Chinese Journal of Electronics
Vol.XX, No.X, Jan. 20XX

Robot Intelligence for Real World Applications*

Yunhui LIU[1], Fan Zheng[1], Ruibin Guo[2], Jiangliu Wang[1], Qiang Nie[1], Xin Wang[1] and Zerui Wang[1]

(1. *T Stone Robotics Institute, The Chinese University of Hong Kong, Hong Kong, China*)

(2. *National University of Defense Technology, Changsha 410000, China*)

Abstract — This paper presents a brief review on recent works on machine intelligence for real-world applications of robots. To act in a real world environment, a robot should possess a broad sense of intelligence including speech, perception, reasoning, action, etc. In this paper, we particularly deal with the intelligence involving action or body motion. The intelligence related to robot action/motion can be classified into two categories: manipulation intelligence and mobility intelligence. The manipulation intelligence means the skill/intelligence of reliably manipulating objects according to tasks and the mobility intelligence corresponds to the ability of autonomously moving, or flying, and or jumping in a natural environment. Human-robot interaction is another important topic for real-world applications. In addition to reviewing the major approaches, this paper also gives an overview on our efforts in these important topics.

Key words — Machine intelligence, Simultaneous localization and mapping (SLAM), Medical robots, Human-robot interaction.

I. Introduction

The objectives of Artificial Intelligence (AI) are to realize human intelligence by computers, robots and other systems. The area is broad and involves in both scientific understanding and technological realization of human intelligence. The term Artificial Intelligence was invented by John McCarthy at the second Dartmouth Conference in 1956[1]. The development of AI has experienced several booms and bottoms since that. In recent years, AI is booming again due to the success of the AlphaGo, which beat the world champions in 2017. The success stimulated exciting imagination on futures and potential applications of AI and led to tremendous research and development efforts.

It should be pointed out that the technologies used by AlphaGo are mainly reasoning, decision-making, learning, etc., which purely depend on computing algorithms. The success is devoted to the significant advancement of computing power, big data, database, etc. To physically interact with a real world, physical systems such as robots need to play more active roles. In addition to intelligence purely relying on computing, a robot should have the intelligence of physically interacting with or changing settings in the real world. Therefore, robots play one of the most important roles in the development of artificial intelligence.

The intelligence that supports physical interactions of robots with the world is associated with the bodily kinesthetic intelligence defined by Howard Gardner in his book Multiple intelligence: theory in practice[2]. The kind of intelligence involves body or limb movement of a person. We think that the bodily kinesthetic intelligence for a robot can be further classified into mobility intelligence and manipulation intelligence. The mobility intelligence means the ability of a robot to reliably move in a natural environment, and manipulation intelligence is the skills of robots for manipulating objects. When interacting with humans in the physical world, a robot should also be equipped with intelligence of tracking and recognizing human actions and expressing itself, in addition to speech. This paper presents a brief review on the research development in those exciting areas and introduces some of our efforts as well.

II. Mobility Intelligence

Mobility intelligence enables robots to navigate in a natural environment. The critical problems in developing mobility intelligence include design of mobile mechanisms, motion control and navigation. Different mechanisms such as wheels, legs, biomimetic mechanisms, etc.

*Manuscript Received June XX, 20XX; Accepted July XX, 20XX. This work is supported by the National Natural Science Foundation of China (No.U1613218), Hong Kong Research Grant Council(No.14204814) and Hong Kong Innovation and Technology Commission (No.ITS/112/15FP).

 DOI:10.1049/cje.20XX.0X.0XX

have been developed. Motion control is also crucial for mobile robots, in particular for legged robots. This paper focuses the discussion on navigation of mobile robots.

Simultaneous localization and mapping (SLAM) is one of the core technologies for navigation of mobile robots in unknown environment[3−4]. SLAM is the problem to jointly estimate the state of a robot equipped with onboard sensors and the structure of the environment (the map) that the sensors are perceiving using the sensor information. SLAM is one of the fundamental problems in mobile robotics.

A typical SLAM system consists of a front-end that builds a map from different sensor measurements, a map-refinement back-end that reduces local errors, and optionally a loop closing module to overcome globally accumulated errors. For map refinement, there are two common approaches: filtering based methods[3] and graph optimization based methods[4]. While the former uses control inputs or ego-motion sensor measurements for state propagation and exteroceptive sensory observations for state update in an iterative framework, the graph based methods jointly minimize the errors originating from all measurements or constraints. For the front-end LIDAR and vision are commonly used sensing modalities that perceive the environment.

LIDAR is the mainstream sensor used in the early works of SLAM[5−8], most of which use filtering based approaches. To improve the accuracy, state-of-the-art SLAM systems using LIDAR usually adopt the graph based approaches, such as LAGO by Carlone et al.[9] and Cartographer[10] by Google. Carlone et al. demonstrated that their pose graph optimization has a peculiar structure in planar scenarios, and exploited this observation to design the estimation framework: LAGO, which could reduce the risk of being trapped in local minima, thus achieving better accuracy as well as improved efficiency. Cartographer combines scan-to-submap matching with loop closure detection and graph optimization to achieve 2D SLAM. In the background, all scans are matched to nearby submaps to create loop closure constraints, which forms a constraint graph that is periodically optimized. This system eventually provides real-time mapping and loop closure at a 5-cm resolution.

Visual SLAM (V-SLAM) estimates the robot states and environmental structures from images. Depending on how image data is utilized in constructing the estimation problems, a variety of V-SLAM methods can be classified into feature-based and direct methods. The feature-based methods are traditionally the mainstream approaches in V-SLAM. The idea is to detect feature points and match them between subsequent frames, and then to improve the states estimation to minimize the reprojection errors. Davison et al. proposed the first V-SLAM system, MonoSLAM[11], using a feature-based filtering framework. V-SLAM using particle filter was also proposed[12]. An important graph based V-SLAM system is PTAM by Klein and Murray[13], which creatively introduced the idea of running camera tracking and mapping in parallel threads, bringing the first real-time performance to optimization-based V-SLAM. The ORB-SLAM, based on the idea of PTAM and proposed by Mur-Artal et al.[14], is currently considered as the state-of-the-art of feature based SLAM method. ORB-SLAM introduced an automatic map initialization with model selection on homography or fundamental matrix based ego-motion calculated using RANSAC. It used the ORB feature detector and descriptor instead of the FAST corners and image patches in PTAM, improving the robustness of image tracking and feature matching under scale and orientation changes. Furthermore, a visual bag-of-words based loop closing module was implemented in parallel to tracking and local mapping threads. To handle large-scale maps, multi-scale mapping strategies were introduced, including a local graph for pose bundle adjustment, a co-visibility graph for local bundle adjustment, and an essential graph for global bundle adjustment after loop closure.

Direct methods skip the feature extraction step and work with the raw pixel intensities. Specifically, the states are estimated by minimizing the photometric errors, i.e. the pixel intensity differences. DTAM by Newcombe et al.[15] is a dense-direct system which exploits all the pixel values in the image, even from areas where gradients are small. Detailed textured depth maps at selected keyframes are estimated to produce a surface patchwork with millions of vertices, and a global spatially regularized energy function is minimized in an optimization framework. Although it can outperform feature-based methods in scenes with poor texture and motion blur, a GPU is required for real-time performance. Semi-dense methods like LSD-SLAM[16] can run real-time on a CPU, by only estimating depth at pixels only near strong-gradient boundaries. The tracking is performed by SE(3) image alignment using a coarse-to-fine algorithm with a robust Huber loss. The map optimization is executed using conventional graph optimization. Exploiting sparsity further, a direct sparse method, DSO, was proposed by Engel et al[17]. This method does not consider the smoothness prior used in other direct methods but carries out even sampling of pixels throughout the images. DSO also considers lens vignetting, exposure time, camera calibration and non-linear response functions. DSO runs faster than both feature based and semi-dense direct methods. Meanwhile, there are also hybrid methods that combines the strategies of feature based and direct methods. SVO by Forster et al.[18] is a semi-direct approach that tracks and triangulates pixels characterized by high image gradients

by direct methods, but jointly optimizes structure and motion using feature-based methods. Semi-direct methods avoid extracting features in every frame during tracking, but maintains the good invariance to viewpoint and illumination changes of image features in keyframes, and hence demonstrated efficient and robust performance.

Apart from the traditional model based methods, in recent years SLAM methods were developed based on learning approaches or semantic information. The following several directions can be observed in this area. Firstly, at the level of landmarks, semantic objects rather than simple 3D points are used to represent the environmental structure and form the graph, as performed by SLAM++[19]. Secondly, depth information can be learned directly from single monocular images using deep convolutional neural networks (CNN)[20]. Finally, at the whole-graph level, deep CNN feature descriptions are utilized to do loop closure detection[21]. Semantic SLAM is still under enthusiastic investigation.

To achieve more accurate and robust estimation, vision is usually fused with other sensors. One popular combination is to use RGBD (RGB image + Depth) sensors, the most famous one of which is Microsoft Kinect. ORB-SLAM2[22] extends its original monocular version to both stereo and RGBD modalities, offering the similar map reuse, loop closing, and relocalization capabilities. ElasticFusion[23] achieves real-time dense localization and mapping without common pose graph optimization, using dense frame-to-model camera tracking and windowed surface-based fusion coupled with frequent model refinement through non-rigid surface deformations. Fig. 1 shows the dense 3D structure of an office using the RGBD SLAM. RGBD SLAM, however, due to the restricted working range of depth sensors, is not applicable in large-scale applications.

