High-resolution Tactile Sensing for Robotic Perception

by

Wenzhen Yuan

Submitted to the Department of Mechanical Engineering
in partial fulfillment of the requirements for the degree of

Doctor of Philosophy in Mechanical Engineering

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

September 2018

Author ..... Signature redacted ....
Department of Mechanical Engineering
June 18, 2018

Certified by ..... Signature redacted .....
Edward H. Adelson
John and Dorothy Wilson Professor of Vision Science
Thesis Supervisor

Certified by ..... Signature redacted .....
Mandayam A. Srinivasan
Senior Research Scientist
Thesis Supervisor

Accepted by ..... Signature redacted
Rohan Abeyaratne
Quentin Berg Professor of Mechanics
Chairman, Department Committee on Graduate Theses

High-resolution Tactile Sensing for Robotic Perception

by

Wenzhen Yuan

Submitted to the Department of Mechanical Engineering
on June 18, 2018, in partial fulfillment of the
requirements for the degree of
Doctor of Philosophy in Mechanical Engineering

Abstract

Why is it so difficult for the present-day robots to act intelligently in the real-world environment? A major challenge lies in the lack of adequate tactile sensing technologies. Robots need tactile sensing to understand the physical environment, and detect the contact states during manipulation. A recently developed high-resolution tactile sensor, GelSight, which measures detailed information about the geometry and traction field on the contact surface, shows substantial potential for extending the application of tactile sensing in robotics. The major questions are: (1) What physical information is available from the high-resolution sensor? (2) How can the robot interpret and use this information?

This thesis aims at addressing the two questions above. On the one hand, the tactile feedback helps robots to interact better with the environment, i.e., perform better exploration and manipulation. I investigate various techniques for detecting incipient slip and full slip during contact with objects, which helps a robot to grasp them securely. On the other hand, tactile sensing also helps a robot to better understand the physical environment. That can be reflected in estimating the material properties of the surrounding objects. I will present my work on using tactile sensing to estimate the hardness of arbitrary objects, and making a robot autonomously explore the comprehensive properties of common clothing. I also show our work on the unsupervised exploration of latent properties of fabrics through cross-modal learning with vision and touch.

Thesis Supervisor: Edward H. Adelson
Title: John and Dorothy Wilson Professor of Vision Science

Thesis Supervisor: Mandayam A. Srinivasan
Title: Senior Research Scientist

Acknowledgments

I would like to thank my two advisors, Prof. Edward Adelson and Dr. Mandayam Srinivasan, for guiding me through this fascinating field of research, and offering the big supports to my career, and Prof. Alberto Rodriguez, and Prof. Katherine Kuchenbecker for the long-term support and advising during the thesis preparing. I would like to thank my collaborators on those thesis work: Siyuan Dong, Shaoxiong Wang, Rui Li, Yuchen Mo, Chenzhuo Zhu, and Andrew Owens. The collaboration with other researchers: Roberto Calandra, Sergey Levine, Shan Luo, Robert Platt, Abhijit Biswas, Mohan Thanikachalam and Jiajun Wu, also largely helped me to better understand the field of robot tactile sensing. I also wish to thank Andrea Censi, Srikumar Ramlingam, Ruth Rosenholtz, Micah Kimo Johnson, Bei Xiao, Xiaodan "Stella" Jia, Russ Tedrake, Gregory Izatt, Elliott Donlon, Ziwei Wang, Shaiyan Keshvari, Erik Hemberg, Ben Wolfe, Phillip Isola, and Tianfan Xue for the help and support during the research. And thank my family and friends, for the support and the good time we spend together.

MIT is a great school. It is fun to attend PhD program here.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

In 2017, the AI program AlphaGo beat the best human player in the complicated board game Go, which reminded the public of the old sci-fi scene: Could robots act as humans to accomplish tactful and delicate tasks, like doing the house-keeping and washing the clothes? Unfortunately, it turns out that it is more difficult for a robot to pick up a dish from a messy dish pile than play Go. The physical world has much more variables and uncertainties, and is hard for robots to deal with. A major barrier lies in the robot perception system: how to understand the physical world, and how to interact with it accordingly. The fast development of computer vision in recent years has enabled robots to well understand visual information, but the perception from other sensory modalities, especially tactile sensing, which is crucial for physical-based perception, has been largely underdeveloped.

In the past decades, researchers have developed many different tactile sensors for robots [10, 70, 9, 29], and the core part of those tactile sensors is to detect the contact and contact force, or force distribution over the fingertip area. For example, a successfully commercialized sensor is the tactile sensor array from Pressure Profile Systems, which measures the normal pressure distribution over the robot fingertip, with a spatial resolution of 5mm. The sensor has been applied to multiple commercialized robots, including PR2, Barrett hands, and it successfully assisted common robotic tasks, such as contact detection and gripping force control. With the force measurement from the fingertip tactile sensors, a robot is much less likely to break

delicate objects. The contact detection and localization also refine the robots' performance in grasping and in-hand manipulation.

Nevertheless, compared to dexterous sensing system of humans, the tactile sensing technologies for robots are limited. Humans use tactile sensing for a wide range of tasks. Humans get abundant information from tactile sensing[53, 60, 34, 62], such as the objects' shape[54], texture[32], material and physical properties including mass[35], compliance[55, 59], roughness, friction and thermal conductivity. Tactile sensing is an important part of humans' closed-loop manipulation as well. With touch, we can know whether the cup in the hand is going to slip, and thus adjust the gripping force accordingly; we can know whether a USB connector is going to be correctly plugged into the socket, because we get the feedback of the impedance force.

The major challenges for robotic tactile sensing technologies come from both hardware and software: how to design sensor devices that can obtain adequate tactile information, and how to interpret the raw signal into the relevant information for understanding the world. The information offered by the traditional tactile sensors – force and pressure distribution over a small area, although helpful for perceiving contact location and magnitude, is very insufficient to help robots to well understand the physical world or interact with it.

In the recent years, the emerging of the high-resolution tactile sensing technology – the vision-based soft sensor GelSight, provides new possibilities for extending the boundary of what robot tactile sensing is capable of. The detailed shape information about the contact surface, the dynamic change of the soft sensory medium, are both informative of the contact surface. A robot could easily learn the local texture or shape details about the object being contacted, thus easily recognize the possible material or item. But more information could be contained in the input. The new questions are, what can the new sensor help with? How to make it work?

My research goal is to build an intelligent robotic tactile perception system for the sake of exploration and manipulation. This thesis work focuses on the possibility of applying the new high-resolution tactile sensor. Regarding the application of tactile

sensing, exploration means understanding the environments and objects, and manipulation means interacting with the physical world, where sensory feedback is very important. With the tactile perception, a robot can know whether a chair is rigid or is covered by comfortable and soft cushions, it can know whether an avocado is ripe enough to eat by estimating its hardness, and it can know whether the ground surface is slippery that does not suit walking. When interacting with the environment, the tactile perception will enable a robot to secure a grasp by detecting potential slip, it will prevent a robot from crushing a rotten tomato, and it will enable a robot to pick an M&M from a bag of snack mix. Those tasks can only be accomplished by the perception through physical contact with the objects.

1.1 Tactile sensors

Some thorough reviews of the existing tactile sensors in the past decades are given by [10, 70, 9, 29]. The sensors use different sensing principles, such as resistance, capacitance, piezoelectricity, optic component, or magnetics. The major focus of the sensors have been measuring the pressure distribution, or contact force and location, over a certain area. For the tactile sensors applied on robots, most of the sensors are designed for the fingertips or gripper end-effectors (an example is [24]), which measures the force or contact during grasping; some other sensors designed for body (an example is [43]), which detect the contact over a much larger area and are commonly used for contact or collision detection during the robot motion.

Among the tactile sensors, the optical sensors based on vision stand out because they are of easier wiring and fabrication processes, and can mostly provide a relatively high spatial precision in locating contact area. The vision-based tactile sensors typically use a deformable body, either a piece of rubber or a fluid balloon, as the sensing medium, and apply a camera to capture the deformation of the medium. In most cases, the deformation is indirectly measured from other visual cues, such as the deformation of the pre-printed patterns on the medium.

In the 1980s, Schneiter and Sheridan [50] and Begej [2] used optical fiber to capture

the light reflection on silicone surface, and used cameras to capture the change of optic signal from the optical fibers. The deformation of the medium would cause a change in the light reflection. A successful example is [61], which has already been commercialized. The sensor used hollow hemispherical rubber dome as the contact medium, the dome has a reflective inside surface. The sensor uses three receivers in the bottom to measure the reflective light from the deformed dome, thus estimating the 3-axis contact force. But the spatial measurement is not available with the sensor.

Another popular design for the vision-based sensors is to print marker patterns on or in the sensing medium, and track the motion of the markers from the embedded camera. Some examples of the designs include [15, 27, 6, 22, 69]. Ferrier and Brockett [15] developed an analytical model to calculate the value and position of the point-contact force on a round shaped fluid finger. But in most cases, an analytical model is hard to build and is restrained by many contact requirements. The non-linearity of the elastomer material and the complicated contact condition with multiple contact points or contact surface greatly increases the difficulty. Kamiyama et al. [27] developed a large flat sensor with two layers of markers in the cover rubber, and they used a simplified mechanical model to calculate normal and shear force during the point contact condition. In their later work[49], they scaled down the sensor design to a robot fingertip. They used experimental methods to calibrate the contact force. Chorley et al. [6] designed another hemispherical sensor with markers close to the surface, and Cramphorn et al. [8] introduced an improved design by adding the core-shaped fingerprint on the sensor surface. The sensor is 3D printed, which makes it much easier to reproduce. TacTip could sensitively discriminate the contact, and can be used to extract the approximate edge of the contact surface. The force was not measured, but the edge detection successfully helped a robot to localize contact and follow the contours [36]. Ito et al. [22] designed a sensor that had a hemispherical shape filled with translucent water, and the marker patterns were printed on the surface. The sensor responded to contact with both marker motion and the change of filling water's reflection, and they built an analytical model to estimate the shapes when the contact shape is simple. They also showed in [23] that by counting the

markers that were 'stuck' on the surface or 'lagged behind', the sensor could estimate partial slip stage. However, their sensor is too large in volume to be properly applied to robots. Yamaguchi and Atkeson [69] designed a sensor that is similar in tracking the motion of the markers on the sensor surface, but with a clear body, so that the camera can also see through the fingertip and obtain images of the outside environment. Their aim is to apply the close-range vision in the fingertip area to aid the manipulation tasks.

The high-spatial-resolution measurement is still largely under-exploited. Maheshwari and Saraf [42] offered one possible solution: they proposed an electro-optical tactile sensor that used a surface film made by metal and semiconducting nanoparticles that converted pressure into optical signals. The sensor is able to sense the pressure distribution at a resolution similar to human fingers. Here are some major challenges for making the desired robotic tactile sensors:

- Measurement of shear force. Only some of the tactile sensors are able to measure shear force as well as normal force, while the shear force is very important in multiple robotic tasks, such as estimating object states in grasping, estimating surface friction of the objects through touch.

- Detecting contact area. Most of the sensors focus on the situation of point contact, which means they are designed to measure the location and magnitude of the contact force. However, in many tactile tasks, the contact is one or multiple areas, instead of a single point. Sensors that can detect more tactile information based on the contact area are desired.

- Hardware optimization. If the sensors can be used on the robots, it must be small in sizes, easy on wiring, and offer real-time feedback. Some sensors have delicate structures and good signals, but the design is either too bulky or too complicated to be mounted on robot hands.

- Fabrication challenge. A major work in the research of the tactile sensing is to develop a method to fabricate the sensors, which is usually non-trivial. Unfortunately, most of the fabrication methods are not well shared – it is hard

for another lab to duplicate the complicated fabrication procedures. TacTip offers a good example of where the 3D printing methods are open source. In other cases, devices have been commercialized and are therefore available via purchase. Two good examples are the Optoforce sensor [61] and the BioTac sensor [63]: the researchers founded startups to improve the product design and produce the sensor commercially, so that other robotic labs have direct access to the sensors.

- Integrating into the robotic system. The designing of the tactile sensor should be well collaborating with 'What kind of sensor is needed by the robots? How could robots use the sensors?', while traditionally the sensor designing community and the robot manipulation community are not closely connected. Applying the sensors on the robots is equally important to designing the sensors, and they two should evolve alongside, instead of independently.

1.2 Thesis Contribution

This thesis explores how tactile sensing, especially high-resolution tactile sensing, could help robots to better understand and interact with the physical world. I address the challenges from three aspects: hardware, algorithm, and integration with other robotic components. For the hardware, I help my labmates to improve the sensor design, to obtain more information about the contact. Especially, our focus includes simplifying the fabrication process, and making the manufacturing procedure open-source, so that other labs could reproduce the sensor as well. For the algorithm, I take advantage of the fast-developing AI and machine learning technologies, and apply the state-of-the-art neural network methods on the tactile signals to learn the embedded information from the high-dimensional input. In addition, I believe a good perception framework is a combination of different robotic components. I try to integrate the tactile sensing with robot motion and vision. The motion performs the contact actively, and the vision provides supplementary information from the perception, that it is much better to capture the global information while touch is

more about local information. The integration will enable robots to learn more.

The thesis structured as following: Chapter 2 introduces the basic principle of the GelSight sensor and the design of the fingertip GelSight, which is used for the robotic tasks in this thesis. It also introduces the principle and capability of slip measurement with GelSight, and show in the robot experiment about how the measurement could help robots to perform stable grasp. Chapter 3 introduces how to use GelSight to estimate the objects' hardness, which is considered one important material property that is related to the understanding of the objects. In addition to demonstrating the working principle, I introduce and compare two algorithms for the measurement: one is using statistical models to directly describe the sensor information, and the other one is using deep learning that is trained with a large dataset of varied contact cases. Chapter 4 introduces the research on making a robot to actively conduct touch on common objects – using clothing as an example, and understand a comprehensive set of material properties through touch. The chapter focuses on how to integrate the tactile sensing in the entire robot system, in order to accomplish some semi-real-world challenges. Chapter 5 tries to explore a possible way to describe the implicit but comprehensive set of property vector to describe the materials, using fabrics as an example. The research uses joint neural-networks and co-training with vision modality, so as to learn a latent vector about the intrinsic properties of the fabrics, which is consistent regardless of the observing bias and perception modality.

Chapter 2

GelSight sensor for robots

For this thesis' work, we apply and improve a vision-based tactile sensor called GelSight, which measures high-resolution geometry, and which can be used to infer local force and shear. In this chapter, I introduce the working principle and designs for the robot fingertip GelSight sensors, and the experiment of force measurement and slip measurement with GelSight. The content in this chapter is published in [69, 70].

2.1 GelSight principle and design

The GelSight sensor uses a deformable elastomer piece as the medium of contact and an embedded camera to capture the deformation of the elastomer surface. The high-resolution 3D geometry of the contact surface can be reconstructed from the camera images. When the sensor surface is painted with small black markers, the motion of the markers provides information about both normal force and shear force.

Figure 2-1(a) and (b) shows an example of an Oreo cookie being pressed against the sensing elastomer, while the reflective membrane takes on the shape of the Oreo's surface. Given that the reflective properties and illumination condition are known, we can reconstruct the depth map of the surface using photometric stereo algorithm [64]. The example is shown in Figure 2-1(c). The spatial resolution of the sensor, when optimized for resolution, can reach 1-2 microns. In the case of compact GelSight devices designed for robot fingers, the spatial resolution is typically in the range of

Figure 2-1: Working principles of the GelSight sensor: when a cookie is pressed against the coated elastomer, the skin distorted, and the cookie's shape can be measured using photometric stereo algorithm. [25]

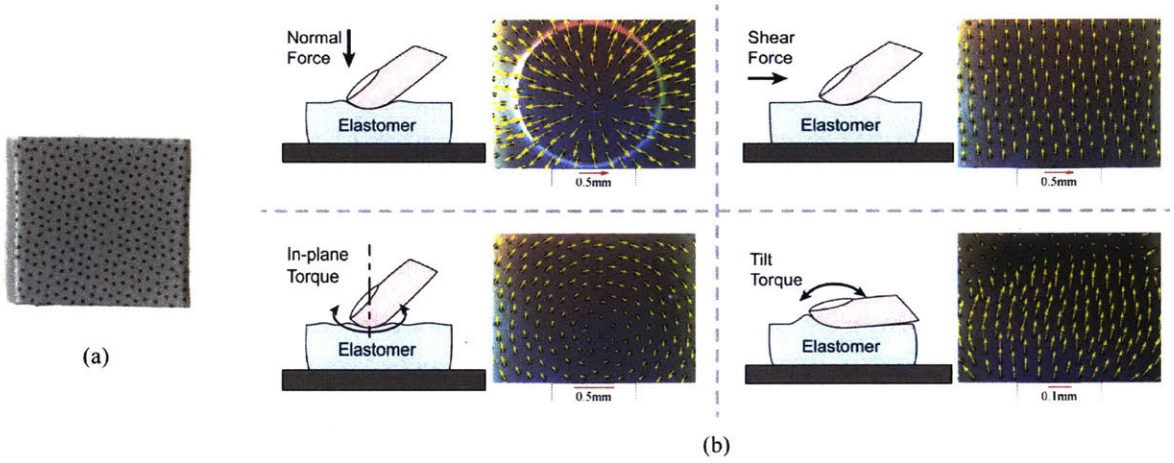
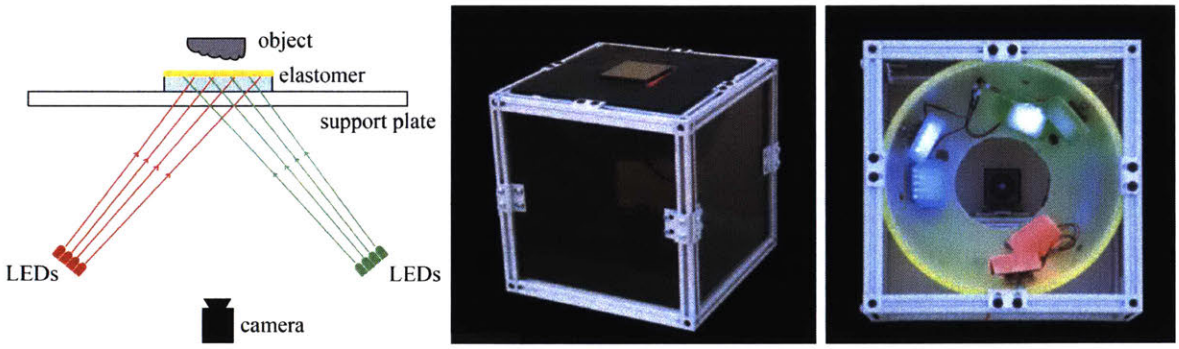


Figure 2-2: The markers on the GelSight surface and their motion. The markers are arranged at an average interval of 1.1mm. They make diverse motion field patterns under different kinds of forces or torques. The magnitude of the markers' motion is proportional to the force/torque value.

