Review: An Investigation of Estimation Techniques for Development of Hand Rehabilitation Devices

Arash Basirat Tabrizi
Dept. of Mechanical and Industrial Engineering
abasirat@torontomu.ca

***Abstract*—Development of a hand rehabilitation device that is accurate, compact, portable, and cost-efficient is a challenging yet novel solution. In this preliminary review, we aim to investigate the role of estimation techniques in the manipulation of vision-based haptic/tactile sensors. Consequently, we will explore the design and implementation of such sensors. The integration of high-resolution and precise vision-based sensors within a hand rehabilitation device has the potential to assume an integral role in the realm of rehabilitative practices.**

***Index Terms*—State Estimation, Rehabilitation Devices, Motion Tracking, Pose Estimation, Factor Graph Optimization, Neural Networks, RGB Images, Tactile Sensing, Haptic Sensors.**

I. Introduction

Disabilities caused by experiencing a stroke and spinal cord injuries severely impact patients' quality of life and ability to perform activities of daily living (ADLs). Early rehabilitation management after a stroke or spinal injury is more effective and shortens hospitalization stays [10]. In practice, such high-intensity therapy is labor and resource intensive, and ultimately limited by the availability of the therapists.

Ideally, rehabilitation devices can provide patients with high-intensity training, carried out in their home environment with fewer resources required. However, despite the increasing number of such devices being developed in the recent years for upper-limb rehabilitation, the majority of these solutions have yet to be tested in clinical settings [11]. In addition, most of these devices lack a user friendly interface, are often too complex to operate, face durability challenges, have limited range of motion (ROM), and are costly. Hence, a novel hand rehabilitation device must be developed to overcome these challenges.

The design of a novel rehabilitation device is divided into three main modules: 1. Actuation 2. Sensing 3. Data Processing. The purpose of each module is summarized as follows:

1) **Actuation**: Actuation is the most challenging part of the design of a mechanical rehabilitation device for the hand. In practice, actuators provide the necessary force and displacement to simulate the natural movements of the hand. Actuation also allows the devices to apply controlled forces to each finger.
2) **Sensing**: Sensor technology can be used to measure forces, torques, and object poses. A fingertip-sized sensor that provides high-resolution measurements is suitable for rehabilitation applications.
3) **Data Processing**: Processing of the data collected by sensors is an essential step in the evaluation and testing of hand rehabilitation devices. The processing of this data involves the use of data-driven algorithms and deep-learning techniques to extract meaningful information about the patient's hand function. This information can then be used to track the rehabilitation training progress over time.

II. Problem Statement

We address the problem of the design and implementation of a portable non-wearable hand rehabilitation device that is cost-efficient and easy to operate. The device is expected to feature a 1-DOF linear actuator system inspired by the linear voice coil actuator (LVCA) design. The user should be able to operate all five fingers simultaneously using the proposed device. High-resolution vision-based haptic/tactile sensors will be an integral part of the device's sensor system. The preliminary contributions of this study are:

1) Design and implementation of a finger-sized LVCA.
2) Design and implementation of a fingertip-sized vision-based haptic/tactile sensor.
3) Empirical evaluation of the actuator and sensor systems using both simulation and real-world experiments.

While the design of the linear actuator system is in the early research phase, we aim to investigate the literature with respect to the design, implementation, and manipulation of vision-based haptic/tactile sensors. Section III provides a comprehensive review of the recent advancements in the field of vision-based sensors tailored to estimation techniques. For future reviews, we aim to introduce works that investigate the intersection of linear actuation and estimation algorithms.