Fig. 1. The constructed office model using RGBD SLAM.

Visual and inertial or odometry measurements offer complementary properties for fusion, and visual inertial SLAM (VI-SLAM) has been an active research topic in recent years. The MSCKF, namely Multi-State Constraint Kalman Filter[24], represents an important contribution in the area of filtering-based VI-SLAM. This approach uses nonlinear triangulation of landmarks observed from a set of camera poses over time to determine their positions so that the landmark positions are not included in the state vector in the EKF update. The state-of-the-art graph optimization based method is OKVIS by Leutenegger et al.[25], which uses a rigorously probabilistic cost function defined on the re-projection errors of landmarks and the inertial terms. To ensure real-time operation, the optimization is limited to a bounded window of keyframes through marginalization, and the consistency in the marginalization is maintained by using first-estimate Jacobians. Mur-Artal et al.[26] used the similar optimization framework to OKVIS, and additionally added the capability to close loops and reuse a map of an already mapped environment. Foster et al.[27] complemented the inertial measurement propagation model used in graph optimization, by introducing a pre-integration theory that properly addresses the manifold structure of the rotation group including its noise, which can be seamlessly integrated into a graph-based visual-inertial pipeline to improve both the efficiency and accuracy.

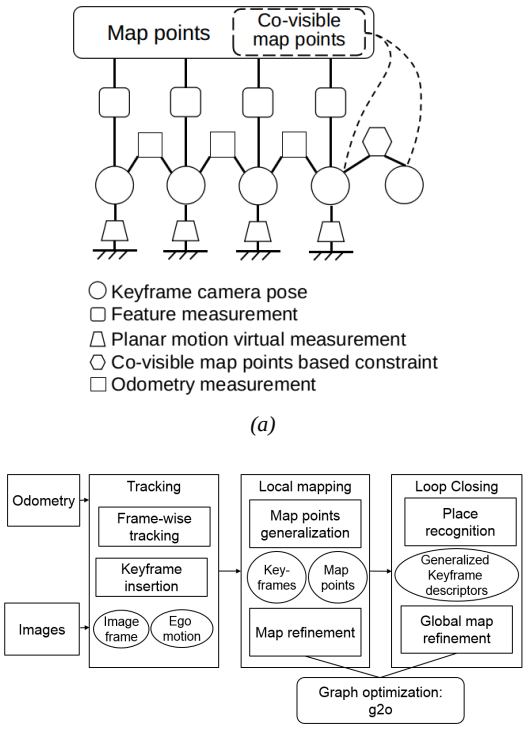


Fig. 2. *(a)* The graph implemented in the system; *(b)* The overall structure of the system.

To apply SLAM methods to real-world applications [28], the estimation accuracy and robustness remain challenging. In real-world applications, customization of the algorithm with the nature and constraints of the problem is important. For this, we developed an odometry-vision based SLAM system for ground vehicles. Unlike most vi-

sual SLAM systems modeling the vehicle poses in the general SE(3) space, we utilize the planar motion constraints of ground vehicles as system priors and propose a novel SE(2)-constrained SE(3) pose parameterization method. In this approach, a graph optimization framework is built to estimate the vehicle poses and the environmental landmark positions simultaneously. As validated by real-world experiments, our method produced better accuracy than the previous ones.

As shown in Fig. 2*(a)*, we implemented four types of constraints in the graph optimization, including 1) the feature based constraint which utilizes the measurement of the detected image feature locations; 2) the planar motion constraint modeled by projecting the SE(3) pose to a virtual SE(2) plane but allowing certain perturbations of motion around the plane; 3) the odometry based constraint between two consecutive keyframes; and 4) the co-visible map points based constraint between two keyframes that observe a bundle of same landmarks. These constraints are generated and optimized in a real-time system as illustrated in Fig. 2*(b)*, which runs in three parallel threads. The tracking thread takes in the sensor data of images and odometry measurements, detects and tracks image features and generates initial values for the graph optimization. The local mapping thread performs the optimization on a local map, which consists of certain number of latest keyframes and their observed map points. To reduce accumulated errors, the loop closing thread detects whether the vehicle visits a place it has visited before, and corrects the whole pose graph if so. The proposed SLAM system is tested on a forklift AGV as shown in Fig. 3*(a)*. The AGV is equipped with 2 SICK encoders and one Fizoptika fiber optic gyro to generate odometry measurement, and one PointGrey camera to capture images of the ceiling at 25 Hz. The system can run in real time on an Intel i7 laptop CPU. We tested the algorithm in an industrial warehouse of around $40 \times 60\ m^2 \times 6\ m$, named Dataset Warehouse. To validate the better performance of our system, we compare it against one state-of-the-art visual SLAM method, e.g. ORB-SLAM.

Fig. 3*(b)* shows one example of the feature tracking results in real time, and Fig. 3*(c)* demonstrates the mapping result in the Dataset Room. The mapping result indicates that the keyframes are constrained to a plane as expected. The statistical results are demonstrated in Fig. 4. It should be noted that the z-coordinates estimated by our method is well constrained around zero, which conforms to the real indoor environments, while ORB-SLAM gives deviating values. Overall, the results by our method exhibit much better accuracy and smaller upper limit of errors. In terms of Root-Mean-Square of the errors (RMSE) in the estimated locations, the results of our method yield one-loop accuracy of around 0.086%, which is applicable in many industrial indoor navigation tasks.

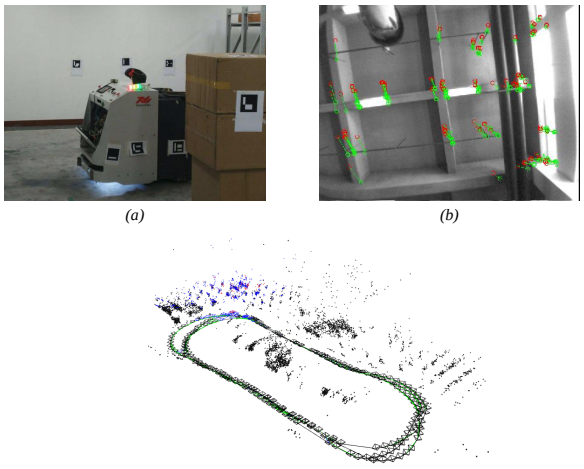


Fig. 3. *(a)* The AGV platform; *(b)* Image features tracked; *(c)* the map constructed.

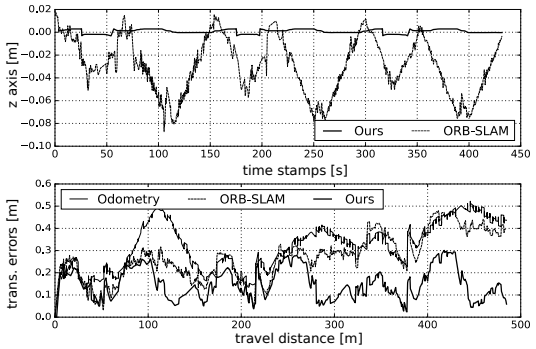


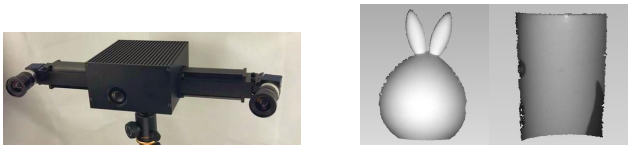
Fig. 4. Estimated z-coordinates and translation errors in the warehouse.

III. Manipulation Intelligence

Manipulation intelligence for a robot includes the capability and skills of grasping, transporting, orienting, positioning, and assembling objects using hands. The research topics include design of robot hands, grasp planning, and manipulation planning and control. Designs of robot hands are certainly crucial in performing manipulation tasks, but in this paper we focus the discussions on planning and control.

Grasp planning is a problem of planning the grasp points faces of a robot on an object. When the position and orientation of a rigid object to be grasped are known, many literatures can be found on planning of the stable grasps using parallel grippers and multi-fingered hands[29−36]. There are still two challenging problems in this area: grasp planning on deformable objects and robust grasp of randomly positioned objects. Grasp planning on soft objects is challenging due to the difficulty of

modeling their deformation and quantitatively defining a stable grasp on them. A possible method to solve grasp planning of soft objects could be to use learning from human-intelligence. By using a dataset to train a stable grasp network based on deep learning or other methods, it could be possible for robots to know the grasp immediately. The robust grasp of randomly positioned objects involves in sensing and perception, and grasp planning in a natural environment. The famous Amazon challenges are to grasp both soft and rigid objects in store settings. Sensing is to measure 3D geometry and position of an object in a real world, and the measurement accuracy is crucial for success planning and execution of the grasping tasks. The accuracy for successful grasping is at the level of 1-2 millimeters. 3D measurement is an old problem in computer vision. Stereo vision provides a simple solution, but its reliability and low accuracy are not suitable for real world applications. The method using structured light provides a solution for accurate 3D measurement, but its computational cost is a burden for real-time applications. We have recently developed a system that combines stereo vision with structured light for real-time and precise 3D measurement[37]. The system can generate 3 million points with an accuracy of 20 micro meters at a speed of 50fps. Fig. 5 shows the system and the point clouds captured by it.