20-30 microns.

The vision-based design of the sensor also makes the hardware accessible and the installation much simpler, and the software for processing the raw data easier to develop by using the algorithms in computer vision. With the help of GelSight, a robot can easily capture the detailed shape and texture of the object being touched, which makes the touch-based object or material recognition much easier.

The first GelSight prototype was developed in 2009 by Johnson and Adelson [25]. The picture of the sensor and design is shown in Figure 2-3. Its high-resolution capabilities were further demonstrated in [26]. Unlike other optically based approaches, GelSight works independently of the optical properties of the surface being touched.



(a) (b) (c)

Figure 2-3: Basic schematic of the Gelsight and the desktop GelSight design [25]. (b)(c) show the overview and LED / camera arrangement of the sensor.

The ability to capture material-independent microgeometry is valuable in manufacturing and inspection, and GelSight technology has been commercialized by a company (GelSight Inc., Waltham, MA).

Design of the Fingertip GelSight sensors

For applying the tactile sensor on robot hands, it is necessary that the sensor be compact enough to be mounted on an end effector. At the same time, there is no need for micron-scale resolution. We have developed two versions of the Fingertip GelSight sensors that are compact, yet have resolution far exceeding that of human skin. The sensors are particularly designed for two kinds of robot parallel grippers. The work in this thesis applies those two sensors.

The first Fingertip GelSight sensor is built by Li et al.[37], and is designed for a Baxter Robot gripper (Figure 2-4(a)(e)). The sensor is close to a cubic shape, with the compact size of 35mm $\times$ 35mm $\times$ 60mm. The design uses perpendicular acrylic plate sets to guide the lights from the top to the sensing elastomer in the bottom, as shown in Figure 2-4(b). The sensing elastomer has a semi-specular coating to reveal the details and small fluctuations on the object's surface. After the internal reflection in the acrylic plate, the lights will be at the angles close to the parallel direction of the elastomer surface. The LEDs on the four sides of the sensors are in the form of a line array (1$\times$8), and are of four colors: red (R), green (G), blue (B) and white (RGB). The hue and saturation of each pixel indicate the direction (yaw angle) of the

surface normal, since the light sources from different directions are different in color, and the intensity corresponds to the magnitude (pitch of surface normal).

This sensor has 2 major deficiencies: firstly, although the sensor is very sensitive to small curvatures, the measurement of surface normal is not precise, because of the non-parallel illumination and semi-specular coating of the gel. The size reduction decreases the illumination quality. Secondly, the fabrication of this sensor is over-complicated. Massive accurate manual work is required, so that the product is hard to be standardized, and the fabrication is time-consuming.

In 2017, Dong et al.[13] proposed another version of the fingertip GelSight sensor (Figure 2-5(a)), with largely improved precision of shape measurement and simplified fabrication process. The new sensor is of approximately the same size and spatial resolution with the previous fingertip sensor, but is in a hexagonal shape and has a new illumination system using LEDs of three colors (RGB). The new LEDs (Osram Opto Semiconductor Standard LEDs - SMD) have small collimating lenses, and the emitted light are within a viewing angle of 30°. They are tightly arranged into 2×4 arrays with a customized circuit board, and are tilted at the angle of 71° to the sensing elastomer surface from the sides. The LED and elastomer are supported by a semitransparent tray, which homogenizes the LED light while allowing a high transmission rate. The sensor coating is matte, and the matte coating, as well as the illumination system, favor a more accurate surface normal measurement. Most of this sensor's parts are 3D printed with a Formlab 2 printer, and the clear acrylics are cut by a laser cutter. The parts are shown in Figure 2-5(c). The fabrication of the new sensor is highly standardized, and manual labor is significantly reduced.

In practice, the GelSight sensor measures the geometry by building a lookup table that matches the color of each pixel to the surface normal. The lookup table is calculated from a calibration process: we press a small sphere on the sensor surface, so that the surface normal of the geometry is known at every pixel, and we build the table that matches the surface normal to the color. Figure 2-6(a) gives two examples of the GelSight images of calibration and the matched surface normal comparing to the groundtruth. Figure 2-6(b) shows some examples of the reconstructed geometry

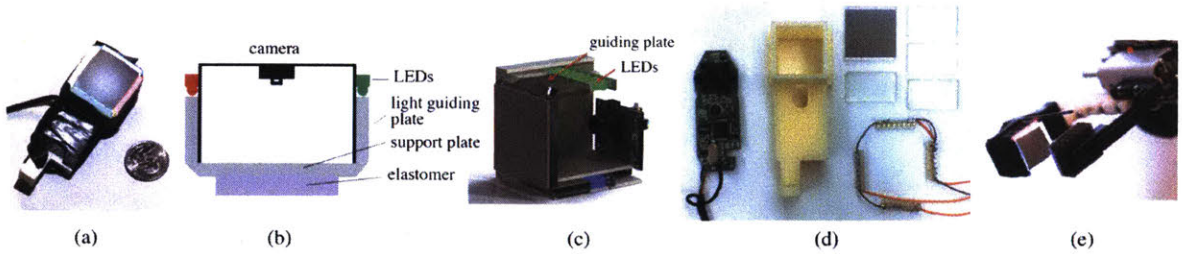


Figure 2-4: The Fingertip GelSight sensor introduced in [37, 36]: the illumination is of four colors (red, green, blue and white), and the entire sensing elastomer through the guiding plates are made of clear acrylic boards.

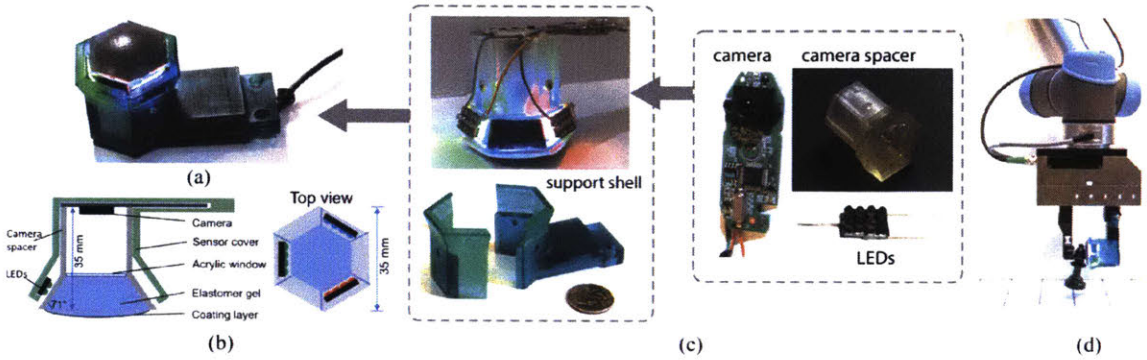


Figure 2-5: The new GelSight fingertip sensor introduced in [13]. The sensor has three LED arrays of different color to illuminate the elastomer surface from a tilted angle of 71°.

of common objects. The ground truth of the geometry is hard to get, but the reconstructed 3D structures capture both the overall shape and local textures of the objects.

2.2 Force and shear measurement with GelSight

The markers' motion well represents the contact force in two ways: the pattern of the motion field indicates the type of the force or torque, and the magnitude of the motion is roughly proportional to the force [68]. Examples of the different motion field patterns are shown in Figure 2-2: under the normal force, the markers spread outwards from the contact center; under the shear force, the markers all move in the shear direction; under the in-plane torque, which means the torque axis is perpendicular to the surface, the markers move in a rotational pattern. When there is a

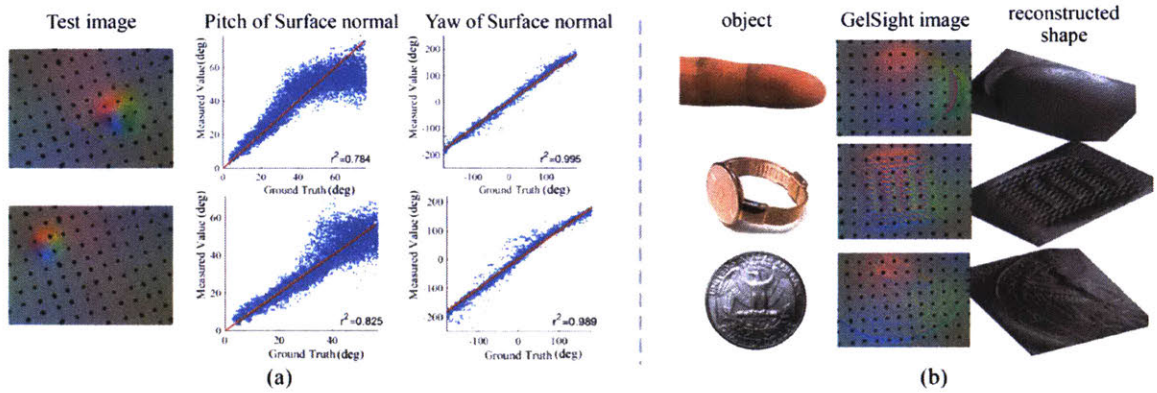


Figure 2-6: (a)Calibration images and results of surface normal matching with the new fingertip GelSight sensor. (b) Reconstructed depth maps.

combination of multiple force types, the motion field can be approximated as a linear combination of the individual forces.

We use the ATI-Nano 17 force/torque sensor for measuring the ground truth of the contact force and torque, and contact the GelSight sensor with different indenters. When using a flat indenter on the fingertip GelSight sensor, that the contact geometry does not change along the contact force, the force-deformation plot is shown in Figure 2-7. Figure 2-7(a) shows the normal force change in both the loading and unloading period, and the gap between the two curves is caused by the elastomer's viscoelasticity. Figure 2-7(b) shows that when the shear load is small, the force is linear to the loading displacement; when the load increases, partial slip or slip occurs, which stops the shear force from growing. Figure 2-7(c) is from the same shear experiment, and it shows that the average motion magnitude of the markers within the contact area is proportional to the shear force, regardless of whether slip or partial slip occurs. In fact, the linear relationship remains even before force reaches equilibrium.

Both the experiment and simulation results prove that, for the thin and flat elastomer piece on the fingertip GelSight sensor, when the geometry of the contact surface remains the same, the displacement of the markers on the surface, is to the linear relationship of the external force or torque. It is also a linear combination of the displacement field under each kind of the load. In other words, the overall displacement

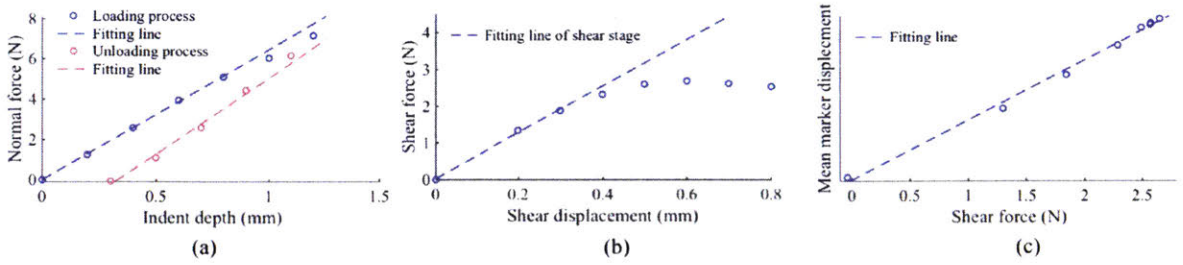


Figure 2-7: Force-displacement curves of the fingertip GelSight elastomer when a flat indenter is pressed against it. (a): The normal force in the loading and unloading period. (b): The increasing shear force under the increasing displacement of the indenter. (c): The force-displacement curve under shear loads.

field Φ can be approximated as

$$\mathbf{\Phi} = \begin{bmatrix} \mathbf{E}_N & \mathbf{E}_x & \mathbf{E}_y & \mathbf{E}_T \end{bmatrix} \begin{bmatrix} F_N \\ F_x \\ F_y \\ T \end{bmatrix}, \tag{2.1}$$

Where $\mathbf{E}_N, \mathbf{E}_x, \mathbf{E}_y, \mathbf{E}_T$ denote the displacement field of the markers under unit force in normal or shear direction, or unit torque; F_N, F_x, F_y, T denote the scaler of force and torque on each axis. Note that for different contact geometries, $\mathbf{E}_N, \mathbf{E}_x, \mathbf{E}_y, \mathbf{E}_T$ are different.

Estimating force using neural network

Although there is a simple linear relationship between the displacement field and the load, it is hard to decompose the motion field into the component fields, especially when considering the shape of the objects will have an unknown influence on the field.

A straightforward way to estimate the force and torque, regardless of the object geometry, is to use the deep neural network. In recent years, Convolutional Neural Network (CNN) [33, 31] has been widely applied in computer vision, that it worked well in modeling the complicated spatial relationships between the pixels in the images. Information on both the geometry and marker motion is contained in the pixel information, thus can be modeled by CNN.

In this experiment, we train the force measurement neural network with a dataset

of GelSight contacting objects with basic shapes, including spheres, cylinders, and flat planes, and then test the network's performance on the cases of contacting other similar but unseen objects The forces and torque we try to measure is the normal force, shear force and direction, and the in-plane torque (the torque along the Z axis in Figure 2-8(a)). The experimental setup is shown in Figure 2-8(a): a fingertip GelSight sensor is mounted on an ATI Nano-17 force/torque sensor on a fixed table, and we manually push or twist different objects against the GelSight sensor with different force amounts and directions. So that the force and in-plane torque on the GelSight surface is equal to the load on the ATI sensor, and we use the measurement from the ATI sensor as the ground truth. The objects include 6 balls (diameters from 12mm to 87mm), 5 cylinders (diameters from 10mm to 70 mm), and 2 flat surfaces of different rigid materials. The total size of the training dataset is around 28815. We only use the data in the loading process to reduce the influence of viscoelasticity.

The CNN model for measuring force and torque is adjusted from VGG-16 net [51], pre-trained on the computer vision dataset ImageNet [11]. We replace the network's last fully-connected layer with an output layer of 4 neurons, corresponding to the forces and torques in 4 axes (Fx, Fy, Fz, Tz). The input of the network is the 3-channel difference image of the current GelSight image and the initial image, when nothing is in contact. We train the network with the mean squared error loss function for the regression problem. To test the model, we use the GelSight data of contacting three novel objects: a ball (d =25mm), a cylinder (d =22mm), and a flat plane. The test set contains 6705 GelSight images under different forces and torques.

The comparison between the output of the neural network and ground truth from the force/torque sensor is summarized in Figure 2-8(b)-(e). The coefficient of determination (R^2) and root mean square error (RMSE) for the results of three different objects are also listed in the figure. The plots show that the model output of forces and torques by GelSight sensor are correlated to the ground truth measured by the force-torque sensor. For the force measurements, R^2 is higher than 0.9. The results also show that the GelSight measurement of force can be robust regardless of the geometry of the contact objects.