III. Literature Review

The idea at the base of vision-based sensors is to measure contact forces as changes in images recorded by a camera, typically through the use of a deformable elastomer gel. In [1] the authors discussed the development of a vision-based optical tactile sensor, *GelSight*. The study introduced the design of the sensor's optical system, the algorithm for shape, force and slip measurement, and the hardware designs and fabrication. GelSight's ability to capture high-resolution images, that encoded measurements of shape and contact force, made it a popular fingertip-sized vision-based tactile sensor for robot grippers. For shape measurements, the proposed algorithm

TABLE I
A COMPARISON BETWEEN THE STATE-OF-THE-ART HAPTIC SENSORS

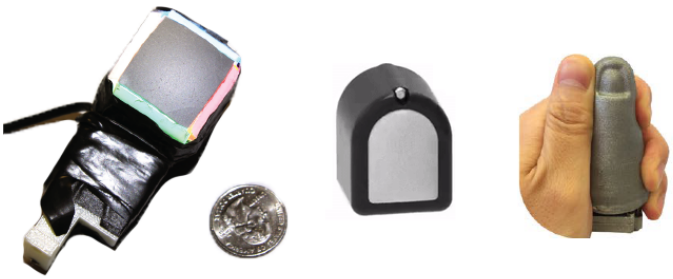
	Fingertip GelSight [1]	DIGIT [2]	Insight [4]
Size [mm]	35x60x35 (WxHxD)	20x27x18 (WxHxD)	20x70 (RxH)
Weight [g]	NA	20	NA
Sensing Field Area [mm^2]	252	304	4,800
Simulation Support	—	TACTO	—
Image Resolution	1920x1080	640x480	1640x1232
Image FPS	30	60	40
Cost components [$]	∼30	15 (for manufacturing of 1000 pieces)	∼100
No. of Layers	4	3	1
Data Processing	CNN	ResNet	ResNet
Output Format	Shape + Force	Shape	Force Map
Notes	Complex to manufacture	Difficult to extend to 3D	Costly and bulky

estimated the surface normals on each pixel of the output RGB image, and then integrated the surface normal to get the 3D shape of the surface. For force measurements, GelSight was initially calibrated using an *ATI Nano-17* force/torque sensor to find that the minimum perceivable force of the sensor, when contacting different types of objects, is less than 0.05 N. Since the sensor's surface was painted with small black markers, the pattern of the markers' motion field indicated the type of the force or torque, and the magnitude of the motion was roughly proportional to the force. The work calculated the motion of the markers from images using a low-pass Gaussian filter in combination with analytical image processing techniques for segmenting and locating the markers. Consequently, by producing force-deformation curves of the fingertip GelSight, the authors found that when an indenter —a small hard object— is pressed on the GelSight surface in normal direction, the force demonstrated a linear relationship with the indenting depth. Furthermore, when an indenter moved on the surface in the shear direction, the average marker displacement was proportional to overall shear force. On the other hand, the linear relationship between force and marker motion only existed when the contact geometry remained unchanged. To achieve a shape independent relationship between force and displacement the authors applied a *Convolutional Neural Network* (CNN) on the GelSight's high-dimensional, noisy, and highly nonlinear images. The CNN network allowed them to extract further useful features regarding contact forces from raw tactile images. With the integration of the CNN, the authors were able to show that GelSight's measurement of force can be robust regardless of the geometry of the contact objects. Furthermore, the model's output of forces and torques correlated to the ground truth measurements made by the force/torque sensor.

Sensing and reasoning about contact forces are crucial to accurately control interactions of a rehabilitation device with the user's fingers. As a step towards enabling better in-hand manipulations, [2] introduced a compact, lightweight, and high-resolution tactile sensor named *DIGIT*. While DIGIT inherited the rich and sensitive measurement characteristics of previous vision-based sensors, it was designed to have a smaller form factor and leveraged a streamlined manufacturing process that facilitated large-scale production of its hardware. DIGIT improved over existing GelSight sensors as seen by the comparison provided by Table I. Furthermore, while GelSight sensor lacked simulation support, making it difficult for deep-learning training, DIGIT was compatible with the open-source TACTO [3] simulator. In [2] the authors focused on presenting the design and manufacturing process of DIGIT, and in conjunction with their paper, they released the design files of the sensor to facilitate tactile sensing research. In practice, high-resolution vision-based tactile sensor like DIGIT come with a computational cost of interpreting the output tactile images with the aid of deep neural networks.