(a) (b)

Fig. 5. *(a)* The real-time 3D imaging system; *(b)* The point clouds captured by the system.

Deforming soft objects using robots is highly demanded in many industrial and service applications such as surgery, food processing, cloth handling, soldering and assembly of flexible PCBs, etc.[35−39]. The major obstacles to robotic deformation control include kinematic and dynamic modeling of soft objects, which is crucial to the controller design and stability analysis. The deformation of a soft object depends on not only the materials but also its geometry, boundary conditions, etc. Deformation control has been studied since early 1990s for their many potential applications in industry and other sectors[35−40]. Sun and Liu studied modeling, position control and impedance control of flexible beam using multiple robots[41]. They proposed a simple and effective controller for regulating position of the beam without using any deformation or model, and proved the stability based on the deformation model. They further extended the controller to position control of a general flexible object[42]. It should be pointed out that they control position of the object rather than the deformation. Most of existing controllers for explicit deformation control of soft objects is designed on the basis of a deformation model. Provided that the model parameters are identified, the controller in[43] controls the shape of a rheological object using different approaches in the elasticity and plasticity phases. Shibata and Hirai[44−45] developed model-free methods for controlling one dimensional linear deformation of soft objects. We also designed a controller for controlling deformation of a 2D beam using two manipulators[46]. However, the controllers can be hardly extended to general 2D or 3D deformation control. There are works on robotic control of 2D or 3D deformation based on different approximated models on the deformation. For example, based on an approximated model, Hirai and his group proposed vision-based approaches to indirectly control deformation of a soft object by controlling a number of points of interest on the image plane of the vision system[47−48]. The work in [49] is similar. Tokumoto and Hirai, in [50], further investigated the problem of shaping food dough using a forming machine subject to known deformation parameters of the object. In [51], Das and Sarkar employed the finite element model (FEM) to design the controller for deforming the shape of a body by multiple manipulators. The difficulties in using a FEM model include identification of the physical parameters and high computational cost. Zacharia et al.[52] investigated vision-based cloth-handling problems in which dynamic interactions between cloth and a robot can be neglected for small mass of cloth. Foresti and Pellegrino[53] studied vision-based grasping of deformable objects, which is different from manipulation. The controllers in [54 − 58] are based on an exact deformation model and a model on interaction forces between the robots and the object. The work[59] investigated real-time tracking of deformation of objects using an eye-in-hand system, but did not consider the control problem. The stability can be hardly proven for the trajectory tracking controller of a flexible plate based on the neural network[60]. An essential procedure in robotic surgery is to manipulate soft tissues, and hence efforts are being extensively made to sensing, modeling and manipulating soft tissues[61−64]. Due to difficulties in deformation modeling, the research on robotic manipulation of soft tissues is still in the preliminary stage and there is no effective solution to this difficult problem.

An important effort being made by us is to design a model-less deformation controller using visual servoing. Visual servoing is an approach of controlling robot motion using visual information captured by vision system of the robot[65−71]. This approach is similar to what human does. We can easily deform a soft object as we want just

by observing deformation of the object using eyes. The objective of our effort is to develop intelligence similar to human for manipulation of soft objects. Vision-based deformation control is illustrated in Fig. 6. A robot manipulator is to manipulate a soft object to a desired shape with assistance of a vision system, which can be an either 2D or 3D system. There are several critical issues in controlling the deformation: description or representation of the shape, kinematic and dynamic model of the deformation, controller design, and stability analysis.

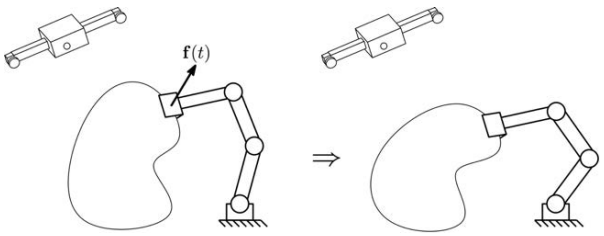


Fig. 6. Deformation control of soft objects using visual feedback.

As mentioned previously, it is difficult to obtain a model for modeling the deformation kinematics and dynamics, so deformation control without models is highly desirable. When the models are not used, it is necessary to develop algorithms to estimate the relationship between robot motion and deformation of the object, i.e. the deformation Jacobian matrix. As for shape description, it is not trivial to give a complete representation of the 3D shape of a soft object. Moreover, since a soft object is of an infinite number of DOF, it is not possible to deform a 3D surface to an arbitrary shape using a manipulator with a limited number of active joints. It is only possible to control deformation of a soft object at local areas or points of interest, which leads to a shape with the minimum deformation energy. We proposed to introduce a shape descriptor s(t) to represent the deformation of interest, which could be coordinates of a number of feature points, centroids of areas, angles between lines, curvatures, etc. Global features such as contour of the projection of a soft object can be also used as the shape descriptor.

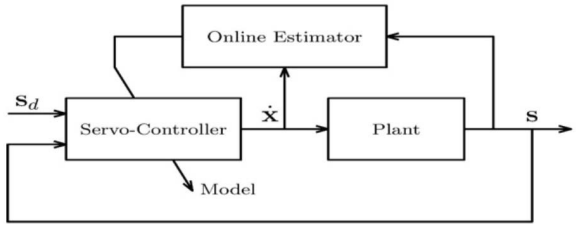


Fig. 7. The block diagram of the vision-based deformation control.

Fig. 7 shows the block diagram of a model-free deformation controller using the shape descriptor. Vector $\mathbf{s}(t)$ is a shape descriptor and $\mathbf{s}_d$ is the desired descriptor corresponding to the desired shape. An on-line estimator needs to be developed to estimate the deformation Jacobian matrix, which relates the deformation and the robot's motion as follows:

$$\dot{\boldsymbol{s}}(t) = \boldsymbol{J}_s \dot{\boldsymbol{x}}(t) \tag{1}$$

where represents the velocity of the end-effector of the manipulator. Navarro-Alarcon and Liu[72] proposed to employ the Broyden formula to estimate the deformation Jacobian matrix from 2D visual feedback so as to achieve model-free regulation of the feature points on the image plane[72] :

$$\widehat{\boldsymbol{J}}_s^T(t) = \widehat{\boldsymbol{J}}_s^\intercal(t-\Delta t) + \boldsymbol{\Gamma}\frac{\Delta \boldsymbol{s}(t) - \widehat{\boldsymbol{J}}_s^\intercal(t-\Delta t)\Delta \boldsymbol{x}(t)}{\boldsymbol{x}^\intercal(t)\boldsymbol{x}(t)} \tag{2}$$

where $\boldsymbol{Gamma}$ is a positive-definite adaptive gain. This algorithm iteratively estimates the deformation Jacobian matrix without knowing any deformation model. The concern is that the iterative estimation may stop at local minimums and hence the stability cannot be rigorously proved. Another estimator was developed based on the assumption that the deformation is approximated by nonlinear functions of position $\boldsymbol{x}(t)$ of the end-effector of the robot at qusai-static motion[73]. That is, the descriptor $\boldsymbol{s}(t)$ is approximately modeled as

$$\boldsymbol{s}(t) = \boldsymbol{\alpha}\boldsymbol{b}(\boldsymbol{x}(t)) \tag{3}$$

where $\boldsymbol{b}(\boldsymbol{x}(t)) \in \mathbb{R}^3$, being a known polynomial vector functions of $\boldsymbol{x}(t)$. $\boldsymbol{\alpha}$ is a constant $m \times 3$ coefficient matrix, where m is the dimension of $\boldsymbol{s}(t)$. Eq. (3) can be written as

$$\boldsymbol{s}(t) = \boldsymbol{W}(t)\boldsymbol{\theta}_1 \tag{4}$$

where $\boldsymbol{\theta}_1$ is a constant parameter vector corresponding to $\boldsymbol{\alpha}$, and matrix $\boldsymbol{W}(t)$ does not depend on the parameters $\boldsymbol{\theta}_1$. With this approximation, for any vector $\boldsymbol{\gamma}$,

$$\boldsymbol{J}_s^\intercal(t)\boldsymbol{\gamma} = \boldsymbol{Y}_s(\boldsymbol{x}(t), \boldsymbol{\gamma})\boldsymbol{\theta}_1 \tag{5}$$

where $\boldsymbol{Y}_s(\boldsymbol{x}(t), \boldsymbol{\gamma})$ is the regression matrix. Define the deformation error as follows:

$$\Delta \boldsymbol{s}(t) = \boldsymbol{s}(t) - \boldsymbol{s}_d \tag{6}$$

The deformation Jacobian matrix is estimated by estimating the parameters $\boldsymbol{\theta}_1$ on-line. Let $\widehat{\boldsymbol{\theta}}_1$ be an estimation of $\boldsymbol{\theta}_1$ and the corresponding estimation of the deformation Jacobian matrix be $\widehat{\boldsymbol{J}}_s^\intercal(t)$. Furthermore,