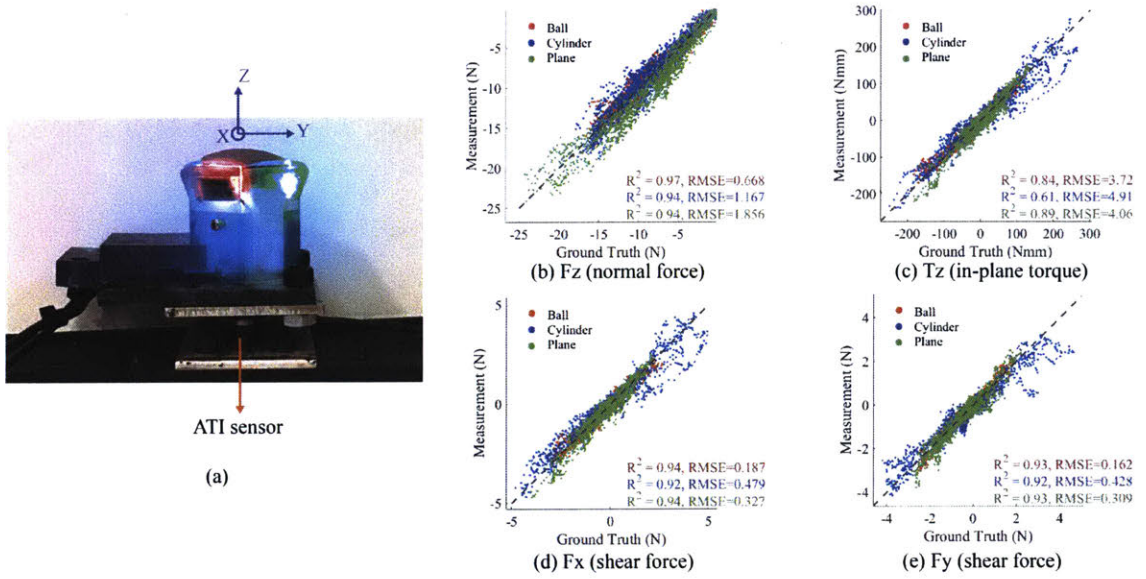


Figure 2-8: Results of the contact force and torque estimation using CNN. (a) Experiment setup. An ATI Nano-17 sensor is attached to GelSight for measuring ground truth, and the contact is conducted manually with different indenters. (b)-(e) Experiment results of the force torque measurement with different unseen indenters.

2.3 Slip detection with GelSight

The GelSight sensor can detect signals related to slip and incipient slip by analyzing the distribution of the markers' motion. Slip can be described as a relative displacement between the tactile sensor's surface and the object in the hold. Before slip occurs, during the increase of the shear force, there is a state called 'incipient slip' that indicates slip is occurring very soon. The incipient slip state can also be described as, part of the object is free from the contact surface, that there has been some relative displacement between the object and the sensor surface, but some other part remains stuck.

A straightforward way to measure slip is to directly measure the movement of the objects and the sensor surface. The planar motion of sensor surface is equal to the motion of the markers, and the motion of the objects is equal to the motion of the geometry on GelSight view. Figure 2-9 gives two examples of the relative displacement of the object geometry and the markers, which indicate the translational slip or rotational slip. We firstly crop a small window of the contact geometry, and

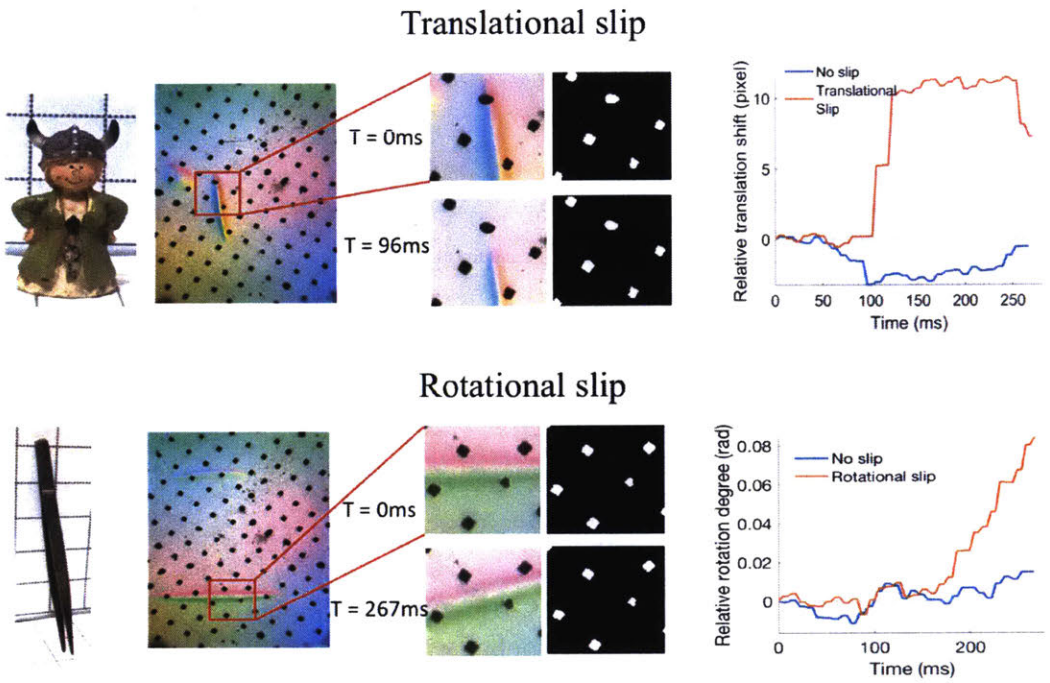


Figure 2-9: Slip detections by detecting the motion of object shapes. The plots are about the relative displacement or rotation angles between the geometry and marker motions along the time in the cropped patch.

then measures the motions or rotation of the geometry, which inferred from the color of the image, as well as the marker patterns. If there is a gap between them, slip has occurred.

In another case, the object surfaces are flat or near flat, with little shape textures. So, the contact area is usually large, but the motion of the object can hardly be tracked by measuring the geometry. But the incipient slip information can also be inferred from the motion distribution of the marker field. By intuition, the incipient slip occurs from the border of the contact, so that the markers in the peripheral contact area have smaller movement in the shear direction compared to the markers in the central contact area, since these markers are still moving with the object. This difference in the motion causes an inhomogeneous distribution of the marker motion within the contact area. [23] used a similar feature to detect slip with a vision-based soft sensor.

Examples of how the displacement field changes as the shear load increases are

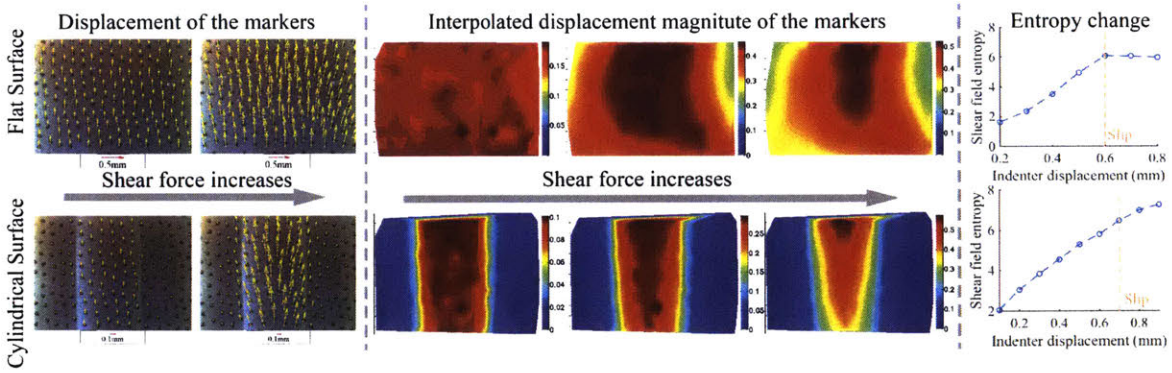


Figure 2-10: The marker displacement fields under the increasing shear force. Incipient slip can be indicated by the smaller motion of the markers in the peripheral contact area. Inhomogeneity of the motion can be measured by entropy, and is shown in the right plots.

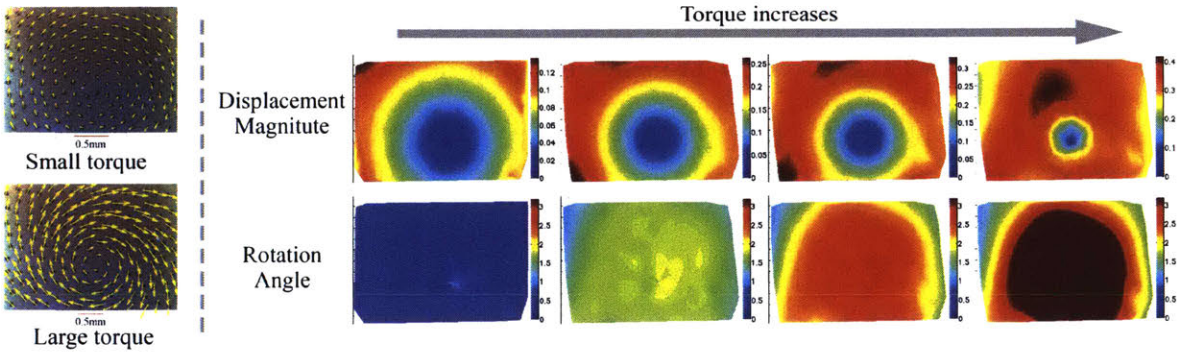


Figure 2-11: The marker displacement fields under the increasing in-plane force. Incipient slip causes the inhomogeneity of the markers' rotation motion around the rotation center.

shown in Figure 2-10 (experiments done with the fingertip GelSight sensor with flat sensor elastomer piece). The degree of partial slip can be inferred from the inhomogeneity degree of the marker displacement distribution. We use the entropy to describe the inhomogeneity degree of the displacement field. The entropy of a histogram X is

$$H(X) = -\int_X p(x)\log p(x)dx \tag{2.2}$$

The entropy H increases as the partial slip degree increases, as shown in the last column in Figure 2-10. To prevent the slip occurrence, a possible way is to set a warning threshold on H.

Rotational slip, which is commonly seen in the failure of grasping, can be inferred

in a similar way. The incipient slip also starts from the peripheral area, and causes an inhomogeneous distribution of the markers' rotational angles along the rotation center. Figure 2-11 shows examples of how the marker motion and rotation angle change as the in-plane torque increases.

The slip detection helps a robot to better perform grasping. We conducted a grasping experiment with general objects [13]. The robot used the combination of two methods to detect slip: for small objects with obvious surface curvatures or textures, the robot detected slip by measuring the relative movement between the object shape and the marker motion; for large objects with smooth surfaces, the robot detected slip by measuring the inhomogeneity of the marker motion within the contact area. Experiments show that, for a dataset of 37 different natural objects, the robot can predict slip, either translational slip or rotational slip, in 84% of the cases. The robot can also conduct re-grasp with larger normal force, when the slip is detected. In this experiment, the robot successfully lifted all the objects in 89% of the cases.

Chapter 3

Material property perception: Hardness estimation

Humans learn a significant amount of information about the objects around them through touch [34, 59], including hardness. We use it to understand the objects, such as whether a tomato is ripe, or whether a sofa cushion is comfortable to sit on. The knowledge of object hardness helps humans to quickly recognize some special objects, evaluate objects for multiple goals, or select corresponding manipulation strategies. Hardness is defined as the resistant force of a solid matter when a compressive force is applied, or in other words, the ratio between the displacement created by an indentation and the contact force. And indeed, hardness-measuring devices such as durometers generally work by measuring the indentation produced by a known force.

In this chapter, I introduce the principle and methods of using GelSight sensor to estimate hardness of arbitrary objects in a loosely controlled setting. To process the raw data from GelSight, we compared the statistical models and the deep learning models that were designed for computer vision. The content of this chapter is published in [72, 74].

3.1 Background

3.1.1 Existing work

Robotic researchers have been working on enabling robots to learn the material properties of objects as well. Two examples are Drimus et al. [14] and Chu et al. [7], who introduced methods to infer multiple object properties by analyzing touch sensor input during several controlled exploration procedures. There has also been work on specifically estimating hardness of objects.

Some of the existing work on hardness estimation uses the basic principle of hardness, i.e., by measuring the contact displacement and force when poking the objects with a robot, thus to infer the object hardness. Su et al. [55] measured the hardness of flat rubber samples using a BioTac touch sensor, which measures multiple tactile signals including the pressure on fixed points. In the work, the sensor was installed on a robot fingertip, and it is pressed onto flat silicone samples in a strictly controlled motion. Changes in the force were then used to discriminate between 6 samples with different hardness values ranging from 30 Shore00 rubbers to rigid aluminum. For this method to be applicable, however, the sensor movement and object geometry must be strictly controlled.

Some other existing works focus on designing special touch sensors for measuring the hardness of specific object categories. The special sensors can measure the force and displacement in a constrained way. For example, Shimizu et al. [50] designed a piezo-resistant cell with a gas-filled chamber, which was used to measure the indentation of the mesa on its top surface. The cell thus measures the material's hardness from the force measured by the pressure change in the chamber and the indentation depth measured by piezo-resistance. The sensor makes measurement easier, under the condition that the surface geometry is certain. These limitations constrain the use of the sensor for more general touch tasks. Okamoto et al. [44] introduced a round shaped soft tactile sensor with strain gauges embedded that measures object roughness, friction, and hardness. They showed that the sensor had distinctive output signals when testing three samples with different Young's moduli. There are also

some sensors designed specifically for measuring hardness for medical use, such as [45, 40]. Another method is to correlate the ultrasonic signal and tissue hardness and researchers have designed structures to measure vibration and resonance frequency for tissue tests [57]. Unfortunately, those sensors are not generalized for measuring the hardness of common objects or using in other tactile tasks. Those sensors, although effective for the specific tasks, can hardly be used on other general tactile tasks. Also, they have constraints on the shape of the target objects as well.

3.1.2 Challenges and contribution

The existing methods on measuring object hardness with robot tactile sensors are largely constrained, in that they only work on a single shaped object (usually flat surfaces), and they require either precisely controlled contact situation or some specialized sensor design. Those constraints made the technology not practical to be applied to real robots in the real world environment. In the real-world scenario, the contact can hardly be well controlled considering the noise in the robot motion control system and the arbitrary alignment of the objects. The complicated geometries of the natural objects also make a big barrier on estimating the object hardness, and most of the common sensors have a measurement noise that could produce a big error in the estimated hardness value. However, humans can figure out the basic properties under all the scenarios. They use touch in a different way. Srinivasan and LaMotte [53] showed with experiments that humans can estimate hardness very well with a passive fingertip, via cutaneous touch alone, evidently based on the deformation pattern of the fingertip, and that kinesthetic information is not essential. Taking the lesson from human system, it should also be possible to estimate the hardness from the local surface change on the shape.

Similar to humans' environment, if a robot wants to estimate the material properties like hardness in the real-world environment, the necessary conditions include:

1. Being able to estimate the hardness of objects with arbitrary shapes.

2. Being able to estimate the hardness of objects when the contact motion is not

precisely controlled.

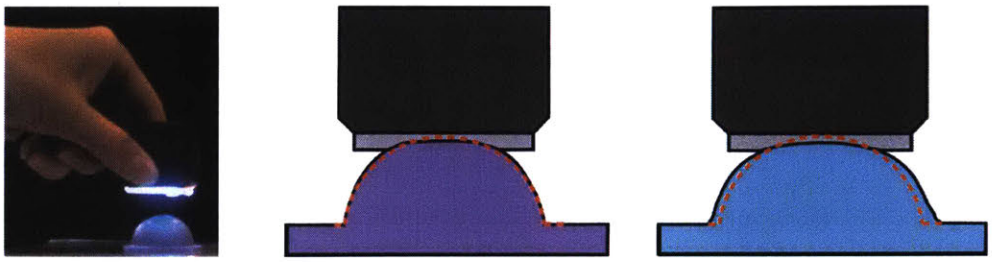
In this part of the work, we focus on exploring the possibility of estimating the hardness of arbitrary objects using the high-resolution tactile sensor GelSight, and under the condition of loosely controlled contact. Particularly, we collect a large training dataset of human testers pressing the GelSight sensor against different objects, where the contact trajectory and velocity are uneven and unknown. Experiments show that the GelSight data contains information about the object hardness regardless of the contact mode; thus by building a numerical model, we are able to measure the hardness of objects from only the GelSight data.

3.2 Principle

The GelSight sensor can estimate the hardness of target objects by measuring the deformation and contact force during the normal contact. An example is shown in Figure 3-1: the GelSight sensor is pressed against a hemispherical object, and under the normal pressure, the object deforms. However, if the object is softer, the deformation will be larger, under the same or smaller normal force. The deformation of the object resulted in a change in the surface geometry. The GelSight sensor measures object's geometry at the contact surface, as well as an estimation of the normal force, thus the information can be used to estimate the hardness.

Examples of the GelSight data when contacting silicone samples of different hardness levels are shown in Figure 3-2. In the image data, the change in the color intensity is correlated with the local curvature of the object surface, and the magnitude of the marker motion is correlated to the normal force. As a comparison, when contacting objects of the same geometry, since the softer objects make a larger deformation, it naturally causes a flatter surface and a smaller normal force. Thus, in the GelSight data, the intensity change of the image is smaller, and the magnitude of the marker motion is smaller. Intuitively, it is possible to build a numerical model to map the change of the image intensity and marker motion to the hardness of the objects.

Note that in this system, the relationship between the object hardness and the



(a) (b) (c)

Figure 3-1: When a GelSight sensor contacts a deformable object, the object deforms. But for a harder sample shown in (b), the deformation is smaller than the one when contacting a softer object in (c).

