Mao, X. Tong, H. Tan, W. B. Griffin, B. Kannan, and L. A. DeRose, "Robotic handling of surgical instruments in a cluttered tray," *IEEE Transactions on Automation Science and Engineering*, vol. 12, no. 2, pp. 775–780, 2015.

[13] N. Otsu, "A threshold selection method from gray-level histograms," *Automatica*, vol. 11, no. 285-296, pp. 23–27, 1975.

[14] M. Kass, A. Witkin, and D. Terzopoulos, "Snakes: Active contour models," *International journal of computer vision*, vol. 1, no. 4, pp. 321–331, 1988.

[15] R. Achanta, A. Shaji, K. Smith, A. Lucchi, P. Fua, and S. Susstrunk, "SLIC superpixels compared to state-of-the-art superpixel methods," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 11, pp. 2274–2282, 2012.

[16] J. Long, E. Shelhamer, and T. Darrell, "Fully convolutional networks for semantic segmentation," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pp. 3431–3440, 2015.

[17] R. Girshick, J. Donahue, T. Darrell, and J. Malik, "Rich feature hierarchies for accurate object detection and semantic segmentation," in *Proceedings of the IEEE conference on computer vision and pattern recognition*, pp. 580–587, 2014.

[18] K. Hausman, C. Corcos, F. Sha, and G. S. Sukhatme, "Towards interactive object recognition," 2014.

[19] N. Lyubova, S. Ivaldi, and D. Filliat, "From passive to interactive object learning and recognition through self-identification on a humanoid robot," *Autonomous Robots*, vol. 40, pp. 33–57, June 2015.

[20] C. Yu, H. Zhang, and L. B. Smith, "Learning through multimodal interaction," in *Proceedings of the Fifth International Conference on Development and Learning (ICDL06)*, 2006.

[21] C. Rosales, J. M. Porta, and L. Ros, "Global optimization of robotic grasps," *Proceedings of Robotics: Science and Systems VII*, 2011.

[22] I. Lenz, H. Lee, and A. Saxena, "Deep learning for detecting robotic grasps," *The International Journal of Robotics Research*, vol. 34, no. 4-5, pp. 705–724, 2015.

[23] A. Rodriguez, M. T. Mason, and S. Ferry, "From caging to grasping," *The International Journal of Robotics Research*, p. 0278364912442972, 2012.

[24] K. Hsiao, S. Chitta, M. Ciocarlie, and E. G. Jones, "Contact-reactive grasping of objects with partial shape information," in *2010 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*, pp. 1228–1235, Oct. 2010.

[25] K. Hsiao, P. Nangeroni, M. Huber, A. Saxena, and A. Y. Ng, "Reactive grasping using optical proximity sensors," in *IEEE International Conference on Robotics and Automation, 2009. ICRA '09*, pp. 2098–2105, May 2009.

[26] J. Yang, Y.-G. Jiang, A. G. Hauptmann, and C.-W. Ngo, "Evaluating bag-of-visual-words representations in scene classification," in *Proceedings of the international workshop on Workshop on multimedia information retrieval*, pp. 197–206, ACM, 2007.

[27] P. Domingos and M. Pazzani, "On the optimality of the simple bayesian classifier under zero-one loss," *Machine learning*, vol. 29, no. 2-3, pp. 103–130, 1997.

[28] D. H. Ballard, "Generalizing the hough transform to detect arbitrary shapes," *Pattern recognition*, vol. 13, no. 2, pp. 111–122, 1981.

[29] Z. Zivkovic and F. van der Heijden, "Efficient adaptive density estimation per image pixel for the task of background subtraction," *Pattern Recognition Letters*, vol. 27, pp. 773–780, May 2006.

[30] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, vol. 1, pp. 886–893, IEEE, 2005.

[31] F. P. Kuhl and C. R. Giardina, "Elliptic fourier features of a closed contour," *Computer graphics and image processing*, vol. 18, no. 3, pp. 236–258, 1982.

[32] M. A. Hearst, S. T. Dumais, E. Osman, J. Platt, and B. Scholkopf, "Support vector machines," *IEEE Intelligent Systems and their Applications*, vol. 13, no. 4, pp. 18–28, 1998.

[33] S. R. Safavian and D. Landgrebe, "A survey of decision tree classifier methodology," 1990.

[34] A. Liaw and M. Wiener, "Classification and regression by randomForest," *R news*, vol. 2, no. 3, pp. 18–22, 2002.

[35] G. Rtsch, T. Onoda, and K.-R. Mller, "Soft margins for AdaBoost," *Machine learning*, vol. 42, no. 3, pp. 287–320, 2001.

[36] B. C. Russell, A. Torralba, K. P. Murphy, and W. T. Freeman, "LabelMe: a database and web-based tool for image annotation," *International journal of computer vision*, vol. 77, no. 1-3, pp. 157–173, 2008.

[37] S. Alpert, M. Galun, A. Brandt, and R. Basri, "Image segmentation by probabilistic bottom-up aggregation and cue integration," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 34, no. 2, pp. 315–327, 2012.

[38] C. Rother, V. Kolmogorov, and A. Blake, "Grabcut: Interactive foreground extraction using iterated graph cuts," in *ACM transactions on graphics (TOG)*, vol. 23, pp. 309–314, ACM, 2004.

[39] P. Marquez-Neila, L. Baumela, and L. Alvarez, "A morphological approach to curvature-based evolution of curves and surfaces," *Pattern Analysis and Machine Intelligence, IEEE Transactions on*, vol. 36, no. 1, pp. 2–17, 2014.

[40] A. Krizhevsky, I. Sutskever, and G. E. Hinton, "Imagenet classification with deep convolutional neural networks," in *Advances in neural information processing systems*, pp. 1097–1105, 2012.

[41] F. Pedregosa, G. Varoquaux, A. Gramfort, V. Michel, B. Thirion, O. Grisel, M. Blondel, P. Prettenhofer, R. Weiss, V. Dubourg, *et al.*, "Scikit-learn: Machine learning in python," *Journal of Machine Learning Research*, vol. 12, no. Oct, pp. 2825–2830, 2011.

Author Biography

Tian Zhou is currently a Ph.D. student at the School of Industrial Engineering at Purdue University, working in Intelligent Systems and Assistive Technologies (ISAT) Lab. He received his M.S. in Electrical and Computer Engineering at Purdue University in 2016, and B.S. in Information Science and Engineering at Southeast University (Nanjing, China) in 2013. His research focuses on gesture recognition, machine vision, robot teleoperation and human-robot collaboration. He is the recipient of Robotics Fellowship of 2016 AAAI, and best presentation award of 2015 AAAI poster session.

Dr. Juan Wachs is an Associate Professor in the Industrial Engineering School at Purdue University. He is the director of the Intelligent Systems and Assistive Technologies (ISAT) Lab at Purdue, and he is affiliated with the Regenstrief Center for Healthcare Engineering. He completed postdoctoral training at the Naval Postgraduate Schools MOVES Institute under a National Research Council Fellowship from the National Academies of Sciences. Dr. Wachs received his B.Ed.Tech in Electrical Education in ORT Academic College, at the Hebrew University of Jerusalem campus. His M.Sc and Ph.D in Industrial Engineering and Management from the Ben-Gurion University of the Negev, Israel. He is the recipient of the 2013 Air Force Young Investigator Award, and the 2015 Helmsley Senior Scientist Fellow. Currently he is working at UBA as part of his 2016 Fulbright U.S. Scholar Program to Argentina. He is also the associate editor of IEEE Transactions in Human-Machine Systems, Frontiers in Robotics and AI.