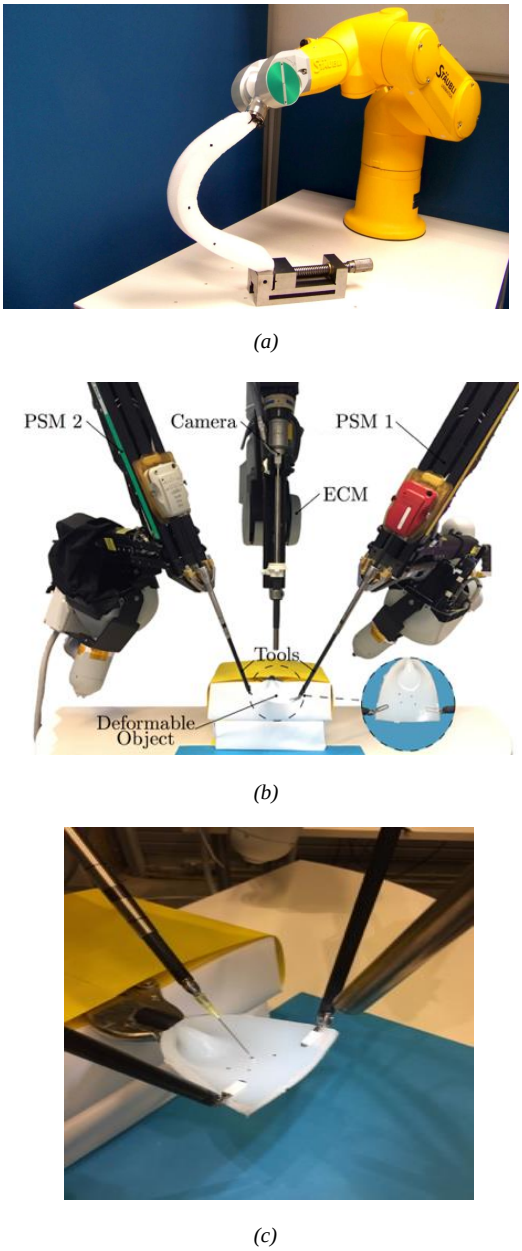
$$\left(\boldsymbol{J}_s^\intercal(t) - \widehat{\boldsymbol{J}}_s^\intercal(t)\right)\boldsymbol{K}\Delta \boldsymbol{s}(t) = \boldsymbol{Y}_s(\boldsymbol{x}(t))\Delta\boldsymbol{\theta}_1(t) \tag{7}$$

where $\Delta\boldsymbol{\theta}_1(t) = \widehat{\boldsymbol{\theta}}_1(t) - \boldsymbol{\theta}_1$, being the estimation errors. The controller is designed as follows:

$$\dot{\boldsymbol{x}}(t) = -\boldsymbol{K}\widehat{\boldsymbol{J}}_s^{\intercal}(t)\Delta\boldsymbol{s}(t) \tag{8}$$

where K is a positive-definite gain matrix. Substituting the controller (8) into eq. (1) leads to:

$$\dot{\boldsymbol{s}}(t) = -\boldsymbol{J}_s(t)\boldsymbol{K}\boldsymbol{J}_s^{\intercal}(t)\Delta\boldsymbol{s}(t) + \boldsymbol{J}_s(t)\boldsymbol{K}\left(\boldsymbol{J}_s^{\intercal}(t) - \widehat{\boldsymbol{J}}_s^{\intercal}(t)\right)\Delta\boldsymbol{s}(t) \tag{9}$$



(a) *(b)* *(c)*

Fig. 8. *(a)* Deformation control experiments conducted in our lab; *(b)* Experiments using an industrial arm; *(c)* Experiments using the da Vinci research kit.

Based on the closed-loop system, the unknown parameters are estimated by:

$$\dot{\widehat{\boldsymbol{\theta}}}_1 = -\boldsymbol{\Gamma}^{-1}\boldsymbol{Y}_s^{\intercal}(\boldsymbol{x}(t))\boldsymbol{J}_s(t)\Delta\boldsymbol{s}(t) \tag{10}$$

where $\boldsymbol{\Gamma}$ is a positive-definite adaptive gain. The asymptotic stability of the system can be proved by introducing the following positive-definite function:

$$V = \frac{1}{2}\Delta\boldsymbol{s}^{\intercal}(t)\Delta\boldsymbol{s}(t) + \frac{1}{2}\Delta\boldsymbol{\theta}_1^{\intercal}(t)\Delta\boldsymbol{\theta}_1(t) \tag{11}$$

From eqs. (9) and (10), it is possible to obtain

$$\dot{V} = \Delta\boldsymbol{s}^{\intercal}(t)\boldsymbol{J}_s(t)\boldsymbol{K}\boldsymbol{J}_s^{\intercal}(t)\Delta\boldsymbol{s}(t) \tag{12}$$

Using the Barbalat Lemma, it is possible to prove the convergence of the deformation error when the matrix $\boldsymbol{J}_s(t)$ is full rank. The controller was further extended to control of 3D positions of the feature points of a soft object by coordination of multiple robots[74]. An energy-based method was also developed to estimate the deformation Jacobian matrix using the visual deformation flow other than the visual displacements. In a recent work[75], we proposed to use the Fourier coefficients of the 2D contour of an object as the visual features to control the 2D deformation. Fig. 8 shows the deformation control experiments using an industrial arm and the da Vinci research kit[76−77]. The experiments conducted by us demonstrated good performance of the model-free visually servod controller for deformation control of soft objects.

IV. Intelligence for Human-Robot Interactions

Human-robot interaction (HRI) is crucial for robots co-working with humans in an environment. HRI involves two aspects: recognition of human actions/behaviors and expression of robot's ideas, etc. to humans. While the second topic is certainly important, this paper focuses the discussions to recognition of human actions.

Human action recognition has attracted lots of attention from different research fields, such as computer vision, robotics, etc. for its various potential applications ranging from elderly caring robots to surveillance. Action recognition was first studied based on RGB video inputs[78−79]. The research is alleviated to 3D action recognition by the development of the low-cost and real time depth cameras such as the Kinect sensor providing both color and depth images. Shotton[80] proposed a method for calculating 3D coordinates of the human skeleton using depth images. Since then, research of action recognition based on skeleton data has become popular[81−86]. Compared with color images, skeleton data is more robust to changes of the view and the light conditions. This paper uses dangerous behavior recognition of a child as the example to review the relevant works and demonstrate our efforts in this area as well. Fig. 9

shows the overall design of the system for detecting dangerous behaviors using a Kinect sensor. The RGB-D images and the human skeleton data of the Kinect sensor are employed to detect the dangerous objects, track motion and pose of the child, and recognize his/her dangerous behaviors.

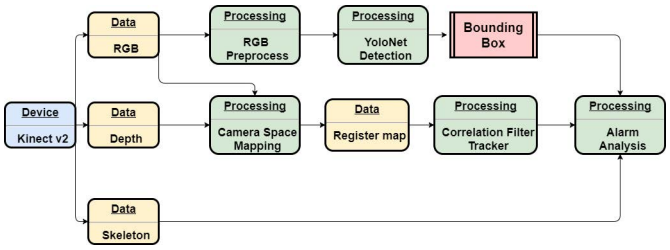


Fig. 9. The design of the dangerous behavior detection system.

One of the key issues in the system is to estimate pose of a child. Human pose estimation has been investigated for decades and most early research efforts were focused on locating the joints using 2D RGB images[87−91]. With the availability of low-cost depth cameras, extracting human skeletons from depth images become popular. Schwarz et al.[92] proposed a human skeleton tracking method from depth data using geodesic distances and optical flow. Shotton et al.[80] studied real-time tracking of a skeleton model with 25 joints from a single depth image using a randomized decision forests algorithm. This approach has been applied to Kinect v2, a commercial sensor of the MicroSoft. However, motion limits of human joints and occlusions are not taken in account.

With the success of convolution neural network (CNN) in image recognition and classification, various CNN-based methods have been developed for skeleton pose estimation with better performance than traditional methods. Tompson et al.[93] used monocular images to estimate human pose based on a hybrid architecture composed of a deep CNN and a Markov Random Field (MRF). A heat-map representing the per-pixel likelihood for key joint locations of the human skeleton is generated from a deep ConvNet in this method. Wei et al.[94] introduced a method to recurrently use the generated heat-map and original images for multi-stages training. Their method has improved the performance significantly.

No matter what methods are used, robust and stable 2D skeleton pose tracking is yet a problem to be solved in computer vision and robotics. Invalid poses are generated from time to time, especially when occlusions occur. To avoid invalid skeleton estimation, Akhter et al.[95] collected a motion capture dataset including an extensive variety of stretching poses and studied how joint-limits vary with human pose. An over-completed pose dictionary is built based on the dataset, and then employed to generate an estimated pose with a parametrization method. It is computationally expensive and costly to collect all possible stretching poses. Instead of extensive data collection, the joint limits can be actually estimated based on a kinematics model and the bio-constraints of human body.

In practice, skeleton data obtained from Kinect often contains a lot of errors, especially in the case of occlusions, as shown in Fig. 10. The skeleton model used in Kinect is not a strictly bio-constrained kinematics model. Here the bio-constraints mean the kinematic parameters of human body, such as joint motion limits, etc. Using the constraints will help ensure generation of valid poses.