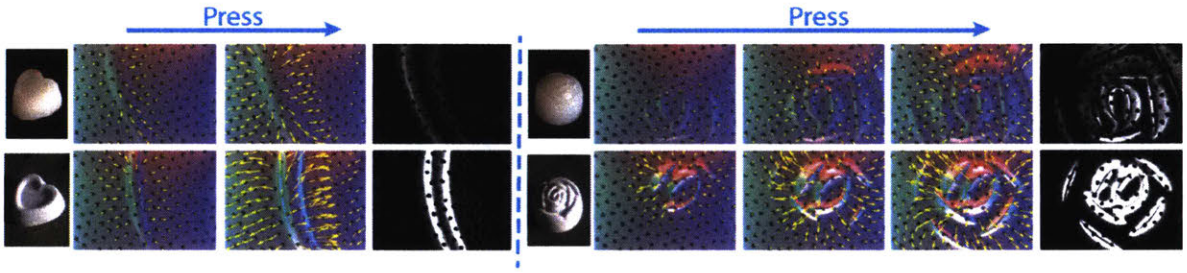


Figure 3-2: GelSight data when contacting softer silicone samples (first row) and the harder ones (second row), in the temporal order. The colored figures show the raw outputs from GelSight and the motion field of the markers, and the gray images show the intensity change of the GelSight images under the largest contact force.

GelSight data is intrinsic and caused by the material property of the target objects and the GelSight sensor, so that the contact trajectory and velocity have little influence on the correlation. So that it should be possible to estimate the hardness of the objects from merely the GelSight data, while the contact motion is unknown. As a result, the method could be applied on cheaper robots and more real-world contact situations as well, where the contact trajectory is hard to control. In the research, to make sure the system is robust, we collect most of the training data manually, as shown in Figure 3-1(a), so that both the trajectory and velocity of the contact motion are uneven and unknown.

3.3 Dataset

We use statistical learning method to estimate the object hardness from the high-dimensional GelSight data. To apply the learning method, building a dataset of different GelSight data when the sensor contacts models of different shapes and hardness levels are necessary. We cast a set of silicone samples as the object stimuli, and collect the dataset of GelSight data by either manually pressing GelSight on the samples, or squeezing the samples with a parallel robot gripper.

The silicone samples are cast by three kinds of materials: Ecoflex® 00-10 (hardness of Shore 00-10), Ecoflex® 00-50 (hardness of Shore 00-50) and Smooth-Sil® 945 (hardness of Shore A-45, or 87 in Shore 00 scale) from Smooth-On Inc. The materials are mixed in different ratios, so that it produces samples of the hardness levels between that of the raw materials. In total, we produced 16 hardness levels from Shore 00-10 to Shore 00-87, and use the single number in Shore 00 scale to denote the hardness. The groundtruth hardness is measured by a PTC® 203 Type OO durometer on the groundtruth-test sample – the thick samples with flat surfaces and made from the same silicone mixture. To reduce the measurement error, we took 5 tests and use the mean value.

As a comparison, we choose the hardness of the GelSight sensor of 17 in Shore 00 scale. It is relatively soft. The hardness of the GelSight sensor influences the sensitive range of the estimation of the object hardness, and the chosen hardness makes the sensor sensitive to the soft objects.

For the shape of the object stimuli, we consider 4 situations:

1. Basic shapes: spheres and cylinders of different radii.

2. Basic shapes: flat surfaces, edges, and corners.

3. Arbitrary shapes.

4. Natural objects with irregular shapes.

For making the silicone samples that are of well-controlled shapes, we make a set of molds of cylindrical and spherical shapes. Most of the molds are 3D printed, as

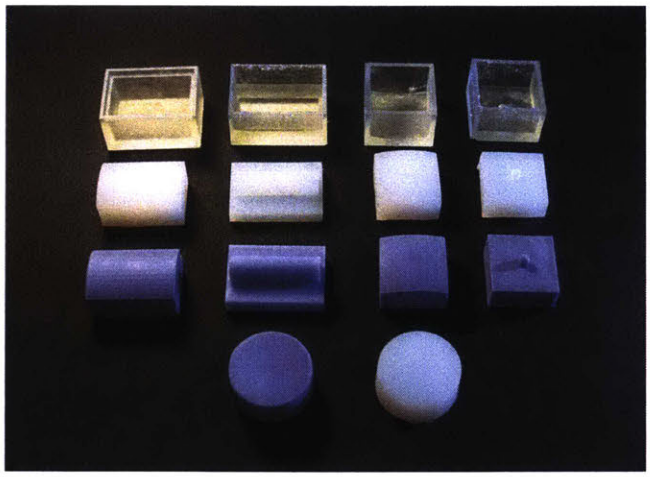
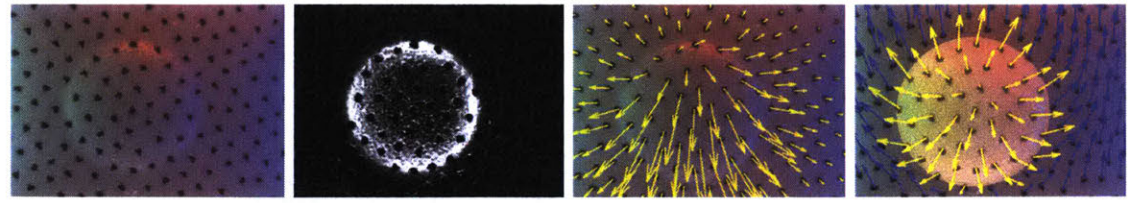


Figure 3-3: 3D printed molds and the casted silicone samples. The silicone samples are used as stimuli objects.

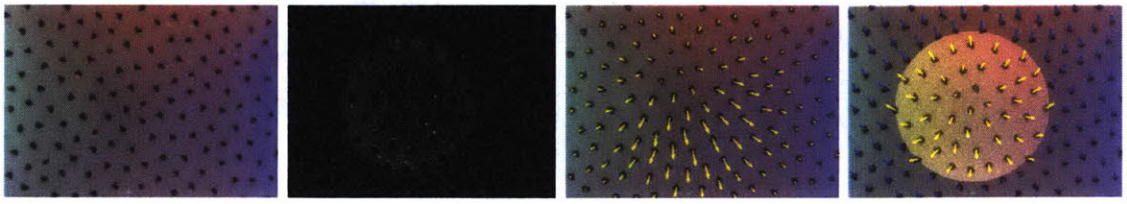
shown in Figure 3-3, and a small portion of the molds are taken from some commercial products that satisfy the shape requirement. The cylindrical and hemispherical silicone samples made from the molds, as shown in Figure 3-3, are of 9 different radii ranging from 2.5 mm to 50 mm, but the heights are similar (around 25 mm). This design is to ensure the sample thickness would have limited influence of on the shape change during the press. The tactile data of contacting edges and corners are collected by touching the edges and corners of the spherical or cylindrical samples.

The samples of arbitrary shapes have two groups: one is of simple and common shapes, that is cast from some daily vessels, like square shaped ice boxes, truncated cone-shaped measuring cups, small beakers; another group is of complicated and special shaped, cast from assorted chocolate molds. They include the shapes of different emboss textures or complicated curvatures, like the shape of seashells.

In total, we made 95 hemispherical silicone samples, 81 cylindrical samples, 15 flat samples and 160 samples of arbitrary shapes for experiments. We also used a set of natural objects as the stimuli, including tomatoes, avocados, mangoes, human body part, etc.. It is impossible to get the ground truth hardness of those objects, but we use the human estimation of their hardness as a reference.



(a) GelSight image (b) dI (c) Marker Motion $\mathbf{\Phi}$ (d) $\mathbf{\Phi}_N$



(e) GelSight image (f) dI (g) Marker Motion $\mathbf{\Phi}$ (h) $\mathbf{\Phi}_N$

Figure 3-4: GelSight data when pressing on hemispherical silicones ($R = 12.7$mm). Stimulus for the fist row is harder than that of the second row. (b) and (f) show the intensity change dI in the GelSight images; (c) and (g) show the displacement field of the markers $\mathbf{\Phi}$; (d) and (h) show the $\mathbf{\Phi}_N$ field decomposed from $\mathbf{\Phi}$, which is caused only by normal force. The contact area is masked in the yellow.

3.4 Statistical learning method for estimating hardness of spherical objects

As stated in Section 3.2, the hardness of the target objects is directly correlated with the intensity change in the intensity change in the GelSight images, and the motion of the markers under the normal force. In this section, I introduce our work on building a statistic model of the change in GelSight image intensity change, marker motion, and the target objects' hardness. As a simplification, we only study the data of contacting hemispherical samples in this section. However, we aim at estimating hardness of samples of different radii, while the model is learned based on the contacting case of the samples of the same radius (12.7mm).

Figure 3-4 shows an intuitive example of how the GelSight data is like. From the raw GelSight data (a) and (e), we calculated the 1-channel intensity change map, denoted as dI, and the displacement field of the markers, denoted as Φ. The motion of the markers is the result of both normal force and shear force, while only normal force is informative in estimating the object hardness and the shear forces are un-

controlled noise. Thus, we decompose Φ and get the displacement field Φ_N that is caused by only the normal force, as shown in (d) and (h).

To estimate the hardness of the target samples, we first study the independent correlations of dI and Φ_N to the object hardness H, and then estimate the hardness from the cues together.

3.4.1 Modeling the shape change

The intensity change in the GelSight image is non-linearly correlated to the gradient of the surface geometry. A theoretically precise way to measure the gradient is using calibration and look-up table, but in this project, we use the intensity directly. This is because the precision of the lookup table for the 1st generation of the GelSight sensor is not high enough for measuring the sensitive information like the hardness of the objects, but the intensity change can preserve more raw information, and thus reduce measurement error.

Supposing the original GelSight image is $\mathbf{I} = (r, g, b)$, where r, g, b corresponds to the intensity value in the 3 color channels, and $\mathbf{I}_0 = (r_0, g_0, b_0)$ denotes the initial GelSight image, where there is no contact. $dI = (r, g, b, r_0, g_0, b_0)$ is a monotonic function that stably denotes the intensity change. The function differs according to the illumination systems of different sensors. For the 1st generation Fingertip GelSight sensor that is used in this project, we choose an experimental function

$$dI = 4.4 \times \mathrm{rgb}_3 + 2.2 \times \mathrm{rgb}_2 + 0.4 \times \mathrm{rgb}_1 \tag{3.1}$$

where

$$[\mathrm{rgb}_1, \mathrm{rgb}_2, \mathrm{rgb}_3] = \mathrm{sort}\left(\frac{r}{r_0}, \frac{g}{g_0}, \frac{b}{b_0}\right) \tag{3.2}$$

A major consideration for choosing the function dI is to make sure the value is robustly mapped to the geometry gradient of the surface, and invariant of the contact location and the gradient direction. A test result on the gradient mapping on different calibration location for the function 3.1 is shown in Figure 3-5.

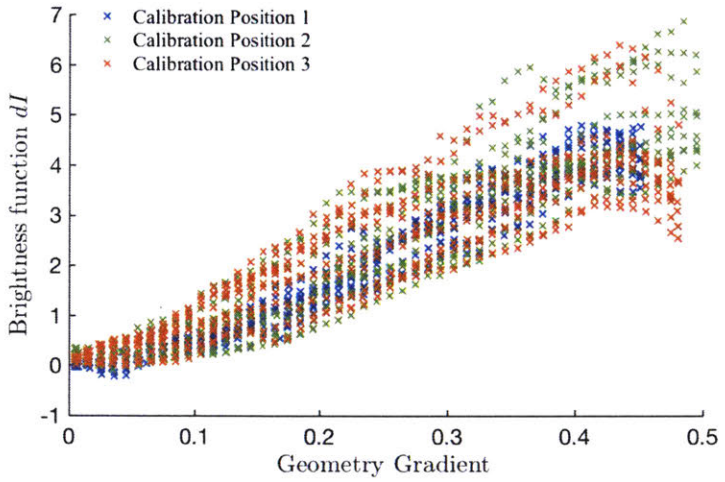


Figure 3-5: Correlation between intensity change function dI (Equation 3.1) and the ground truth gradient of the local geometry. Different plots indicate different contact positions on the sensor surface.

Examples of dI when contacting hemispherical objects is shown in Figure 3-4(b)(f) (Note that the black marker areas are excluded). The circular area of the high dI is the contact area, and it is straightforward to measure its radius $r_{\max}$. $r_{\max}$ is also monotonically related to the pressing force and depth. Within the contact area, dI grows larger when approaching the border, indicating larger surface gradient. To reduce the noise in the measurement, we calculate the mean dI within every ring area along the contact center, and denote it as $\bar{dI}(r)$. Figure 3-6 shows two examples of the growth of $\bar{dI}(r)$ along the radial direction, and it can be fitted with a binomial function

$$\hat{dI}(r) = p_1 \times r^4 + p_0 \tag{3.3}$$

and the peak value is measured as $\bar{dI}_{\max}$. Overall, $\bar{dI}(r)$ of pressing on the softer sample is significantly smaller than that when pressing on a harder sample.

Both $\bar{dI}_{\max}$ and the p_1 in Function 3.3 well indicate the sample hardness. For the target objects of hemispherical silicone with a radius of 12.7mm, the relationship of $\bar{dI}_{\max}$ to the contact radius $r_{\max}$ is shown in Figure 3-7(a), which is close to a linear relationship $\bar{dI}_{\max}(r_{\max}) = f_{dI} \times r_{\max} + p_0$, and the linear coefficient f_{dI} is monotonically related to the hardness. The relation of f_B to the hardness is shown

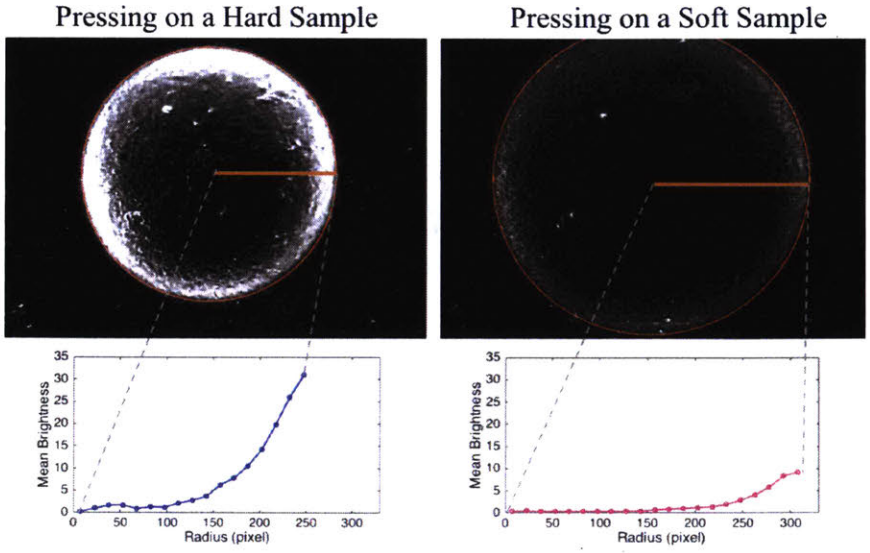


Figure 3-6: The intensity change of GelSight image along the radial direction within the contact area. It can be fit with a binomial function $\hat{dI}(r) = p_1 \times r^4 + p_0$.

in Figure 3-7(b). We fit the data with an offset exponential function

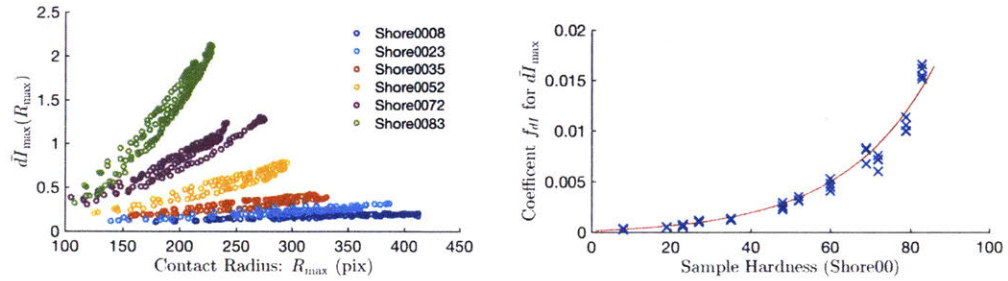
$$\hat{f}_{dI} = A_I \exp(B_I \times H_I) + C_I, \tag{3.4}$$

so that if we measure f_{dI} from a sequence of GelSight data during a press, we can predict the hardness $\hat{H}_I$ using $\hat{f}_{dI}$.

Another hardness indicator is p_1 in Function 3.3. During a press, p_1 changes as the contact radius $r_{\max}$ increases, and the relation differs as the sample hardness differs, as shown in Figure 3-8. The relationship can be approximated as

$$\begin{bmatrix} \log(p_1) & \log(r_{\max}) & 1 \end{bmatrix} \mathbf{b} = \hat{H}. \tag{3.5}$$

We use linear regression to obtain $\mathbf{b}$ from the training data set, and make a prediction of hardness $\hat{H}_p$ by averaging all the $\hat{H}$ in a single pressing sequence, where $r_{\max}$ differs.



(a) (b)

Figure 3-7: (a): Maximum intensity change $\bar{dI}_{\max}$ during multiple presses on silicone samples with different hardness levels. (b): The relationship of the linear coefficient of $\bar{dI}_{\max}$, the f_{dI}, against $R_{\max}$ for different hardness levels. The data is fitted with an exponential function $\hat{f}_{dI}$, shown as the red line.