Sun *et al.* [4] presented a robust, vision-based, thumb-sized three-dimensional haptic sensor named *Insight*, which claimed to provide a directional force-distribution map over its entire conical sensing surface. The force information was deduced through the utilization of a deep neural network, which employed image mapping to discern the spatial arrangement of three-dimensional contact force involving both normal and shear components. Insight possessed an overall spatial resolution of 0.4 mm, a force magnitude accuracy of around 0.03 N and a force direction accuracy of around five degrees over a range of 0.03–2 N for numerous distinct contact episodes. While all competing sensors employed multiple functional layers mechanical designs, Insight claimed to be the only sensor with a single soft layer. Furthermore, Insight appeared to have the largest sensing surface in comparison with competitors. The authors claimed that GelSight and DIGIT sensors, [1] and [2] respectively, specialized for measuring contact area shape, and that a spatially extended map of three-dimensional contact forces over the sensing surface, which they called a force map, was rarely provided in any of the competing designs. Table I provides a detailed comparison of the three vision-based sensors discussed thus far. The study presented a direct contact estimation pipeline for inferring single contact position and force to help evaluate the sensor's performance with respect to accuracy. A real experimental setup allowed the researchers to perform statistical evaluation of the sensor's performance with the help of *histograms* and various force plots. Furthermore, a second pipeline for estimating the force distribution was introduced. The force distribution was needed to infer contact areas and multiple simultaneous contacts, and further evaluate the accuracy of the sensor. The study employed a deep-learning-driven method to estimate the force distribution directly from a raw image, namely an adapted

ResNet, which is favoured over GelSight's CNN architecture.



(a) (b) (c)

Fig. 1. Vision-based haptic/tactile sensors: (a) Fingertip GelSight [1], (b) DIGIT [2], (c) Insight [4].

While the sensors presented in [1], [2], and [4] granted us a road map for the design and implementation of high-resolution vision-based sensors, we must further investigate how these sensors can be manipulated. We shift our focus to the work presented in [5], where the authors addressed the problem of estimating object poses from touch during planar pushing. The vision-based tactile sensor used in the study, DIGIT, only provided local image measurements at the point of contact. However, a single such measurement, encapsulated limited information and multiple measurements were needed to infer latent object state. Hence, they solved for this inference problem by using a *Factor Graph* (FG) interface. In the proposed two stage approach, they first learned local tactile observation models supervised with ground truth data, and then integrated these models along with physics and geometric factors within a factor graph optimizer. Under Gaussian noise model assumptions *maximum a posteriori* (MAP) inference over a factor graph was equivalent to solving a *nonlinear least-squares* problem:

$$\hat{x}_{1:T} = \underset{x_{1:T}}{\operatorname{argmin}} \sum_{t=1}^{T} \Big\{ \|F_{qs}(o_{t-1}, o_t, e_{t-1}, e_t)\|^2_{\Sigma_{qs}} + \|F_{geo}(o_t, e_t)\|^2_{\Sigma_{geo}} + \|F_{tac}(o_{t-k}, o_t, e_{t-k}, e_t)\|^2_{\Sigma_{tac}} + \|F_{eff}(e_t)\|^2_{\Sigma_{eff}} \Big\} \tag{1}$$

Equation (1) was the optimization objective that they solved for every time step. For the planar pushing setup, states in the graph (variable nodes) were the planar object and end-effector (sensor) poses, o_t and e_t respectively at every time step t. Factors in the graph incorporated tactile observations $F_{tac}(\cdot)$, quasi-static pushing dynamics $F_{qs}(\cdot)$, geometric constraints $F_{geo}(\cdot)$, and priors on end-effector poses $F_{eff}(\cdot)$. The goal of learning a tactile observation model using deep-learning networks was to derive the tactile factor cost term in (1). The work's learnt tactile observation model consisted a transform prediction network trained using a *mean-squared loss* (MSE) against ground truth data and implemented using PyTorch. The study demonstrated that the proposed method was able to reliably track object poses for over 150 real-world planar pushing trials for 3 distinct object shapes using tactile measurements alone.