Fig. 10. Corrupted skeletons obtained from Kinect sensor.

We developed a method to recover the skeleton from the corrupted one from the Kinect. Fig. 11 shows the framework of the method. Kinect sensor generates two data streams: RGB stream and skeleton stream. The RGB stream is fed into a convolutional neural network to estimate the 2D poses by the method in [94] . A confidence value C1 of the RGB-based prediction is calculated by averaging the confidence values of all the predicted joints. The skeleton data is fed into an error detection module for detecting the error joints. Another confidence value C2 for the skeleton data is calculated as a ratio between the number of the correct joints number and the total number of the joints. The two skeletons are fused based on the confidence values. Then the final 3D pose is generated by optimally fitting the fused skeleton with a standard human skeleton model.

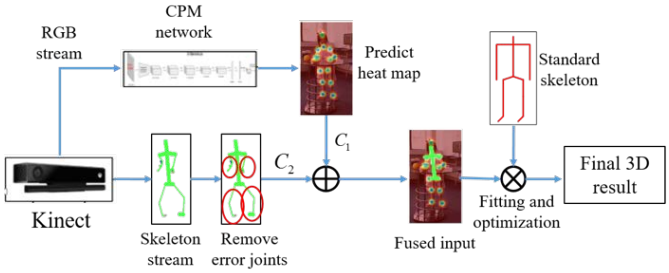


Fig. 11. The framework of proposed method that combines CNN and skeleton fitting.

An experimental result of this approach is shown in Fig. 12. The left picture shows the skeleton obtained from Kinect where the upper body is corrupted due to occlusion. As shown in the right picture of Fig. 12, the occluded left arm was estimated and the skeleton was recovered by the proposed approach. The results confirmed

that bio-constraints helped extract human skeleton more robustly.

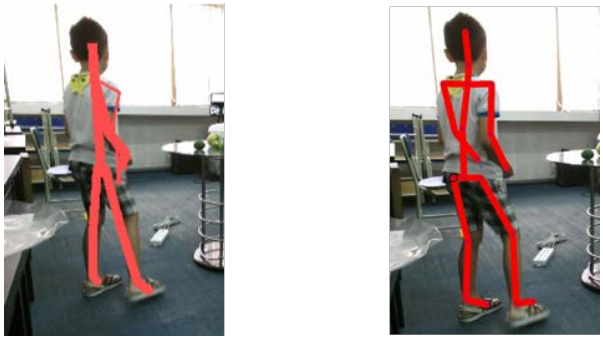


Fig. 12. Left: the original skeleton data, and right: the recovered skeleton.

Take the dangerous behavior of touching electrical sockets by a child in an indoor scenario as the example. Detection of the target object, i.e. electrical sockets, is another crucial technical problem. The state-of-the-art method for object detection uses the deep learning algorithms. The most classical algorithm is the RCNN series: RCNN[96] , fast-RCNN[97] , and faster-RCNN[98] . We can regard this type of methods as classification of many sliding windows using full convolution. There are also end to end object detection works, such as YOLO[99] and SSD[100] . In order to apply to a three-dimensional case, the inventor of the RCNN, Girshirk[101] extended their method to the depth image, and concluded that the raw depth image, as an input of the network, would produce better results than using depth features. Song[102] used the 3D reconstruction method to obtain the TSDF model of the scene and train it through a relatively shallow region proposal network. However, it produced poor accuracy for small objects and hence is not suitable for detecting the indoor sockets.

In the detection module, we used YOLONet2[103] network structure. For the feature extraction part, the network is based on the Darknet-19 basic classification model, including 19 full convolution layers and 5 maximum pooling layers. Darknet-19 requires 5.58 billion operations to process a photo. Like VGG, the network uses more convolution kernels, and doubles the number of channels after each pool. The network employs the global average pooling method to predict, and places the convolution kernel between the convolution kernels to compress the features. In addition, we used batch normalization to stabilize the model, speed up the convergence rate, and regularize the model. After using ImageNet to pre-train, we modified it to a detection network. After removing the output layer of the last convolution kernel, adding two layers of convolution kernel, and implementing a passthrough layer, the output in the 16th layer convolution layer of Darknet-19 and the last convolution layer are concatenated. The passthrough layer superimposed high and low-resolution feature maps according to adjacent features at different channels. After the feature is concatenated, the fused feature map passes through a layer of convolution kernel. Finally, we obtain the location and category of the object.

In the filtering tracker module, we use the correlative filter to predict and locate the object. In this application, the child in the scene will be an interference to the measured socket, resulting in a fluctuated detection rate. We use the correlative filtering method to solve the problem. Relevance is mainly used to describe the relationship between the two factors. For the detection window, we need to find a filter to achieve the maximum response in the detection window. In the time domain, this calculation is a convolution, which is time consuming, while in the frequency domain the calculation becomes a matrix pixel-wise multiplication. Therefore, we can find the desired filter in the Fourier frequency domain to maximize the response.

The task is to find an expected filter :

$$\boldsymbol{H^*} = \frac{\boldsymbol{G}}{\boldsymbol{F}} \tag{13}$$

where $\mathbf{G_i}$ is a Gaussian kernel and $\mathbf{F_i}$ is an image patch in the frequency domain. Considering the problem of the attitude transformation and update of time series, we take m time-sequence detection windows as references to improve the filter robustness,

$$\min_{\boldsymbol{H^*}} \sum_{i=1}^{m} |\boldsymbol{F_i} \circ \boldsymbol{H^*} - \boldsymbol{G_i}|^2 . \tag{14}$$

By deriving every element in filter , we have,

$$\frac{\partial}{\partial H^*_{wv}} \sum_{i=1}^{m} |F_{iwv} H^*_{wv} - G_{iwv}|^2 = 0 \tag{15}$$

In details, we have,

$$\frac{\partial}{\partial H^*_{wv}} \sum_{i=1}^{m} (F_{iwv} H^*_{wv} - G_{iwv}) (F_{iwv} H^*_{wv} - G_{iwv})^* = 0 \tag{16}$$

$$\frac{\partial}{\partial H^*_{wv}} \sum_{i=1}^{m} (F_{iwv} H^*_{wv}) (F_{iwv} H^*_{wv})^* - (F_{iwv} H^*_{wv}) G^*_{iwv} - G^*_{iwv} (F_{iwv} H^*_{wv})^* + G_{iwv} G^*_{iwv} = 0 \tag{17}$$

Computing the partial derivatives leads to

$$\sum_i [F_{iwv} F^*_{iwv} H_{wv} - F_{iwv} G^*_{iwv}] = 0 \tag{18}$$

Therefore, we have H_{wv} as follows:

$$H_{wv} = \frac{\sum_i F_{iwv} G^*_{iwv}}{\sum_i F_{iwv} F^*_{iwv}} \tag{19}$$

Finally, we rewrite the expression in matrix form,

$$\boldsymbol{H} = \frac{\sum_i \boldsymbol{F_i} \boldsymbol{G_i^*}}{\sum_i \boldsymbol{F_i} \boldsymbol{F_i^*}} \tag{20}$$

$\boldsymbol{G_i}$ is a Gaussian kernel, and the initial $\boldsymbol{F}_i$ can be obtained by the random affine transformation. To improve the robustness of the filter to deformation and illumination, and the computational efficiency, we separate the template update strategy into $\boldsymbol{A}$ and $\boldsymbol{B}$:

$$\boldsymbol{H}_i^* = \frac{\boldsymbol{A}_i}{\boldsymbol{B}_i} \tag{21}$$

$$\boldsymbol{A}_i = \eta \boldsymbol{F}_i \boldsymbol{G}_i^* + (1-\eta)\boldsymbol{A}_{i-1} \tag{22}$$

$$\boldsymbol{B}_i = \eta \boldsymbol{F}_i \boldsymbol{F}_i^* + (1-\eta)\boldsymbol{B}_{i-1} \tag{23}$$

For the depth image of the scene, we have projected the coordinates from the registration mapping to the camera coordinate. For each point in the registration map, we map it to a 3-dimensional vector.

$$\boldsymbol{P_i} = [x_i \quad y_i \quad z_i] \tag{24}$$

For a single patch j with size of $m \times n$ in the registration map, we first calculate the mean value of a patch with a size of $m \times n \times 3$, and then the offset matrix $\boldsymbol{K}$ from the mean value:

$$\boldsymbol{K} = \left\{ [\boldsymbol{P}_i - \begin{bmatrix} x_c \\ y_c \\ z_c \end{bmatrix} \right\}_{m \times n} \tag{25}$$

Fig. 13. The detection results: normal light (top left); dim light (top right); small objects (bottom left); occlusion (bottom right).

After regularizing K, we achieve the approximate surface feature of the object. The feature has three coordinates x, y, and z with respect to the camera coordinate frame. We use the feature maps in the y-direction (height) and z-direction (depth) and calculate the ratio of the peak and the adjacent peak (PSR) to judge the status of sockets. When the center of the object detection area moves forward, the z-direction changes significantly, and the PSR is greatly reduced, we consider the socket in an occlusion state. When the motion of the center point of the object detection area and the change in the z-direction are not obvious, and the PSR significantly reduced, we consider the socket is in a moving state. When the motion of the center point of the object detection area and the change in the z-direction are not obvious, and the change of the PSR is not large, we consider that the socket does not change. In addition, the filter templates refresh every 40 frames.