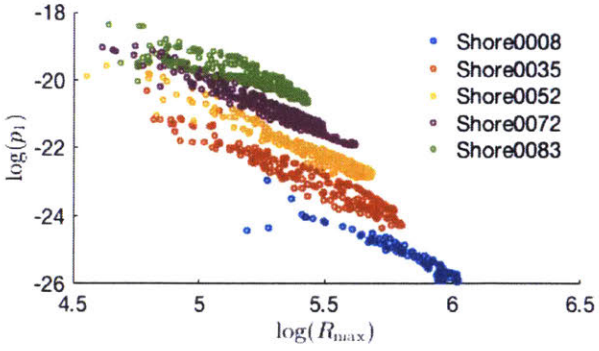


Figure 3-8: Relationship of $\log(p_{1,R_{\max}})$ and $\log(R_{\max})$ when contacting silicone of different hardness levels.

3.4.2 Modeling the marker motion

The motion field of the markers on the GelSight surface indicates the type and magnitude of the contact force. Particularly, for sensors surfaced with a thin elastomer, the resultant vector field can be approximately considered as the linear sum of the fields caused by different forces. Thus, the overall displacement field $\mathbf{\Phi}(x, y)$ has the expression

$$\mathbf{\Phi} = \mathbf{\Phi}_N + \mathbf{\Phi}_S + \mathbf{\Phi}_T, \tag{3.6}$$

where $\mathbf{\Phi}_N$ is the field caused by normal force, $\mathbf{\Phi}_S$ is the field caused by shear force, and $\mathbf{\Phi}_T$ is the field caused by torque. In this set of experiments of pressing GelSight on hemispheres, the contact geometry is rotationally symmetric, and the shear force is relatively small so that no partial slip exists. So, the displacement field can be written in polar coordinates $\mathbf{e}_r, \mathbf{e}_\theta$ with the origin as the contact center, and within the contact region, there are

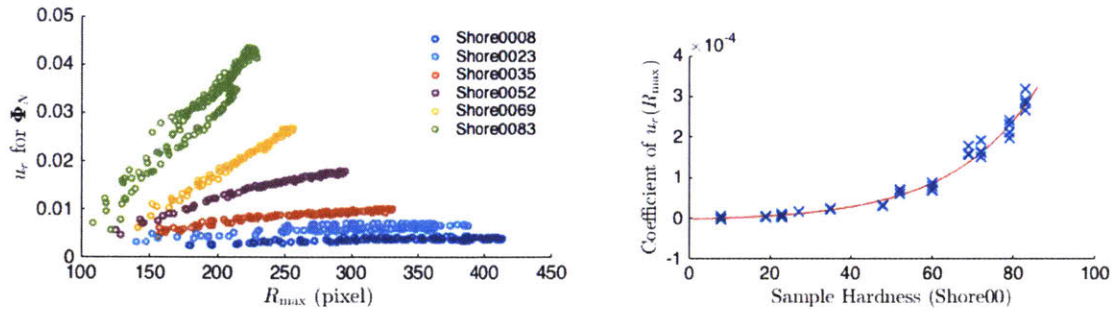
$$\begin{aligned} \mathbf{\Phi}_N(r, \theta) &= U_r(r)\mathbf{e}_r \\ \mathbf{\Phi}_S(r, \theta) &= u_x\mathbf{e}_x + u_y\mathbf{e}_y \\ \mathbf{\Phi}_T(r, \theta) &= U_\theta(r)r\mathbf{e}_\theta \end{aligned} \tag{3.7}$$

According to experimental results, $U_r(r)$ and $U_\theta(r)$ can be simplified as

$$U_r(r) = u_r r, U_\theta(r) = u_\theta \tag{3.8}$$

u_r, u_θ, u_x, u_y are all constants related to the magnitudes of the external force or torque. As an approximation, they are linearly related to the force or torque magnitudes. We decompose the displacement field $\mathbf{\Phi}$ within the contact area according to (3.8) and (3.7). For estimating hardness, only normal force, or the field $\mathbf{\Phi}_N$ is in concern.

We decompose $\mathbf{\Phi}$ using a method that combines image registration and image pyramids. First, we mark the contact area for the press according to the image intensity change, and consider $\mathbf{\Phi} = [u_x, u_y, u_r, u_\theta]^T$. The vector is calculated using



(a) (b)

Figure 3-9: (a)The relationships of u_r and the contact area radius $R_{\max}$, for samples of different hardness levels. (b)The coefficient f_M for $u_r(R_{\max})$, and the exponential fit function $\hat{f}_M$.

image registration, from different scales of the image.

Figure 3-4(c)(g) shows examples of the marker displacement field $\mathbf{\Phi}$, and the corresponding $\mathbf{\Phi}_N$ in the contact area shown in (d)(h). The two examples are pressing experiments with different levels of hardness, so that although the contact area is the same, the magnitudes of $\mathbf{\Phi}$ differ. Pressing on softer samples makes smaller u_rs, indicating that the normal forces are relatively smaller. In the loading period, u_r also increases as the contact radius $r_{\max}$ and the normal force increase. The relationship of u_r to $r_{\max}$ is shown in Figure 3-9(a), that when pressing on the same sample, there is an approximately linear relationship $u_r(r_{\max}) = f_M r_{\max} + p_0$. The linear coefficient f_M is positive with respect to sample hardness, as shown in Figure 3-9(b). We fit an offset exponential function

$$\hat{f}_M = A_M \exp(B_M \times H_M) + C_M, \tag{3.9}$$

for f_M, so that we can predict the hardness $\hat{H}_M$ from measured f_M in a pressing sequence.

3.4.3 Estimating hardness of sphere objects

According to Section 3.4.1, 3.4.2, we can obtain three hardness prediction $\hat{H}_B$, $\hat{H}_p$ and $\hat{H}_M$ from a sequence of GelSight images during a single press. We make a final

Table 3.1: Results on training and tests set with a single-shaped spherical samples

	$\hat{H}_B$	$\hat{H}_p$	$\hat{H}_M$	$\hat{H}_A$
R^2 train (human data)	0.9946	0.9954	0.9921	0.9978
RMSE train (human data)	3.4123	3.1539	4.1147	2.1846
R^2 test (robot data)	0.9910	0.9951	0.9794	0.9952
RMSE test (robot data)	4.4349	3.2777	6.7078	3.2218

hardness prediction $\hat{H}_A$ as the linear combination of the three predictions, such that

$$\hat{H}_A = \begin{bmatrix} \hat{H}_I & \hat{H}_p & \hat{H}_M & 1 \end{bmatrix} \mathbf{b}_A \tag{3.10}$$

Vector $\mathbf{b}_A$ is trained through linear regression.

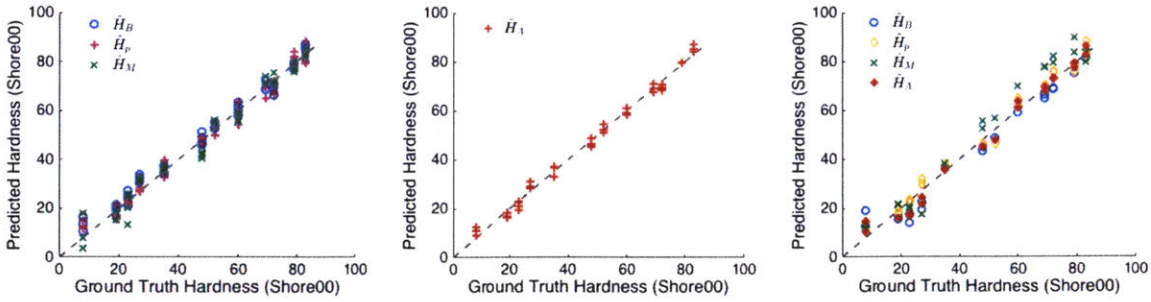
3.4.4 Experiment

For training the statistical model, we press the GelSight sensor manually on 12 silicone samples, which have the same shape: hemisphere with the radius of 12.7mm. The hardness of the objects ranges from Shore 00-08 to Shore 00-83. We make 46 press sequences on the samples.

On the training dataset, the prediction $\hat{H}_B$, $\hat{H}_p$, $\hat{H}_M$, which are for different features, are shown in Figure 3-10(a), and the fitting error shown in Table 3.4.4. The overall prediction $\hat{H}_A$ is shown in Figure 3-10(b), and the error shown in Table 3.4.4. Those results show that the separate and overall hardness predictions are all very close to the ground truth on the training dataset, with an R^2 (coefficient of determination) of 0.9978 and a root mean square error (RMSE) of 2.18 on Shore 00 scale. In a 10-fold cross-validation test, the RMSE ranges from 0.8441 to 3.8849, with the mean being 2.4279 and the standard deviation of 0.6771.

We conduct two experiments for testing: one on the same set of silicone samples, but different contact mode; the other one is on hemispherical silicone samples of different radii.

For the test on the same set of silicone samples, we make a Baxter robot squeeze on the samples with GelSight sensor on the fingertip, and collect 24 data sequences.



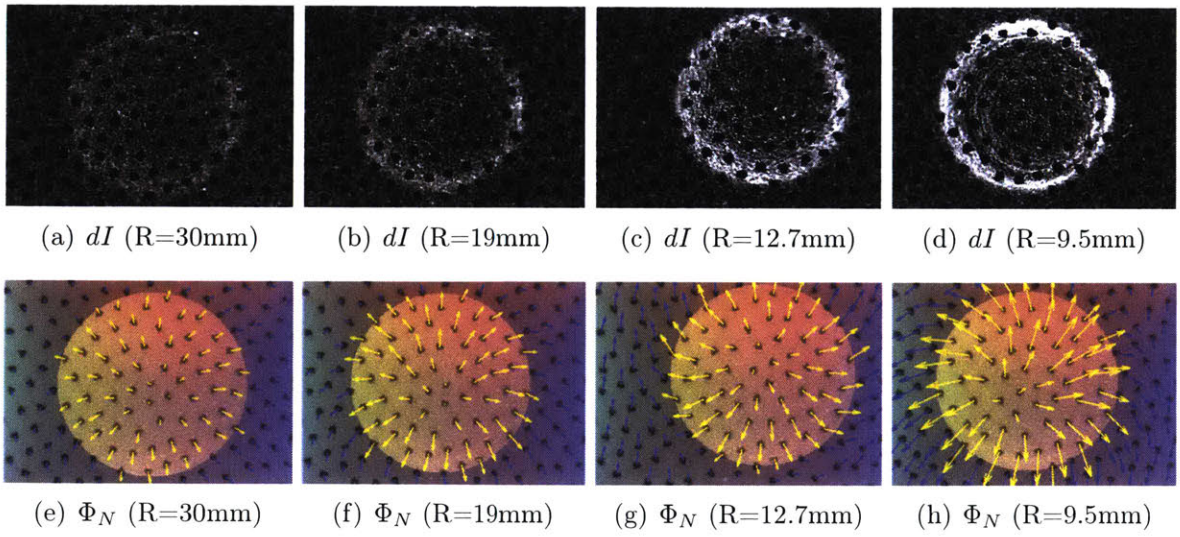
(a) (b) (c)

Figure 3-10: Hardness estimation results with the single-shaped hemispherical samples (R=12.7mm). (a): Independent estimation $\hat{H}_B$, $\hat{H}_p$, $\hat{H}_M$ for different on training set. (b): Overall prediction $\hat{H}_A$ on training set. (c): Results on test set, that the data is collected by a robot, which is different from the manually collected data in the training set.

The robot is controlled in open-loop, and the force and displacement are not well controlled. So, there is large and unpredictable variability between each contact instance, and the variability is larger than human conducted experiments. The result is shown in Figure 3-10(c), and the error is reported in Table 3.4.4. It can be seen that the estimation results well match the ground truth, with R squared of 0.9952 and RMSE of 3.22 in Shore 00, although the data noise is of different types. Among the predictions, $\hat{H}_p$ is the most stable, and $\hat{H}_M$ makes the largest error, most likely because the robot introduces large noise in pressing force during the measurement, leading to a u_r measurement that makes a much larger error. The final result shows that the model is very robust regardless of contact modes.

In the second test experiment, we consider the stimuli objects that are of hemispherical shape but 3 different radii: 30 mm, 19 mm, and 9.5 mm, while the model is trained based on the contact examples of samples of 12.7 mm radius. In general, the GelSight data has similar features and correlation to the sample hardness for stimuli of different radii, but when the target sample is of smaller radii, the intensity change of GelSight images tends to be larger, and the force is likely to be larger when the contact area is the same. Examples of the GelSight data are shown in Figure 3-11.

In general, when pressing on samples of different dI and u_r have a similar relation-



(a) dI (R=30mm) (b) dI (R=19mm) (c) dI (R=12.7mm) (d) dI (R=9.5mm)

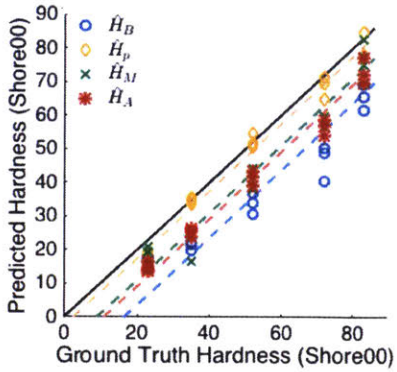
(e) Φ_N (R=30mm) (f) Φ_N (R=19mm) (g) Φ_N (R=12.7mm) (h) Φ_N (R=9.5mm)

Figure 3-11: GelSight data when pressing on hemispherical samples of different radii but same hardness level (Shore 00-35).

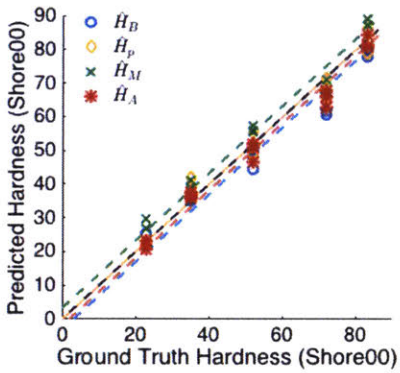
ship to $r_{\max}$ as shown in Figure 3-7, Figure 3-8, and Figure 3-9, but the parameters differ. If using the same parameter set, the hardness estimation result is shown in Figure 3-12(a)(b)(c) and Table 3.2, Table 3.3 (lines of 'before fixing'). It can be seen that there is a linear deviation between the prediction and the ground truth.

The shift value can be calculated from the target sample's radius, if it is known. That is based on the assumption that when the shape of the stimuli remains the same while the dimension differs, it can be assumed that the GelSight data is also scaled on the spatial dimension. Supposing the stimuli silicone samples in the training set has a radius of R_0, which is 12.7mm in this experiment, and the radius of the target stimuli object is R. So that the dimension ratio between the test case and the training case is $\alpha = \frac{R_0}{R}$. For the measurement on the GelSight data, we denote a normalized radical measurement $\hat{r} = \alpha r$, and use it instead of r in Equation 3.4, 3.5, and 3.9. Considering $C_I \ll \hat{f}_d I$ and $C_M \ll \hat{f}_M$, the predicted hardness value H^α after the fix of the sample radius can be denoted by

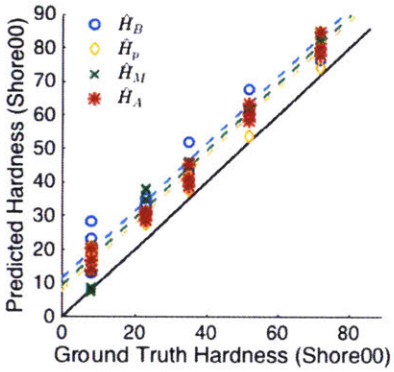
$$\begin{aligned} \hat{H}_I^\alpha &= \hat{H}_I + \tfrac{1}{B_I}\log\alpha \\ \hat{H}_p^\alpha &= \hat{H}_p + \log\alpha \begin{bmatrix} 4 & 1 & 0 \end{bmatrix} \mathbf{b} \\ \hat{H}_M^\alpha &= \hat{H}_M + \tfrac{1}{B_M}\log\alpha \end{aligned} \tag{3.11}$$



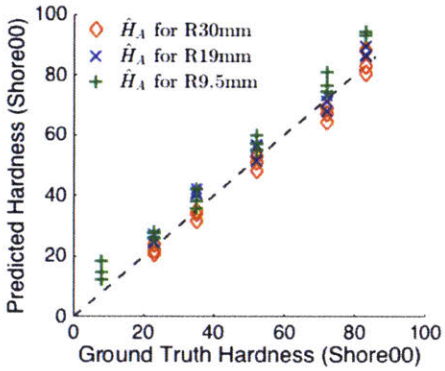
(a) Sample R=30mm



(b) Sample R=19mm



(c) Sample R=9.5mm



(d)

Figure 3-12: For the model trained on samples with R=12.7mm, the hardness estimation of the objects with different radii. (a)(b)(c) show the prediction using the same model; (d) shows the prediction after linear fixing using Equation 3.12.