Sodhi *et al.* [6] addressed the problem of tracking small 3D object poses from touch during in-hand manipulations. While previous works have looked at the object tracking problem primarily in the context of planar pushing, such as in [5], or have relied on a-priori information about the object being localized, the work presented in [6], *PatchGraph*, created a 3D local patch map of the object online within a factor graph for inferring the latent 3D object poses. The states in the graph (variable nodes) were the same as in [5], however, the optimization objective changed to the following:

$$\hat{x}_{1:T} = \underset{x_{1:T}}{\operatorname{argmin}} \sum_{t=1}^{T} \Big\{ \|f_{im2im}(o_{t-1}, o_t, e_{t-1}, e_t)\|^2_{\Sigma_{im2im}} + \|f_{im2pc}(o_t, e_t)\|^2_{\Sigma_{im2pc}} + \|f_{vel}(o_{t-2}, o_{t-1}, o_t)\|^2_{\Sigma_{vel}} + \|f_{eff}(e_t)\|^2_{\Sigma_{eff}} + \|f_{vis}(o_t)\|^2_{\Sigma_{vis}} \Big\} \tag{2}$$

Factors in the graph included image-to-image factors $f_{im2im}(\cdot)$, image-to-patch factors $f_{im2pc}(\cdot)$, velocity smoothness priors $f_{vel}(\cdot)$, end-effector pose priors $f_{eff}(\cdot)$, and vision priors for re-localization at the beginning of a contact episode $f_{vis}(\cdot)$. In the first stage of PatchGraph, surface normals were predicted from the tactile RGB images generated by DIGIT using a pix2pix (U-net) image translation network. They trained the model on both simulated and real data. For simulated trials and dataset collection they imported DIGIT in the TACTO environment. In the second stage, they integrated the learned surface normals inside a factor graph to build a local patch map and then use that to find the object poses. The authors evaluated the performance of PatchGraph on simulated trials for 4 distinct objects, and on real trials for 2 objects. Although PatchGraph was able to reliably track rotations of most objects, it struggled to track the simulated toy brick object. Additionally, tracking rotations for a sphere object demonstrated high errors, which was expected since rotations of a sphere are unobservable from tactile images due to its unique geometric characteristics.

[9] presented a learnable *Bayes filters* models with the objective to estimate the position of a robotic gripper with respect to the global map of the environment using tactile feedback. In their localization problem the robot was assumed to have a non-uniform prior as the initial *belief*. In addition, the pose of the arm was assumed to be known with respect to the world coordinates. Furthermore, the work formulated the localization problem as a recursive Bayes filtering problem and learned the filter models from data. Unlike localization with visual feedback, tactile localization is not affected by the occlusions between the camera and the end-effector and can be useful in situations where there is no precise forward kinematic model of a robotic arm. While tactile feedback has a limited use in state estimation due to its noisy and high-frequency nature, it can still provide rich information about the contact interactions, which can help determine object properties and locations. Bayes filters are well suited for state estimation problems with uncertain observations and transition functions. The Bayes filters work by maintaining a belief of the state (s_t) and use an observation and a motion model. The observation and motion models update the belief using the measurement