As shown in Fig. 13, by pre-processing the image, such as random cropping and color shifting, we could cope with problems due to various illumination changes and partly occlusions. However, to detect small objects, especially the sockets on the wall, the algorithm cannot meet the requirements. Thus, we applied a trick that crops the original image into 4 parts and then used non-maximal suppression to deal with redundant detection bounding boxes. The results are shown in Fig 14. We have carried out experiments in which a child is to touch nine sockets five times in four different scenarios. The total detection rate was 86.7% and the details are given in Table 1. The failures occurred mainly when the sockets were small seen from the camera or occluded. More efforts must be made to improving the success rates.

Fig. 14. The original image (top left); the cropped image (top right); the detection results (bottom left); the detection results with non-maximum suppression (bottom right).

Table 1. The experiments on touching sockets

Area	Socket	True	False	Accuracy
1	1	5	0	100%
	2	5	0	100%
	3	4	1	80%
2	4	3	2	60%
	5	4	1	80%
3	6	5	0	100%
	7	5	0	100%
4	8	4	1	80%
	8	4	1	80%

V. Conclusions

This paper presents a short review on robot intelligence for real world applications, in particular on the intelligence involving action or body motion. Specifically, we reviewed the works on simultaneous localization and

mapping, manipulation of soft objects, and human actions tracking and recognition. Although some important progress is being achieved to those important topics, a lot of efforts must be made to improvement of performance of the approaches including the reliability, the robustness to noises or disturbances, the accuracy, the real-time efficiency in order to be applicable in real worlds. Our efforts in those areas were also briefly introduced.

References

[1] `http://www.wikipedia.org/wiki/Artificial_intelligence.`

[2] H. Gardner, *Multiple Intelligences: Theory In Practice*, 1993.

[3] H. Durrant-Whyte and T. Bailey, "Simultaneous localization and mapping: Part I", *IEEE Robotics and Automation Magazine*, Vol.13, No.2, pp.99-110, 2006.

[4] R. Kmmerle, G. Grisetti, H. Strasdat, *et al.*, "G2o: A general framework for graph optimization", *Proc. of IEEE International conference on Robotics and Automation*, Shanghai, pp.3607-3613, 2011.

[5] S. Thrun, W. Burgard and D. Fox, "A probabilistic approach to concurrent mapping and localization for mobile robots", *Autonomous Robots*, Vol.5, No.3-4, pp.253-271, 1998.

[6] M. Montemerlo, S. Thrun, D. Koller, *et al.*, "FastSLAM: A factored solution to the simultaneous localization and mapping problem", *AAAI/IAAI*, 2002.

[7] P. Newman, J. Leonard, J. D. Tardos, *et al.*, "Explore and return: experimental validation of real-time concurrent mapping and localization", *Proc. of IEEE International conference on Robotics and Automation*, Vol.2, pp.1802-1809, 2002.

[8] J. J. Leonard, R. J. Rikoski, P. M. Newman, *et al.*, "Mapping partially observable features from multiple uncertain vantage points", *The International Journal of Robotics Research*, Vol.21, No.10-11, pp.943-975, 2002.

[9] L. Carlone, R. Aragues, J. A. Castellanos, *et al.*, "A fast and accurate approximation for planar pose graph optimization", *The International Journal of Robotics Research*, Vol.33, No.7, pp.965-987, 2014.

[10] W. Hess, D. Kohler, H. Rapp, *et al.*, "Real-time loop closure in 2D LIDAR SLAM", *Proc. of IEEE International conference on Robotics and Automation*, Stockholm, pp.1271-1278, 2016.

[11] A. J. Davison, I. D. Reid, N. D. Molton, *et al.*, "MonoSLAM: Real-Time Single Camera SLAM", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.29, No.6, pp.1052-1067, 2007.

[12] M.H. Li, B.R. Hong and R.H. Luo, "Mobile robot simultaneous localization and mapping using novel rao-blackwellised particle filter", *Chinese Journal of Electronics*, Vol.16, No.1, pp.3439, 2007.

[13] G. Klein and D. Murray, "Parallel Tracking and Mapping for Small AR Workspaces", *2007 6th IEEE and ACM International Symposium on Mixed and Augmented Reality*, pp.225-234, 2007.

[14] R. Mur-Artal, J. M. M. Montiel and J. D. Tards, "ORB-SLAM: A Versatile and Accurate Monocular SLAM System", *IEEE Transactions on Robotics*, Vol.31, No.5, pp.1147-1163, 2015.

[15] R. A. Newcombe, S. J. Lovegrove and A. J. Davison, "Dtam: Dense tracking and mapping in real-time", *Proc. of IEEE Conference on Computer Vision*, pp.2320-2327, 2011.

[16] J. Engel, T. Schops and D. Cremers, "Lsd-slam: Large-scale direct monocular slam", *Proc. of European Conference on Computer Vision*, pp.834-849, 2014.

[17] J. Engel, V. Koltun and D. Cremers, "Direct Sparse Odometry", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.40, No.3, pp.611-625, 2018.

[18] C. Forster, Z. Zhang, M. Gassner, *et al.*, "SVO: Semidirect Visual Odometry for Monocular and Multicamera Systems", *IEEE Transactions on Robotics*, Vol.33, No.2, pp.249-265, 2017.

[19] R. F. Salas-Moreno, R. A. Newcombe, H. Strasdat, *et al.*, "Slam++: Simultaneous localisation and mapping at the level of objects", *Proc. of IEEE International conference on Computer Vision and Pattern Recognition*, pp.1352-1359, 2013.

[20] F. Liu, C. Shen, G. Lin, *et al.*, "Learning Depth from Single Monocular Images Using Deep Convolutional Neural Fields", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.38, No.10, pp.2024-2039, 2016.

[21] Y. Hou, H. Zhang and S. Zhou, "BoCNF: efficient image matching with Bag of ConvNet features for scalable and robust visual place recognition", *Autonomous Robots*, pp.1-17, 2017.

[22] R. Mur-Artal and J. D. Tards, "ORB-SLAM2: An Open-Source SLAM System for Monocular, Stereo, and RGB-D Cameras", *IEEE Transactions on Robotics*, Vol.33, No.5, pp.1255-1262, 2017.

[23] T. Whelan, R. F. Salas-Moreno, B. Glocker, *et al.*, "ElasticFusion: Real-time dense SLAM and light source estimation", *The International Journal of Robotics Research*, Vol.35, No.14, pp.1697-1716, 2016.

[24] A. I. Mourikis and S. I. Roumeliotis, "A Multi-State Constraint Kalman Filter for Vision-aided Inertial Navigation", *Proc. of IEEE International conference on Robotics and Automation*, pp.3565-3572, 2007.

[25] S. Leutenegger, S. Lynen, M. Bosse, *et al.*, "Keyframe-based visualinertial odometry using nonlinear optimization", *The International Journal of Robotics Research*, Vol.34, No.3, pp.314-334, 2015.

[26] R. Mur-Artal and J. D. Tards, "Visual-Inertial Monocular SLAM With Map Reuse", *IEEE Robotics and Automation Letters*, Vol.2, No.2, pp.796-803, 2017.

[27] C. Forster, L. Carlone, F. Dellaert, *et al.*, "On-Manifold Preintegration for Real-Time Visual–Inertial Odometry", *IEEE Transactions on Robotics*, Vol.33, No.1, pp.1-21, 2017.

[28] H. Wu, Y. Wu, C. Liu, *et al.*, "Visual data driven approach for metric localization in substation", *Chinese Journal of Electronics*, Vol.24, No.4, pp.795-801, 2015.

[29] A. Bicchi, "On the closure properties of robotic grasping", *The International Journal of Robotics Research*, Vol.14, No.4, pp.319-334, 1995.

[30] M. Buss, H. Hashimoto and J. Moore, "Dexterous hand grasping force optimization", *IEEE Transactions on Robotics and Automation*, Vol.12, pp.406-418, 1996.

[31] D. Ding, Y. H. Liu, Y. T. Shen, *et al.*, "An efficient algorithm for computing a 3D form-closure grasp", *Proc. of IEEE/RSJ Conference on Intelligent Robotics and Systems*, Vol.2, pp.1223-1228, 2000.

[32] L. Han, J. C. Trinkle and Z. Li, "Grasp analysis as linear matrix inequality problems", *IEEE Transactions on Robotics and Automation*, Vol.16, pp.663-674, 2000.

[33] M. L. Lam, D. Ding and Y. H. Liu, "Grasp planning with kinematic constraints", *Proc. of IEEE/RSJ Conference on Intelligent Robotics and Systems*, Vol.2, pp.943-948, 2001.

[34] Y. H. Liu, "Qualitative test and force optimization of 3-D frictional form closure grasps using linear programming", *IEEE Transactions on Robotics and Automation*, Vol.15, pp.163-173, 1999.

[35] Y. H. Liu, "Computing n-finger form-closure grasps on polygonal objects", *The International Journal of Robotics Research*, Vol.18, No.2, pp.149-158, 2000.