Where $\hat{H}_I, \hat{H}_p, \hat{H}_M$ are the predicted hardness using the model without fixing. Thus,

$$\hat{H}_A^{\alpha} = \hat{H}_A + \log\alpha \left[\frac{1}{B_I} \quad \begin{bmatrix} 4 & 1 & 0 \end{bmatrix} \mathbf{b} \quad \frac{1}{B_M} \quad 0\right] \mathbf{b}_A \tag{3.12}$$

Figure 3-12(d), Table 3.2 and Table 3.3 (lines of 'fixed'). showed the estimation result of $\hat{H}^{\alpha}$. It can be seen that after fixing the estimation error is reduced largely, with RMSE around 3.1 in Shore 00 scale.

Table 3.2: R squared of hardness estimation on samples of different radii than the training set

Sample	$\hat{H}_B$	$\hat{H}_p$	$\hat{H}_M$	$\hat{H}_A$
R30mm (before fixing)	0.7768	0.9880	0.9237	0.9084
R30mm (fixed)	0.9680	0.9926	0.9818	0.9927
R19mm (before fixing)	0.9819	0.9907	0.9822	0.9908
R19mm (fixed)	0.9890	0.9907	0.9898	0.9931
R9.5mm (before fixing)	0.9273	0.9651	0.9482	0.9555
R9.5mm (fixed)	0.9899	0.9941	0.9895	0.9952

Table 3.3: RMSE of hardness estimation on samples with different radii than the training set

Sample	$\hat{H}_B$	$\hat{H}_p$	$\hat{H}_M$	$\hat{H}_A$
R30mm (before fixing)	17.6611	4.0970	10.3246	11.3127
R30mm (fixed)	6.6850	3.2177	5.0460	3.1913
R19mm (before fixing)	5.0352	3.6117	4.9824	3.5942
R19mm (fixed)	3.9252	3.6099	3.7666	3.1036
R9.5mm (before fixing)	12.3616	8.5665	10.4372	9.6737
R9.5mm (fixed)	4.6055	3.5117	4.6957	3.1872

3.5 Deep learning for estimating hardness of arbitrary objects

A challenge comes from the geometries of the object – when the objects are of some complicated and unknown shapes. The physical correlation between the hardness of the target objects and the GelSight images, which is demonstrated in Figure 3-1 and Figure 3-2, remains the same. However, building the statistical model of the correlation could be hard. Thus, we turn to deep learning method.

Since the input is simply a sequence of images, it can be analyzed using standard computer vision models that learn end-to-end, directly mapping from pixels to hardness values. We apply a neural network model that is similar to [12], which is used for action recognition: each frame of the GelSight image is represented using the deep convolutional neural network (CNN). Then, to represent the temporal changes in the signal, we use a recurrent neural network with long short-term memory units

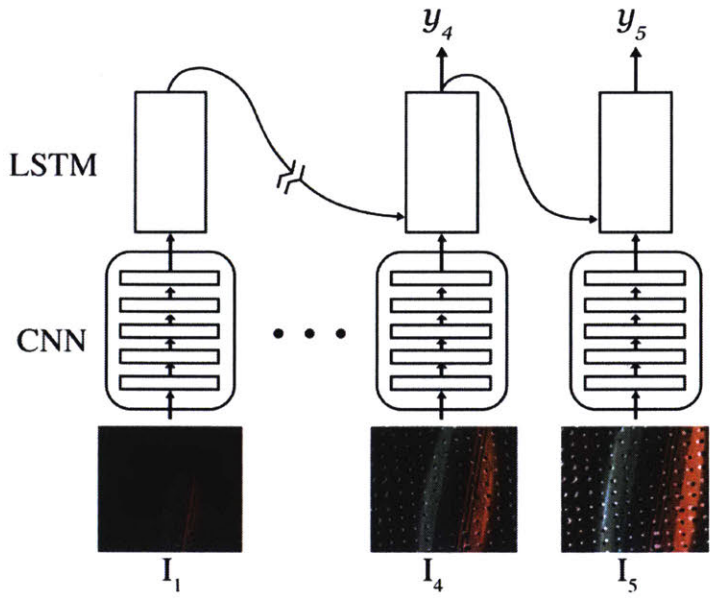


Figure 3-13: Neural network architecture for estimating hardness values from GelSight images. 5 difference GelSight images ($\mathbf{I}(t) - \mathbf{I}(t_0)$) from the sequence of contact, are represented using CNN features *fc7* from a VGG16 net, and feed into an LSTM net. Output values from the last 3 frames, i.e. y_3, y_4, y_5, are used to estimate the hardness.

(LSTM) [21].

3.5.1 Model

The neural network model for estimating hardness is shown in Figure 3-13. The input is a selected sequence of images from the GelSight data during the contact with the object, and each image is represented by a CNN, then the LSTM model, and output a single number of the hardness value in Shore 00 scale. Specifically, for the CNN model, we choose the 16-layer VGG architecture [51] which has been shown to achieve high performance on large-scale object recognition tasks [31, 33].

At each timestep, the model regresses the output hardness value via an affine transformation of the current LSTM's hidden state h_t:

$$\begin{aligned} y_t &= Wh_t + b \\ h_t &= L(h_{t-1}, \phi(I_t)), \end{aligned} \tag{3.13}$$

where W and b define an affine transformation of the hidden state h_t, and L updates h_t based on the previous state h_{t-1} using the current image I_t (here the LSTM's hidden cell state is omitted for simplicity). The prediction y_t is the hardness estimate for the current timestep. We estimate a hardness value for the object as a whole by averaging the predictions from the final 3 frames. This per-frame based regression adds robustness in the hardness estimation. During training, the loss is based on the difference between the predicted and ground-truth hardness values, using a Huber loss.

Choosing the input sequence

The input images to the neural network are from a temporal-ordered selected sequence from the GelSight video of contacting the stimuli objects. In this project, the raw video typically contains 20 to 30 frames, but we choose 5 frames as the input sequence, on the belief that the 5 frames contain enough information about the object hardness. At the same time, choosing a sub-sequence of the images could also make the model invariant to the speed of the pressing motion and to the maximum contact force. For example, different human testers or robots may manipulate objects with different loading speed or maximum force, which result in differences in the sequence distribution. Therefore, we constrain the video sequence so that it begins and ends at times that are consistent across manipulation conditions. The 5 chosen frames are within the loading period of the press, between the trigger moment and the maximum contact moment. They are evenly distributed on the dimension of the pressing proceeding stage. Here the press processing is roughly estimated according to the intensity change of the pixels, that

$$\int dI(t) = \int_{x,y} \big(I_r(t) - I_r(t_0)\big) + \big(I_g(t) - I_g(t_0)\big) + \big(I_b(t) - I_b(t_0)\big). \tag{3.14}$$

where t_0 denotes the moment when there is no contact with the sensor. The process is demoed in Figure 3-14. In the figure, the grey dashed lines denote the point of choosing frames from the original GelSight sequence. Note that, since the frames in the touch sequence is discrete, the selected frames are likely to be not exactly on the

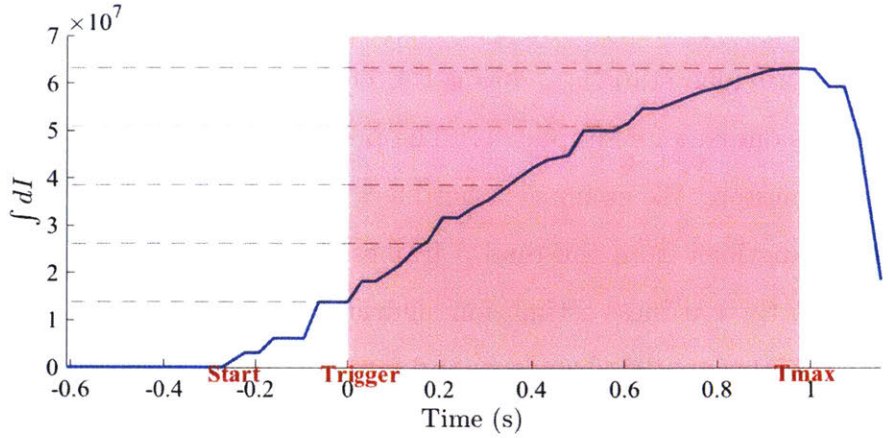


Figure 3-14: Examples of choosing 5 frames from a sequence of the GelSight images of the contact. The blue plot shows the intensity change of the GelSight images during the contact, and the gray lines are the even devision lines between the trigger point and the maximum contact point. The sample frames are chosen on those moments.

$\frac{n}{4}$ division, but they are close to an even division of the contact sequence.

For the input images, we experiment with the raw GelSight image $\mathbf{I}(t)$, the difference image of $\mathbf{I}(t) - \mathbf{I}(t_0)$, and the derivative image of $\mathbf{I}(n) - \mathbf{I}(n-1)$, where n denotes the number of the frames in the 5 chosen ones. As a result, using the difference image $\mathbf{I}(t) - \mathbf{I}(t_0)$ makes the best estimation result. Note that, in the deep learning method, we do not explicitly model the motion of the markers on the sensor – relying instead on the network to learn about their motion via the entire frame.

Training

The training dataset for the network is about 7000 videos obtained by manually controlled contact on silicone samples; each video is an independent pressing sequence. The dataset mainly contains the basic object shapes (Group 1 in Figure 3-15), but also with a large portion of complicated shapes or bad contact conditions (Group 2 and 4 in Figure 3-15). Those irregular shapes help to largely prevent model's overfit. Each video is used for multiple times during the training, with different end points for sequence extract, so that the subsequences represent the data of contacting the samples with different maximum forces. We train the model using stochastic gradient descent, initializing the CNN weights with ImageNet [31] pretraining, jointly training the CNN and LSTM. The algorithm is trained for 10,000 iterations, at a learning rate

of 0.001, and step size of 1000.

3.5.2 Dataset

The dataset of object stimuli is introduced in Chapter 3.3. We contact all the target objects with the Fingertip GelSight sensor, either manually or with a robot, for multiple times, with random contact trajectory or location. The contact is always close to the normal direction. In the human testing scenario, the test object is placed on a flat hard surface, and a tester holds the GelSight sensor and presses on the object vertically; for the robot test, we use a Weiss WSG 50 parallel gripper which has GelSight as one finger. When the objects are within the gripping range, the gripper closes in a slow and constant speed until the gripping force reaches the threshold, making the GelSight sensor squeezing on the object. The speed was randomly chosen between 5 to 7 mm/s, and the gripping force threshold is random between 5 to 9N. In both cases, the sensor is pressed into the objects, while the contact force grows, and the deformation of both the GelSight elastomer and the object increases. The GelSight video is recorded during the contact. In average, there are 20 to 30 frames in the loading period.

Examples of GelSight data when contacting different object stimuli is shown in Figure 3-15. They are divided into 5 groups:

1. *Basic shapes.* Samples of the simplest geometries, including a flat surface, a spherical surface, cylindrical surface, straight sharp edges, sharp corners. Spherical and cylindrical shapes are of 10 levels of radii from 2.5mm to 50mm.

2. *Basic shapes, challenging contact conditions.* The contact objects are the same as in the previous group, but the contact condition is undesired. For example, the sensor contacts the silicone samples in tilted angles, or the sample is included in the contact area.

3. *Simple shapes.* The samples are made into the shapes of natural objects with simple geometries, such as shapes of frustum measuring cup.

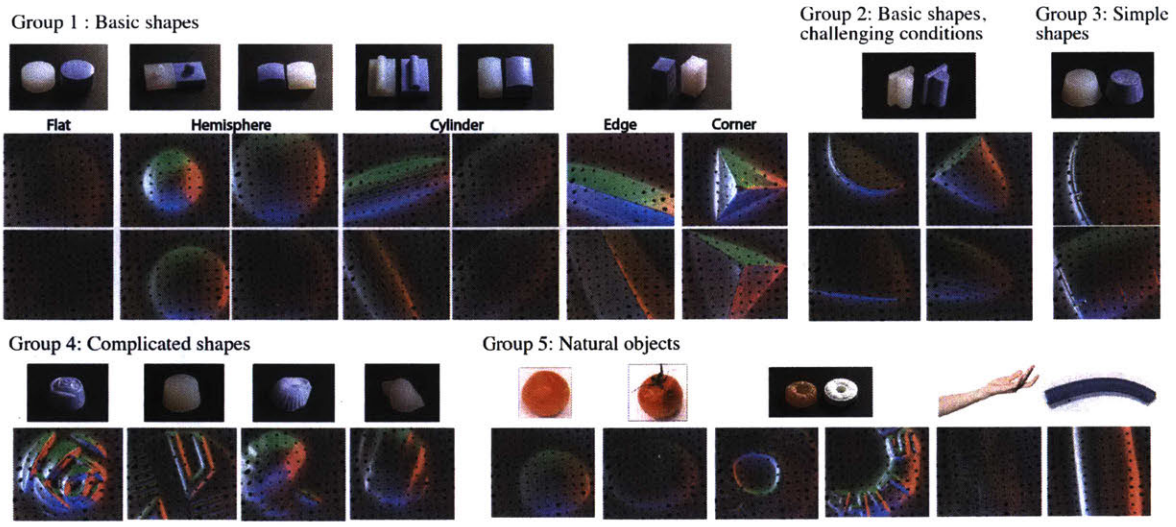


Figure 3-15: Examples of the GelSight dataset when contacting objects of different shapes and hardness levels. Data is divided into 5 groups: the basic shapes, basic shapes when contacted on undesired locations; shapes of natural objects with simple geometry; complicated shapes made by chocolate molds; natural objects.

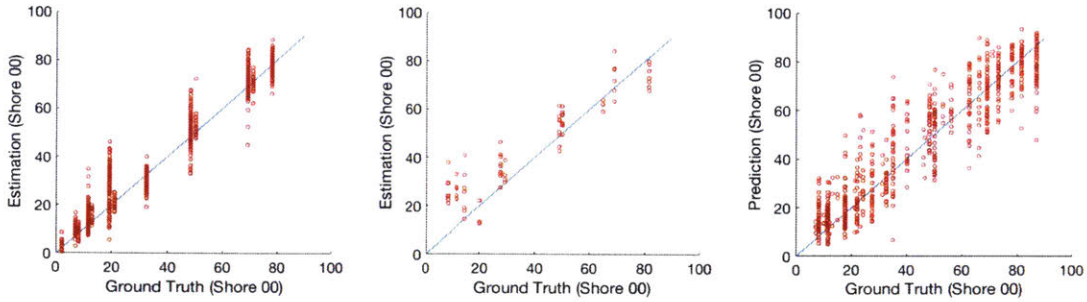
4. *Complicated shapes.* The samples are made of silicone but with complicated textures or shapes. They are made from the chocolate molds.

5. *Natural objects.* These are the soft objects in the everyday life, mostly with relatively simple shapes. Humans can roughly feel whether they are 'very soft', or 'soft', or 'hard'.

We selected some samples in Group 1, Group 2, and Group 4 as the training set, and tested the model's prediction on Group 1, 3, 4 and 5. The data in the training set is the data collected by human testers; in the test set, some data is collected by human testers, while data is collected by a robot.

3.5.3 Experiment

After collecting the dataset of the Fingertip GelSight sensor contacting different objects, either manually or with a robot, we train the neural network with part of the manually conducted data, and experimented on multiple experimental cases to test the neural network's performance.



(a) Seen shapes with unseen hardness

(b) Unseen shapes

(c) Data collected by robot

Figure 3-16: Hardness estimation result with the deep neural network, when the test objects are of basic shapes.

Table 3.4: Hardness estimation results on basic shapes

	number of videos	R^2	RMSE
Trained shape, novel hardness	1398	0.9564	5.18
Novel shapes	73	0.7868	11.05
Trained samples, robot gripper	683	0.8524	10.28

Basic Shapes

The first experiment aims to test whether the network model could generalize to unseen hardness values. The stimuli objects in the test set are of the same shapes as the objects in the training set (Group 1 in Figure 3-15), but with different hardness levels. The second experiment is about target objects of similar shapes in the training set (spheres and cylinders). The third experiment is about different contact mode, that the test set is conducted by the robot, that the velocity and trajectory are totally different. The output result of the neural network is shown in Figure 3-16 and Table 3.4.

Complicated shapes

In this challenging experiment, we test the deep learning model with target objects that are of complicated and unseen geometries. Examples are in Group 3 and 4 in Figure 3-15. For the test data in Group 3, that the objects are of relatively simple shapes, the R^2 of the estimation is 0.57, and RMSE is 19.3. For the data in Group 4, that the objects are of complicated shapes, the R^2 decreases to 0.39, and the RMSE

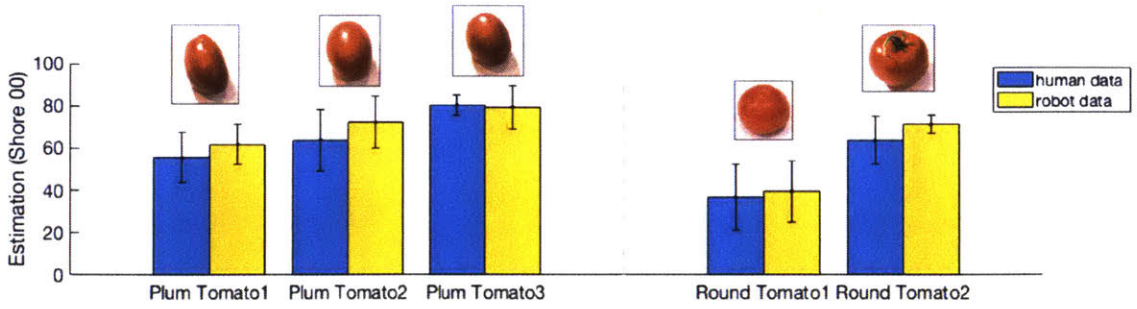


Figure 3-17: Hardness estimation results on tomatoes using the neural network. Data collected by both human and robot for multiple times, on different parts. The display order of the tomatoes is based on human ranking of the tomato hardness.

is 18.2.