received from the environment (o_t) and the actions taken by the robot (a_t). In the study, the authors designed their unknown observation and motion models using neural network layers and implemented Bayes filtering as tensor operations. The proposed observation model was based on the U-net architecture and used depth images (captured with a Structure depth sensor) of the environment as visual maps and tactile feedback as the observation to generate the likelihood probabilities. Additionally, the motion model transitioned the belief to the next time step. In short, the core of the presented touch localization framework was a discrete Bayes filter (*Histogram filter*) with learned models. As mentioned earlier, the work decided to represent the observation and motion models of the Bayes filter as neural networks and learned them using *gradient descent*. They authors let $f_O(\cdot)$ be a neural network that took the environment image, tactile observation and the action as input and generated the likelihood probabilities of the current observation. Additionally, they let $f_M(\cdot)$ be a neural network that took the previous belief and the action as input and predicted the belief at the next time step. The state space was defined as the projected pixel coordinates of the gripper in the environment image: $s_t = (p_x, p_y) \in \mathcal{Z}^{H \times W}$ where H was the height and W was the width of the image, I. Hence, the belief was encoded as a matrix and formulated as:

$$bel(s_t) = \eta f_O(o_t, a_t, I) \odot f_M(bel(s_t - 1), a_t) \qquad (3)$$

In (3), $\odot$ is element-wise multiplication and η is the normalization factor. Overall, the work demonstrated that in addition to successful localization with unseen object configurations and good generalization over different environment configurations, the proposed approach could also localize with novel objects. The main drawback of the method was the assumption that the gripper did not move objects. To overcome this problem, the authors proposed to extend the localization formulation in future work to track the objects as well.

REFERENCES

[1] W. Yuan, S. Dong and E. H. Adelson, "GelSight: High-resolution robot tactile sensors for estimating geometry and force", ***Sensors***, vol. 17, no. 12, pp. 2762, 2017.

[2] M. Lambeta, P.-W. Chou, S. Tian, B. Yang, B. Maloon, V. R. Most, D. Stroud, R. Santos, A. Byagowi, G. Kammerer et al., "DIGIT: A novel design for a low-cost compact high-resolution tactile sensor with application to in-hand manipulation", ***IEEE Robotics and Automation Letters***, vol. 5, no. 3, pp. 3838-3845, 2020.

[3] S. Wang, M. Lambeta, P. -W. Chou and R. Calandra, "TACTO: A Fast, Flexible, and Open-Source Simulator for High-Resolution Vision-Based Tactile Sensors", ***IEEE Robotics and Automation Letters***, vol. 7, no. 2, pp. 3930-3937, April 2022.

[4] H. Sun, K.J. Kuchenbecker and G. Martius, "A soft thumb-sized vision-based sensor with accurate all-round force perception". ***Nat Mach Intell 4***, pp. 135–145, 2022.

[5] P. Sodhi, M. Kaess, M. Mukadam and S. Anderson, "Learning tactile models for factor graph-based estimation", ***IEEE Intl. Conf. on Robotics and Automation (ICRA)***, 2021.

[6] P. Sodhi, M. Kaess, M. Mukadanr and S. Anderson, "PatchGraph: In-hand tactile tracking with learned surface normals", ***IEEE Intl. Conf. on Robotics and Automation (ICRA)***, 2022.

[7] P. Griffa, C. Sferrazza and R. D'Andrea, "Leveraging distributed contact force measurements for slip detection: a physics-based approach enabled by a data-driven tactile sensor", ***IEEE Intl. Conf. on Robotics and Automation (ICRA)***, pp. 4826-4832, 2022.

[8] Perez, N. "Theory of Elasticity", 1–52, Springer, 2017.

[9] T. Keleştemur, C. Keil, J. P. Whitney, R. Platt and T. Padır, "Learning Bayes Filter Models for Tactile Localization", ***IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)***, pp. 9253-9258, 2020.

[10] Y. Liu, and C. Yeow, "A Portable Soft Hand Exerciser With Variable Elastic Resistance for Rehabilitation and Strengthening of Finger, Wrist, and Hand", ***ASME. J. Med. Devices***. September 2015; 9(3): 030920.

[11] R. Rätz, F. Conti, RM. Müri and L. Marchal-Crespo, "A Novel Clinical-Driven Design for Robotic Hand Rehabilitation: Combining Sensory Training, Effortless Setup, and Large Range of Motion in a Palmar Device", ***Front. Neurorobot***. 2021.