[36] Y. H. Liu, M. L. Lam and D. Ding, "A complete and efficient algorithm for searching for 3-D form-closure grasps in discrete domain", *IEEE Transactions on Robotics*, Vol.20, No.5, pp.805-816, 2004.

[37] `www.smarteyetech.com.`

[38] D. Henrich and H. Worn Eds., "Robot manipulation of deformable object (Series advanced manufacturing)", New York, NY, USA: Springer-Verlag, 2000.

[39] V. Mallapragada, N. Sarkar and T. Podder, "Toward a robot-assisted breast intervention system", *IEEE/ASME Transactions on Mechatronics*, Vol.16, No.6, pp.1011-1020, 2011.

[40] J. Smolen and A. Patriciu, "Deformation planning for robotic soft tissue manipulation", *Proc. of IEEE International Conference on Advances in Computer-Human Interactions*, pp.199-204, 2009.

[41] D. Sun and Y. H. Liu, "Modeling and impedance control of a two-manipulator system manipulating a flexible beam", *ASME Journal of System, Dynamics, Measurement, and Control*, Vol.119, pp.736-742, 1997.

[42] D. Sun, J. Mills and Y. H. Liu, "Position control of multiple robots manipulating a general flexible object", *The International Journal of Robotics Research*, Vol.18, No.3, pp.319-332, 1999.

[43] M. Higashimori, K. Yoshimoto and M. Kaneko, "Active shaping of an unknown rheological object based on deformation decomposition into elasticity and plasticity", *Proc. of IEEE International conference on Robotics and Automation* pp.5120-5126, 2010.

[44] M. Shibata and S. Hirai, "Soft object manipulation by simultaneous control of motion and deformation", *Proc. of IEEE International conference on Robotics and Automation*, pp.2460-2465, 2006.

[45] M. Shibata, "Wiping motion for deformable object handling", *Proc. of IEEE International conference on Robotics and Automation*, pp.134-139, 2009.

[46] Y. H. Liu and D. Sun, "Stabilizing a flexible beam handled by two manipulators via PD feedback", *IEEE Transactions on Automatic Control*, Vol.45, No.11, pp.2159-2164, 2000.

[47] S. Hirai and T. Wada, "Indirect simultaneous positioning of deformable objects with multi-pinching fingers based on an uncertain model", *Robotica*, Vol.18, No.1, pp.311, 2000.

[48] T. Wada, S. Hirai, S. Kawamura, *et al.*, "Robust manipulation of deformable objects by a simple PID feedback", *Proc. of IEEE International conference on Robotics and Automation*, Vol.1, pp.85-90, 2001.

[49] J. Smolen and A. Patriciu, "Deformation planning for robotic soft tissue manipulation", *Proc. of IEEE International Conference on Advances in Computer-Human Interactions* , pp.199-204, 2009.

[50] S. Tokumoto and S. Hirai, "Deformation control of rheological food dough using a forming process model", *Proc. of IEEE International conference on Robotics and Automation*, Vol.2, pp.1457-1464, 2002.

[51] J. Das and N. Sarkar, "Autonomous shape control of a deformable object by multiple manipulators", *Journal of Intelligent and Robotic Systems*, Vol.62, pp.3-27, 2011.

[52] P. Zacharia, N. Aspragathos, I. Mariolis, *et al.*, "A robotic system based on fuzzy visual servoing for handling flexible sheets lying on a table", *Industrial Robot: An International Journal*, Vol.36, No.5, pp.489-496, 2009.

[53] G. L. Foresti and F. A. Pellegrino, "Automatic Visual Recognition of Deformable Objects for Grasping and Manipulation", *IEEE Transactions on Systems, Man and Cybernetics*, Vol.34, No.3, 2004.

[54] A.-M. Cretu, "Soft Object Deformation Monitoring and Learning for Model-Based Robotic Hand Manipulation", *IEEE Transactions on System, Man, and Cybernetics*, Vol.42, No.3, 2012.

[55] H. Yoshida, T. Fukuda, M. Sakai, *et al.*, "Manipulation of a flexible object by dual manipulators", *Proc. of IEEE International conference on Robotics and Automation*, pp.318-323, 1995.

[56] T. Yukawa, M. Uchiyama, *et al.*, "Handling of a constrained flexible object by a robot", *Proc. of IEEE International conference on Robotics and Automation*, pp.324-329, 1995.

[57] T. Kashiwase, M. Tabata, K. Tsuchiya, *et al.*, "Shape control of flexible structures", *Journal of intelligent material systems and structures*, Vol.2, No.1, pp.110-125, 1991.

[58] T. Yukawa, M. Uchiyama and D. N. Nenchev, "Stability of control system in handling of a flexible object by rigid arm robots", *Proc. of IEEE International conference on Robotics and Automation*, Vol.3, pp.2332-2339, 1996.

[59] M. J. Sullivan and N. P. Patpanikolopoulos, "Using active deformable models to track deformable objects in robotic visual servoing experiments", Proc. of IEEE International conference on Robotics and Automation, pp.2929-2934, 1996.

[60] F. Arai, R. Rong and T. Fukuda, "Trajectory control of flexible plate using neural network", *Proc. of IEEE International conference on Robotics and Automation*, pp.155-160, 1993.

[61] N. Sugita, F. Genma, Y. Nakajima, *et al.*, " Adaptive controlled milling robot for orthopedic surgery", *Proc. of IEEE International conference on Robotics and Automation* , Roma, Italy, pp.605-610, 2007.

[62] A. Suebsomran and M. Pamichaknn, "Disturbance observer-based hybrid control of displacement and force in medical teleanalyzer for abdominal mass analysis", *Proc. of IEEE International conference on Industrial Technology*, Vol.1, No.7, pp.365-369, 2002.

[63] P. Valdastri1, S. Tognarelli1, A. Menciassi, *et al.*, "A scalable platform for biomechanical studies of tissue cutting forces", *Measurement Science and Technology*, Vol.2673, No.10, pp.175-182, 2003.

[64] C. Mendoza and C. Laugier, "Tissue cutting using finite elements and force feedback", *Lecture Notes in Computer Science 2673, Surgery Simulation and Soft Tissue Modeling*, pp.175-182, 2003.

[65] S. Hutchinson, G. D. Hager and P. I. Corke, "A tutorial on visual servo control", *IEEE Transactions on Robotics and Automation*, Vol.12, No.5, pp.651-670, 1996.

[66] A. Astolfi, L. Hsu, M. Netto, *et al.*, "Two solutions to the adaptive visual servoing problem", *IEEE Transactions on Robotics and Automation* , Vol.18, No.3, pp.387-392, 2002.

[67] Y. H. Liu, H. Wang, C. Wang, *et al.*, "Uncalibrated visual servoing of robots using a depth-Independent interaction matrix", *IEEE Transactions on Robotics*, Vol.22, No.4, pp.804-817, 2006.

[68] H. Wang, Y. H. Liu and D. Zhou, "Dynamic visual tracking for manipulators using an uncalibrated fixed camera", *IEEE Transactions on Robotics*, Vol.23, No.3, pp.610-617, 2007.

[69] H. Wang, Y. H. Liu and D. Zhou, "Adaptive visual servoing using point and line features with an uncalibrated eye-in-hand camera", *IEEE Transactions on Robotics*, Vol.24, No.4, pp.843-857, 2008.

[70] K. Wang, Y. H. Liu and L. Y. Li, "Visually servoed trajectory tracking of nonholonomic mobile robots without direct position measurement", *IEEE Transactions on Robotics*, Vol.30, No.4, pp.1026-1035, 2014.

[71] K. Wang, Y. H. Liu and L. Li, "Vision based tracking control of underactuated water surface robots without direct position measurement", *IEEE Transactions on Control Systems Technology*, 2015.

[72] D. Navarro-Alarcon, Y. H. Liu, J. G. Romero, *et al.*, "Model-free visually servoed deformation control of elastic objects by robot manipulators", *IEEE Transactions on Robotics*, Vol.29, No.6, pp.1457-1468, 2013.

[73] D. Navarro-Alarcon, Y. H. Liu, J. G. Romero, *et al.*, "On the visual deformation servoing of compliant objects: uncalibrated control methods and experiments", *The International Journal of Robotics Research*, Vol.33, No.11, pp.1462-1480, 2014.

[74] D. Navarro-Alarcon, H. M. Yip, Z Wang, *et al.*, "Automatic 3D manipulation of soft objects by robotic arms with adaptive deformation model", *IEEE Transactions on Robotics*, Vol.32, No.2, pp.429-441, 2016.

[75] D. Navarro-Alarcon and Y. H. Liu, "Fourier-Based Shape Servoing: A New Feedback Method to Actively Deform Soft Objects into Desired 2D Image Contours", *IEEE Transactions on Robotics*, 2017.

[76] Z. Wang, S. C. Lee, F. Zhong,*et al.*, "Image-based trajectory tracking control of 4-DOF laparoscopic instruments using a rotation distinguishing marker", *IEEE Robotics and Automation Letters*, Vol.2, No.3, pp.1586-1592, 2017.