It can be inferred that the large estimation error is mostly caused by the confusion of the novel geometry. For example, in most of the large-error cases, the objects have sharp curvatures or dense textures on the surface, which is not included in the training set, and the neural network tends to estimate a much larger hardness value than the ground truth. But for the objects whose shapes resemble the ones in the training set, the neural net can well estimate their hardness.

Natural objects

This experiment aims at testing the generalization of the hardness estimation on natural objects. The ground truth of the objects is unable to be tested, so that we refer to human's estimation for a rough comparison. We compare several plum tomatoes and round tomatoes with different ripeness, and the estimation result is shown in Figure 3-17. The contact is either conducted manually or by a robot, for multiple times and at different locations. In average, each tomato has data collected for 18 times. So, some of the variances of the hardness estimation are caused by the uneven distribution of the tomato. But in general, the output of the neural network matches the order of human estimation. We also compare the hardness estimation of different candies, elastomer tubes, and some random natural objects, and the result is shown in Figure 3-18. The contact is conducted manually for 5 times on each object.

The results indicate that, for the natural objects with simple geometry and smooth

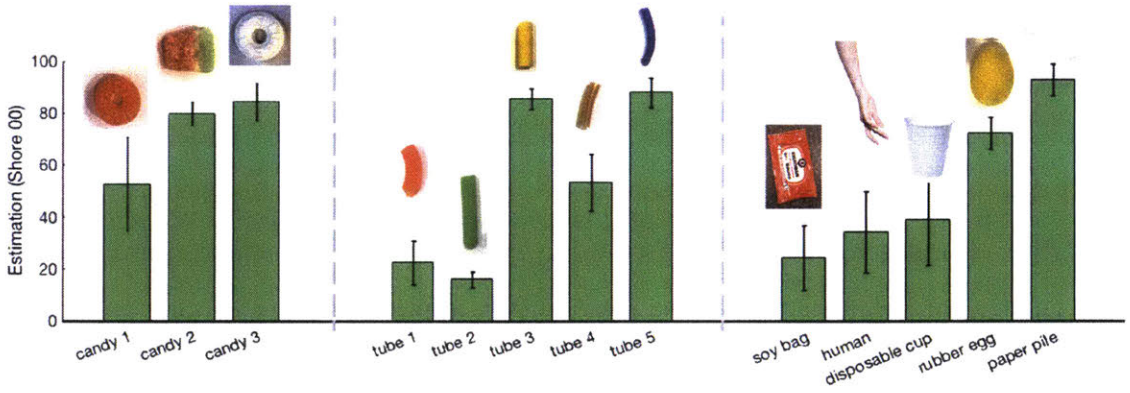


Figure 3-18: Hardness estimations on natural objects using the neural network. In each group, the display order of the objects in each group is based on human perception from soft to hard.

surfaces, the neural network can well estimate their hardness level. The estimation can be used to differentiate ripeness levels of some fruits, like tomatoes. Rigid objects will be estimated with a number larger than 80. However, similar to the previous experiment on the complicated shaped objects, when the shapes are sharp of textured, the network is likely to make a larger estimation hardness value. A typical example is for the candies: the hardness of candy 2, which is a deformable gummy candy, is confusing to the neural network because of the candy grains on the surface, which make a rich surface texture. A similar case is the tube 1 and tube 3 in Figure 3-18, while other tubes in the test set have much smoother surface.

Discussion

The deep learning method provides a possible way to estimate the hardness of arbitrary objects, whose shapes vary and are unknown. The method is relatively easy to apply, without long-time trials on building and comparing the analytical models. However, the deep learning method largely depends on the training dataset. In other words, the current network model is good at remembering the seen examples, but is not good at generalizing the rule to a broader set of samples. The failure in generalization mostly exists in the case of the shape of the objects – when the object shapes or texture types are different from the samples in the training set, the neural network could provide a very wrong prediction.

A common approach to solve this is increasing the training dataset, and adding more shapes that are representative of more general cases of the objects. In practice, this approach may be costly regarding materials, human effort and time to make the physical samples. Another possible solution to this is to build a simulation platform and produce simulated data for the training set. The idea of using simulation to enlarge the training dataset with less cost is widely applied in other robotic research.

3.6 Conclusion

In this chapter, I introduce the research of using a high-resolution tactile sensor to estimate a basic physical property of the objects – the hardness. Particularly, my focus is to make the method robust to different robot scenarios, i.e., the method should work for objects with arbitrary shapes and non-standard contact conditions while the velocities and trajectories cannot be measured. I first showed that by intuition, the material property of hardness, can be indicated by the deformation during the contact, while the measured geometry and marker motion by GelSight is correlated to the deformation. The correlation is general for different object shapes and contact trajectories, and the challenge is building a numerical model the correlation between the GelSight signal and the hardness property.

In the chapter, we propose 2 models for measuring the correlation: one is using a statistical model, one is using deep learning model. The statistical model directly describes the numerical relationship between the features of the GelSight images, which includes the intensity change in the image and the motion field of the markers, and the target samples' hardness. The method can get a relatively precise measurement of the hardness, with a small training set, but is constrained to target objects that are of spherical shapes with known radii. For arbitrary shapes, it is hard to build a direct numerical model for the correlation. For this case, we propose a neural network model for objects with unknown and arbitrary shapes. The network contains a CNN part for the spatial information in the GelSight images, and an LSTM to model the temporal change in the sequence from the contact. We train the neural network

with a training set of GelSight sensor contacting silicone samples of both standard shapes, like cylinders and spheres, and some complicated shapes. The experiments show that the neural network can predict the hardness of the objects with either unseen hardness levels or contact modes.

Chapter 4

Closed-loop tactile exploration on common clothing

The goal of this chapter is to develop a robot system that can autonomously explore common objects through touch, and have a more comprehensive understanding of the objects. In other words, that means learning a broad set of properties about the objects. We take clothing in common life as the target object category, which is an important part of human life. We can easily evaluate an article of clothing largely according to its material properties, such as thick or thin, fuzzy or smooth, stretchable or not, etc. The understanding of the clothes' properties helps us to better manage, maintain and wash the clothing. If a robot is to assist humans in daily life, understanding those properties will enable it to better understand human life, and assist with daily housework such as laundry sorting, clothing maintenance and organizing, or choosing clothes.

However, exploring the extensive properties of common clothing remains a challenge for robots. The materials for common clothing are similar, but well distinguishable to humans. With the help of high-resolution tactile sensing, the robot could learn more about the material properties and feel the subtle difference between different materials.

The previous chapter introduces the principle and method of using high-resolution tactile sensing to estimate object hardness, and the target objects are mostly artificial

silicone samples. In comparison, the challenges of this work are:

1. Exploring a broad set of different properties from tactile reading
2. Target objects are common objects, and in the state of the natural environment
3. The robot exploring the objects by itself

The system in this chapter addresses the three challenges. We use CNN to recognize a set of pre-labeled properties of the clothing from the GelSight data, and another CNN on the external camera image that guides the robot to squeeze on some specific points on the clothing in order to collect tactile data. We also collect a dataset of 153 items of different common clothing for the experiment, and the results show that the system can generalize the exploration of the unseen clothing as well.

The content of this chapter is published in [71].

4.1 Background

The robotics community has been interested in clothing related topics for years, especially for the home assistant robotic tasks. The major focus has been clothing manipulation and recognition/classification. Researches on clothing manipulation are mostly about grasping, folding and unfolding. On the clothing recognition or classification tasks, most of the research uses vision as sensory input, and classifies the clothing according to their rough types, such as pants, t-shirts, coats, etc. [63, 39, 18] introduced methods for clothing classification by matching the 2D or 3D shape of the clothing to the clothing dataset. Sun et al. [56] proposed a method to recognize clothing type from stereo vision, where they applied more local features, such as the clothing's wrinkle shapes and textures.

On multi-modal clothing perception, Kampouris et al. [28] proposed a robotic system to classify clothes' general types and materials. They used an RGBD camera to capture the global shape, a photometric stereo sensor to record surface texture, and a fingertip tactile sensor to measure the dynamic force when rubbing the clothing.

They showed that the multi-modal input, especially the texture perception from the photometric stereo sensors, largely improve the precision of the material recognition. However, recognizing fine-grained clothing properties of common clothing remains a challenge.

4.2 Dataset

This project aims at developing a robotic system that can autonomously perceive clothes and classify them according to material properties. We divide the aim into 2 parts: one is planning a path for the robot to collect the tactile data, and the other one is recognizing the corresponding properties from the tactile data. The robot motion is guided by an external Kinect sensor. So, we collect both the GelSight image sequences and the gripping points on the Kinect depth images. The GelSight images help the robot to recognize the clothing properties, and the depth images and the exploration results help the robot to learn whether a gripping position is likely to generate good tactile data.

4.2.1 Clothing dataset

We collect a dataset of 153 pieces of common clothing with wide varieties. Since we wish the system can be generalized to the real-world environment, the clothing in the dataset should well represent the commonly seen clothes in people's everyday life. The dataset includes both new and second-hand clothes, and differ in sizes, materials and types. A small number of other fabric products, such as scarfs, handkerchiefs and towels. Some examples of the clothing are shown in Figure 4-1.

When a robot touches different clothes with the GelSight sensor, it obtains various tactile signals. Figure 4-2 shows some examples of the GelSight data when touching different kinds of the clothing. The textile textures of the fabric materials are clearly shown in the data, and the overall shapes and the foldings are also informative of the clothing properties as well.

As a ground truth for the clothing understanding, we select 11 clothing properties

Figure 4-1: Examples of the clothes in the dataset. The dataset contains 153 items of clothes, ranging widely in materials, sizes, and types.

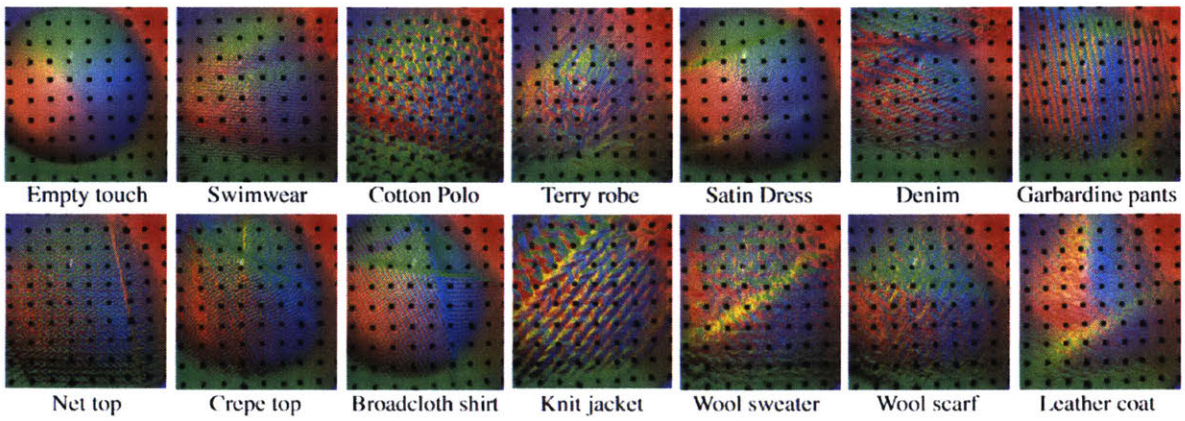


Figure 4-2: Examples of GelSight images when the robot squeezes on clothes (color rescaled for display purpose). Different clothes make different textures on GelSight images, as well as different overall shapes and folding shapes. The top left example is the image when there is no clothing in the gripper.

that humans use to describe clothes, and ask human testers to label the clothes on the properties. Those properties include 8 physical properties of the material, such as thickness, smoothness, fuzziness, and 3 semantic properties of the textile type, the wash method, and wearing season. The semantic properties can guide robots to better handle the clothing in the homes, such as sorting the clothing for different washing baskets or storing drawers. The labels and examples of the classes are shown in Table 4.1.

The properties are in classes, either binary classes (yes or no), or multiple classes. Given that it is hard, and unnecessary to measure the exact value of the material properties, we ask human testers to label them into different levels. For example, for

Table 4.1: Clothing property labels

Thickness (5)	Smoothness (5)
0 - very thin (*crepe dress*) 1 - thin (*T-shirt*) 2 - thick(*sweater*) 3 - very thick (*woolen coat*) 4 - extra thick (*down coat*)	0 - very smooth (*satin*) 1 - smooth (*dress shirt*) 2 - normal (*sweater*) 3 - not smooth (*fleece*) 4 - rough (*woven polo*)
Fuzziness (4)	**Season (4)**
0 - not fuzzy (*dress shirt*) 1 - a little fuzzy (*dress shirt*) 2 - a lot fuzzy (*terry robe*)	0 - all season (*satin pajama*) 1 - summer (*crepe top*) 2 - spring/fall (*denim pants*) 3 - winter (*cable sweater*)
Textile type (20)	**Washing method (6)**
cotton; satin; polyester; denim; garbardine; broad cloth; parka; leather; crepe; corduroy; velvet; flannel; fleece; hairy; wool; knit; net; suit; woven; other	machine wash warm; machine wash cold; machine wash cold with gentle cycles; machine wash cold, gentle cycles, no tumble dry; hand wash; dry clean
Labels with binary classes: Softness, stretchiness, durability, woolen, wind-proof	

the thickness of the clothing, they are divided into 5 classes: 0 for very thin materials like crepe, 1 for common single layer materials like T-shirt, 2 for thick materials like sweaters, 3 for thicker ones like coats, and 4 for extra thick ones like down coat.

4.2.2 Robot setup

The robotic hardware system is shown in Figure 4-3(a), and it consists of four components: a robot arm, a robot gripper, a GelSight tactile sensor, and a Kinect 2 sensor that takes RGBD images of the entire table area. The arm is a 6 DOF UR5 collaborative robot arm from Universal Robotics, with a reach radius of 850mm and payload of 5kg. We use the MoveIt! library for the motion planning. The parallel robotic gripper is a WSG 50 gripper from Weiss Robotics, with a stroke of 110mm, and a rough force reading from the current. We mount GelSight on the gripper as one finger, and the other finger is 3D printed with a curved surface, which helps GelSight get in full contact with the clothes. The GelSight sensor we used is the revised Fingertip GelSight sensor [13], with a soft and dome-shaped surface for sensing, and

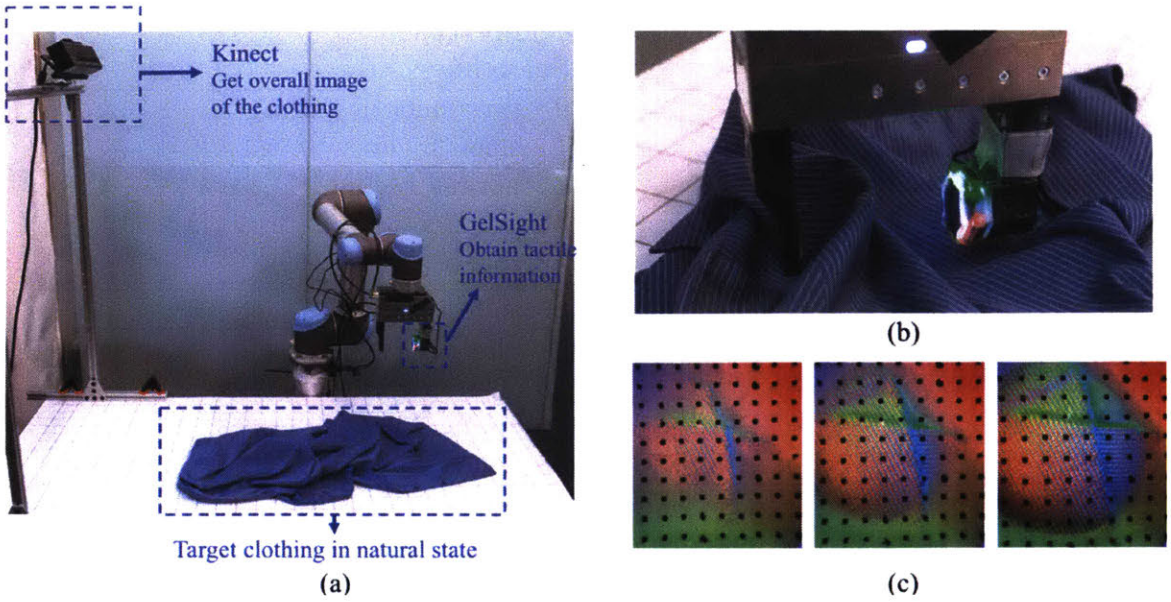


Figure 4-3: (a): The robot system that collects tactile data on clothing. (b): The gripper with a GelSight sensor mounted is gripping the clothing. In the exploration, the gripper squeezes on the clothing foldings to collect a series of GelSight data. (c) Tactile images from GelSight when gripping a piece of clothing with a increasing force.

a sensing range of 18.6mm×14.0mm, a spatial resolution of 30 microns for geometry sensing. The elastomer on the sensor surface is about 5 in Shore A scale, and the peak thickness is about 2.5mm. The sensor collects data at a frequency of 30Hz. The external RGBD camera we used is a Kinect 2 sensor, which has been calibrated and connects to ROS system via IAI Kinect2 [62] toolkit. It is mounted on a fixed supporting frame which is 106mm above the working table and a tilt angle of 23.5°, so that the sensor is able to capture a tilted top view of the clothes.