[77] D. Navarro-Alarcon, S. Singh, T. Zhang, *et al.*, "Developing a compact robotic needle driver for MRI-guided breast biopsy in tight environments", *IEEE Robotics and Automation Letters*, Vol.2, No.3, pp.1648-1655, 2017.

[78] H. Wang, A. Klser, C. Schmid, *et al.*, "Action recognition by dense trajectories", *Proc. of IEEE International conference on Computer Vision and Pattern Recognition*, pp.3169-3176, 2011.

[79] H. Wang and C. Schmid. "Action recognition with improved trajectories", *Proc. of the IEEE International Conference on Computer Vision*, pp.3551-3558, 2013.

[80] J. Shotton, T. Sharp, A. Kipman, *et al.*, "Real-time human pose recognition in parts from single depth images", *Proc. of IEEE International conference on Computer Vision and Pattern Recognition*, pp.1297-1304, 2011.

[81] X. Yang and Y. Tian, "Effective 3d action recognition using eigenjoints", *Journal of Visual Communication and Image Representation*, pp.2-11, 2014.

[82] C. Wang, Y. Wang and A. L. Yuille, "Mining 3d key-pose-motifs for action recognition", *Proc. of the IEEE International conference on Computer Vision and Pattern Recognition*, pp.2639-2647, 2016.

[83] R. Vemulapalli, F. Arrate and R. Chellappa, "Human action recognition by representing 3d skeletons as points in a lie group", *Proc. of the IEEE International conference on Computer Vision and Pattern Recognition*, pp.588-595, 2014.

[84] M. K. Pan, V. Skjervy, W. P. Chan, *et al.*, "Automated detection of handovers using kinematic features", *The International Journal of Robotics Research*, pp.721-738, 2017.

[85] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition", arXiv preprint arXiv:1409.1556, 2014.

[86] J. Wang, X. Nie, Y. Xia, *et al.*, "Cross-view action modeling, learning and recognition", *Proc. of the IEEE International conference on Computer Vision and Pattern Recognition*, pp.2649-2656, 2014.

[87] S. Johnson and M. Everingham, "Clustered pose and nonlinear appearance models for human pose estimation", *Proc. of the British Machine Vision Conference*, pp.1-11, 2010.

[88] M. Eichner, M. Marin-Jimenez, A. Zisserman, *et al.*, "2D articulated human pose estimation and retrieval in (almost) unconstrained still images", *International Journal of Computer Vision*, pp.190-214, 2012.

[89] B. Sapp, A. Toshev and B. Taskar, "Cascaded models for articulated pose estimation", *Proc. of European Conference on Computer Vision*, pp.406-420, 2010.

[90] Y. Yang and D. Ramanan, "Articulated pose estimation with exible mixtures-of-parts", *Proc. of IEEE International Conference on Computer Vision and Pattern Recognition*, pp.1385-1392, 2011.

[91] D. Mehta, S. Sridhar, O. Sotnychenko, *et al.*, "VNect: Real-time 3D human pose estimation with a single RGB Camera", arXiv preprint arXiv:1705.01583, 2017.

[92] L. A. Schwarz, A. Mkhitaryan, D. Mateus, *et al.*, "Human skeleton tracking from depth data using geodesic distances and optical flow", *Image and Vision Computing*, pp.217-226, 2012.

[93] J. J. Tompson, A. Jain, Y. LeCun , *et al.*, "Joint training of a convolutional network and a graphical model for human pose estimation", *Advances in Neural Information Processing Systems*, pp.1799-1807, 2014.

[94] S. E. Wei, V. Ramakrishna, T. Kanade, *et al.*, "Convolutional pose machines", *Proc. of IEEE International conference on Computer Vision and Pattern Recognition*, pp.4724-4732, 2016.

[95] I. Akhter and M. J. Black, "Pose-conditioned joint angle limits for 3D human pose reconstruction", *Proc. of IEEE International conference on Computer Vision and Pattern Recognition*, pp.1446-1455, 2015.

[96] R. Girshick, J. Donahue, T. Darrell, *et al.*, "Rich feature hierarchies for accurate object detection and semantic segmentation", *Proc. of IEEE International conference on Computer Vision and Pattern Recognition*, pp.580-587, 2014.

[97] R. Girshick, "Fast R-CNN", *Proc. of IEEE International conference on Computer Vision*, pp.1440-1448, 2015.

[98] S. Ren, K. He, R. Girshick, *et al.*, "Faster R-CNN: Towards real-time object detection with region proposal networks", *Advances in Neural Information Processing Systems*, pp.91-99, 2015.

[99] J. Redmon, S. Divvala, R. Girshick, *et al.*, "You only look once: Unified, real-time object detection", *Proc. of IEEE International conference on Computer Vision and Pattern Recognition*, pp. 779-788, 2016.

[100] W. Liu, D. Anguelov, D. Erhan, *et al.*, "Ssd: Single shot multibox detector", *Proc. of European Conference on Computer Vision*, pp.21-37, 2016.

[101] R. Girshick, J. Donahue, T. Darrell,*et al.*, "Region-based convolutional networks for accurate object detection and segmentation", *IEEE Transactions on Pattern Analysis and Machine Intelligence*, Vol.38, No.1, pp.142-158, 2016.

[102] S. Song and J. Xiao, "Sliding shapes for 3d object detection in depth images", *Proc. of European Conference on Computer Vision*, pp.634-651, 2014.

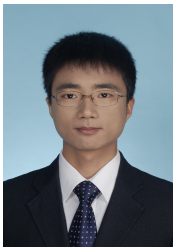
[103] J. Redmon and A. Farhadi, "YOLO9000: Better, faster, stronger", arXiv preprint arXiv:1612.08242, 2016.

LIU Yunhui received the B.Eng. degree from the Beijing Institute of Technology, the M.Eng. degree from Osaka University, and the Ph.D. degree from the University of Tokyo in 1992. After working at the Electrotechnical Laboratory of Japan as a Research Scientist, he joined The Chinese University of Hong Kong (CUHK) in 1995 and is currently Choh-Ming Li Professor of Mechanical and Automation Engineering and the Director of the CUHK T Stone Robotics Institute. He is also an adjunct professor at the State Key Lab of Robotics Technology and System, Harbin Institute of Technology, China. He has published more than 200 papers in refereed journals and refereed conference proceedings and was listed in the Highly Cited Authors (Engineering) by Thomson Reuters in 2013. His research interests include visual servoing, medical robotics, multifingered grasping, mobile robots, and machine intelligence. Dr. Liu has received numerous research awards from international journals and international conferences in robotics and automation and government agencies. He is the Editor-in-Chief of Robotics and Biomimetics and served as an Associate Editor of the IEEE TRANSACTION ON ROBOTICS AND AUTOMATION and General Chair of the 2006 IEEE/RSJ International Conference on Intelligent Robots and Systems. He is an IEEE Fellow. (Email: yhliu@mae.cuhk.edu.hk)

ZHENG Fan received the B.Eng. degree from Department of Mechanical Engineering, Zhejiang University, in 2014. He is currently pursuing his Ph.D. degree with Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong. His research interests include state estimation for robotics, visual SLAM, and visual inertial navigation. (Email: fzheng@link.cuhk.edu.hk)

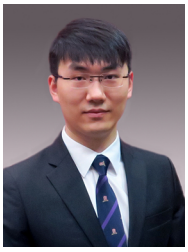
GUO Ruibin was born in Shanxi, China, in 1990. He received the B.S. and M.S. degrees in information and communication engineering from the National University of Defense Technology, Changsha, China, in 2012 and 2014, respectively. where he is currently pursuing the Ph.D. degree. His research interests include Computer Vision, 3D Reconstruction and SLAM. (Email: rbguo@nudt.edu.cn)

WANG Jiangliu was born in 1994. She received the B.E. degree in Automation from Nanjing University in 2015. She is currently pursuing her Ph.D. degree with Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong. Her research interests include human-robot interaction, action recognition and computer vision. (Email: jlwang@mae.cuhk.edu.hk)



NIE Qiang was born in 1990. He received the B.E. degree in Mechanical Engineering from Nanchang University in 2012 and the M.S. degree in same major from Shanghai Jiao Tong University in 2015. Now he is a Ph.D. candidate of Mechanical and Automation Engineering in The Chinese University of Hong Kong . His research interests include robot design and AI about recognition of human behaviors. (Email: qnie@mae.cuhk.edu.hk)

WANG Xin was born in 1992. He received the B.E. degree and M.Sc degree in Control Science and Engineering from Harbin Institute of Technology. He is a Ph.D. candidate of Mechanical and Automation Engineering in The Chinese University of Hong Kong. His research interests include objection detection and grasp pose detection . (Email: xwang2@mae.cuk.edu.hk)



WANG Zerui received the B.E. degree from School of Reliability and System Engineering, Beihang University, in 2013, and the Ph.D. degree from Department of Mechanical and Automation Engineering, The Chinese University of Hong Kong, in 2017. He is currently a research assistant professor of the Department of Mechanical and Automation Engineering and the CUHK T Stone Robotics Institute, The Chinese University of Hong Kong. His research interests are vision-based soft object manipulation, vision-based surgical tool tracking and control, and developments of robotic surgical components and systems. (Email: zrwang@mae.cuhk.edu.hk)