A challenge in this project is that, since the robot needs to frequently lift the clothing, which could be heavy, the coatings on the GelSight elastomer wear off after a while. We have to change the elastomer. However, since the sensor is manually fabricated, each elastomer piece is similar but unique in the shape and the marker patterns. As a result, the GelSight images vary for each elastomer piece, and it may cause confusion in the property recognition. As a possible solution, we try to balance the collected data on each elastomer piece during the network training. The intention

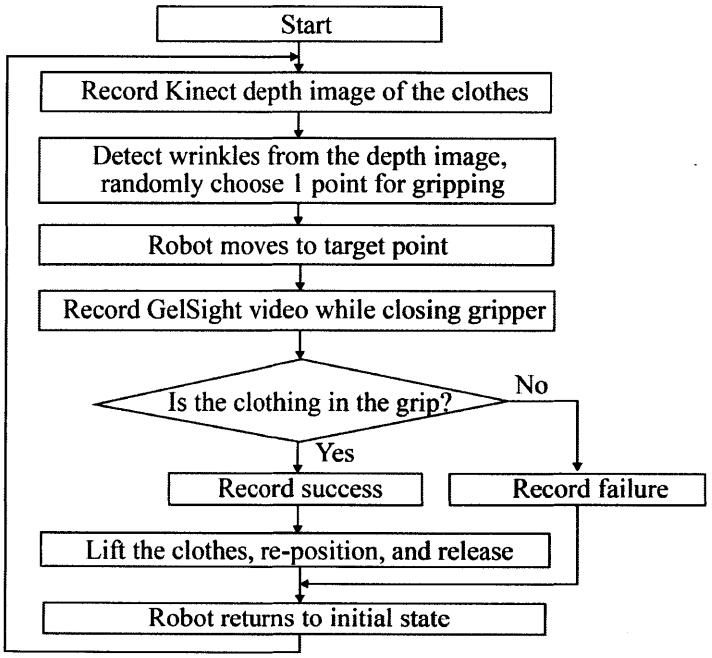


Figure 4-4: The flow chart of the autonomous data collection process.

is to make the property recognition system robust regardless of the gel pad.

4.2.3 Data collection

The training data is autonomously collected by the robot. The flow chart of the process is shown in Figure 4-4. At the start of each iteration, the Kinect takes a depth image of the clothing on the table, and chooses a possible gripping location on the image. Then the robot will move to the target location, and grip on the clothing slowly, while we take a sequence of tactile images with GelSight. If the robot successfully gripped the clothing, it then would lift the clothing and moves it to a random position, then releases it, so that the clothing will be re-positioned. In this procedure, we also record the Kinect image and the gripping location, and record whether the exploration of the given location is successful: which means the robot has gripped the clothing, and the tactile data is explicit. In total, we collect 6616 iterations of the data, with 3762 GelSight sequences valid for training the property classifiers. In the other cases, the robot could not touch the clothing at the position. The data is also recorded to learn what is a good or bad contact location.

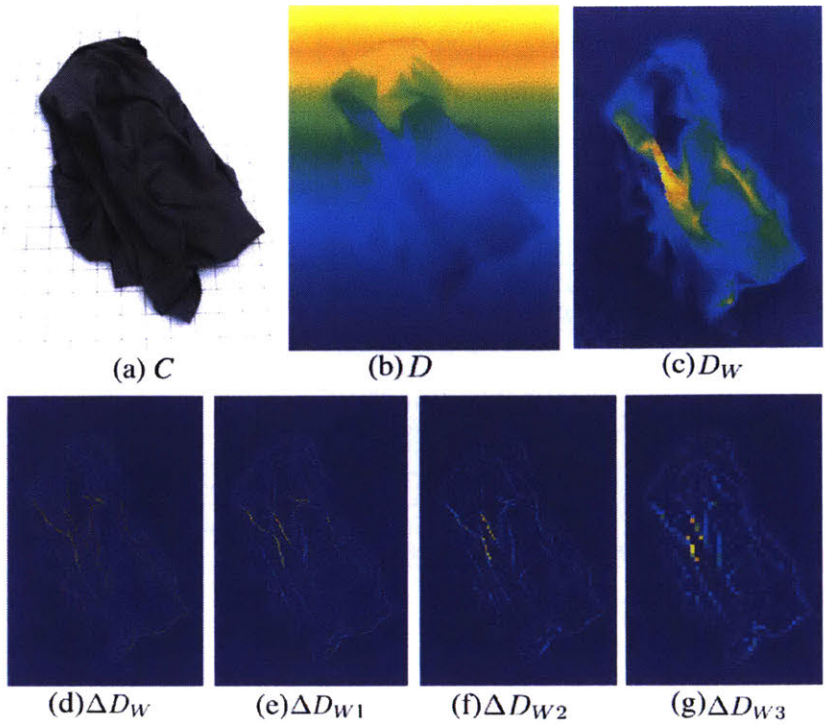


Figure 4-5: Demonstration of finding winkles on the clothes. (a) The RGB image from Kinect. (b) The depth image D from Kinect. (c) D_W: the depth image in the world coordinate, using the table as $x-y$ plane. (d) ΔD_W: the Laplacian operated D_W, where the borders are picked as high-value points. (e)-(g): the Laplacian operated D_W on different pyramid levels. (The color in the figures is re-scaled for display purposes.)

Choosing gripping positions from Kinect images

The robot is most likely to collect good tactile data when it grips on the wrinkles on the clothes. The wrinkles are higher than the surrounding area, which will be captured by Kinect's depth images D. We firstly transfer the depth map into the world frame, thus obtain the depth map D_W using

$$D_W = T_{K2W} \cdot K^{-1} \cdot D \tag{4.1}$$

where K is the camera matrix that expanded to 4×4 dimension, and T_{K2W} is the 4×4 transformation matrix from the Kinect frame to world frame.

We set the $x-y$ plane in the world frame as the table, so that the 'depth' value of D_W, which is represented as z, corresponds to the real height of the clothing on

the table. An example of the transformed D_W is shown in Figure 4-5(c). The edges of D_W, which could be easily picked by Laplacian operation, show the wrinkles on the clothing. We apply the pyramid method to down-sample the image to 3 different levels, therefore the high-derivative areas on different levels represent the wrinkles of different widths. From all the high-derivative points in the 3 levels, we randomly choose 1 point as the target gripping position.

Before gripping, the gripper should rotate to the angle perpendicular to the wrinkle. The rotation of the gripper is decided by the direction of the folding, which can be derived by:

$$Dir(x,y) = \arctan(\frac{\partial D_W(x,y)}{\partial y} / \frac{\partial D_W(x,y)}{\partial x}) \tag{4.2}$$

Gripping on the wrinkles

Once the target point on the wrinkle is selected, the robot will move about the point, with the gripper in a perpendicular direction, and then descend to the position below the wrinkle to grip the clothing, with a low speed of 5mm/s. Due to the simplified setting in this project, the clothing in the 3D space could be simplified as a 2D motion problem: the clothing is placed on the flat surface, and the robot will only contact the wrinkles from the top-down direction, on the $x-y$ plane that is parallel to the table. The low speed is for the safety concern and aims at collecting more touch data during the process. The current sensor uses a USB webcam as the vision acquisition device, which obtains images in a low frequency. Thus, a slow speed of the gripper helps the robot to obtain more frames during one contact. The gripper stops closure when the motor current reaches a threshold, which indicates a large impedance force. The GelSight records videos during the closure. Typically the GelSight records 10 to 25 frames for one gripping iteration.

After the gripping, we judge whether the contact is valid using GelSight images. If the GelSight image shows no contact with the clothing, we mark this tactile data invalid, and mark the gripping location as a failure case.

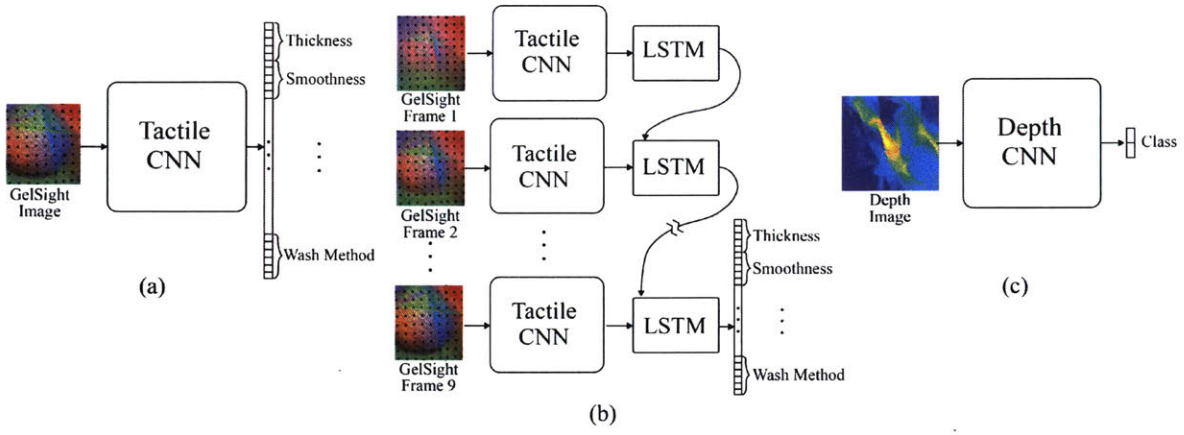


Figure 4-6: (a) The multi-label classification CNN for recognizing different properties from a single GelSight image. (b) The neural network for recognizing different properties from GelSight video, where we choose 9 frames from the video as the input. (c) The neural network for evaluating points on the clothing: whether it would generate effective tactile data.

4.3 Method

In this project, we apply two independent neural networks for 1) selecting a point on the clothing for the robot to explore, and 2) estimating the properties of the clothing from the collected tactile data.

4.3.1 Networks for property perception

To perceive the properties of the clothes, we use a CNN for the multi-label classification of the GelSight images. The labels correspond to the clothing properties, and they are independently trained. We experiment on two networks: one takes a single GelSight image as the input (Figure 4-6(a)), and the other one takes in multiple frames (Figure 4-6(b)). The CNN for GelSight images are VGG19 [51], which is originally designed for object recognition for general images, and pre-trained on the image dataset ImageNet [11].

For the network with a single input frame, we choose the GelSight image when the contact force is the maximum, and use a single CNN to classify the image. For recognizing the multiple properties, we train the same CNN with classification on

multiple labels, which correspond to the clothing properties. The architecture is shown in Figure 4-6(a).

Additional to learning the properties from the single GelSight image, we also try to learn the properties from the GelSight image sequence. The sequence includes a set of images when the sensor squeezes the clothing with increasing forces, thus the frames record the surface shapes and textures under different forces. The image sequences are more informative than the single images. To train on the image sequence, we use the structure connecting CNN and a long short-term memory unit (LSTM) [12] with a hidden state of 2048 dimensions, as shown in Figure 4-6(b). We use the features from the second last layer fc6 from VGG16 as the input of LSTM.

The image sequence contains 9 frames, with an equal time stamp interval until reaching the frame of max contact. We choose the number of 9 as a balance of low computational cost and the sum of information. Since the gripper closes slowly and evenly when collecting the data, the gripper's opening width between the frames is equal. As a result, some of the thick clothes would deform largely in the squeezing process, so that the selected sequence starts after the contact; while when gripping thin clothes, the maximum contact point is easily reached, and the selected sequence starts with several blank images.

4.3.2 Networks for gripping point selection

In the data collection part, the robot selects a set of wrinkle positions on the clothing, but not all the wrinkle points are good positions for touch exploration, and thus some data collection fails. So, we train a CNN (based on VGG16 [51] architecture) to learn whether a location, or the selected wrinkle point, is likely to generate good tactile data. The network architecture is shown in Figure 4-6(c). The input data is a cropped version of D_W, the depth image in the world frame, and the output is a binary class on whether the image represents a potentially successful gripping. To indicate the gripping location in the depth image, we crop the depth image to make the gripping location the center of the new image, and the window size is 11cm×11cm. Using D_W also makes the network robust to the exact experimental setting. In other words, in

another experimental setting, where the Kinect position differs from the current one, we can still derive the D_W with Equation 4.1 but a different T_{K2W}, and feed the D_W into the same neural network.

4.3.3 Offline training of the neural networks

We divided the robot exploration data on the 153 items of the clothing into 3 sets: the training set, the validation set, and the test set. The training set and validation set make of data from 123 items of clothing, and the testing set contains data from the rest 30 items. For the 123 items, we randomly choose data from 85% of collecting iterations as the training set, and 15% of the data as the validation set. The division of clothes for training and testing is manually done ahead of the network training, with the standard that the clothes in the test set should be a comprehensive representation of the entire dataset.

In all the exploration iterations, we consider 2 situations that the exploration is 'failed': 1) the gripper does not contact the clothing, which can be detected automatically from GelSight data; 2) the contact is not good, that collected GelSight images is not clear. Those cases are manually labeled. We train the tactile CNNs with only the data from 'successful' exploration. When training the Depth CNN, the iterations that are considered 'successful' are made class 1.

We train the networks using stochastic gradient descent as the optimizer. The weights of the Depth CNN (Figure 4-6(c)) and single GelSight image (Figure 4-6(a)) is pre-trained on ImageNet [11], and the CNN for the multi GelSight image input (Figure 4-6(b)) is initialized with the weights of Figure 4-6(a). For the video network, we jointly train the CNN and LSTM for 500 epochs, at a dropout rate of 0.5. For training the network for GelSight images, we apply data augmentation to improve the performance of the network, by adding random values to the image intensity in the training. When training with the image sequence, we choose the input sequence slightly differently on the time stamp.

4.3.4 Online robot test with re-trials

We run the robot experiment online with the two networks: at the start of the exploration, the robot generates a set of candidate exploration locations from the depth image, and uses the depth CNN to select the best one. After collecting tactile data by gripping the clothing at the selected location, we use the tactile CNN to estimate the clothing properties. At the same time, the robot evaluates whether the collected tactile data is good, by analyzing the output classification probability of the tactile network. If the probability is low, it is likely the tactile data is ambiguous and the CNN is not confident about the result. In this case, the robot will explore the clothing again, until a good data point is collected. In the practice, we choose the property of washing method and the probability threshold of 0.75.

4.4 Experiment

We conduct both offline and online experiments. For the offline experiments, we test the system on the dataset that is collected in Section 4.2.3.

4.4.1 Property perception

In the experiment of property perception, we test the neural networks' performance on the offline data.

We use 3762 GelSight videos from the 153 clothing items, and classify the tactile images according to the 11 property labels. The training set includes 2607 videos, the validation set includes 400 videos from the same clothes, and the test set includes 742 videos from novel clothes. We try the networks with either a single image as input, or multiple images from a video as input. The results are shown in Table 4.2.

The results show that for both seen and novel clothes, the networks can predict the properties with a precision much better than chance. Specifically, the precision on seen clothes is very high. The network with video input makes slightly better results, especially on the more complicated tasks. However, the precision gap between the

Table 4.2: Result of property perception on seen and novel clothes

	Chance	Seen clothes		Novel clothes	
		Image	Video	Image	Video
Thickness	0.2	0.89	0.90	0.67	0.69
Smoothness	0.2	0.92	0.93	0.76	0.77
Fuzziness	0.25	0.96	0.96	0.76	0.76
Softness	0.5	0.95	0.95	0.72	0.76
Stretchiness	0.5	0.98	0.98	0.80	0.81
Durability	0.5	0.97	0.98	0.95	0.97
Woolen	0.5	0.98	0.98	0.90	0.89
Wind-proof	0.5	0.96	0.96	0.87	0.89
Season	0.25	0.89	0.90	0.61	0.63
Textile type	0.05	0.85	0.89	0.44	0.48
Wash method	0.17	0.87	0.92	0.53	0.56

validation set and test set indicates the model overfits to the training set. We suppose the major reasons for the overfit are:

- The dataset size is limited. Although the dataset has a wide variety of clothing types, the number of the clothes in each refined category is small (2 to 5).
- We used 5 GelSight sensors in data collection, and the difference in the sensor system causes confusion to the network.
- The CNNs are designed for visual images, which is not the optimum for the GelSight images. For example, the networks tend to highly related to the textures from the tactile images, as they are more obvious features, and put small weights to the general shapes that are more directly related to the physical properties.
- Some properties, are not only related to the materials but also the occasions of the clothing. For example, satin is mostly used for summer clothing, but a satin pajama, which feels exactly the same, is worn for all seasons. The touch sensing can only provide local information about the clothing.

We also experiment with other CNN architectures for the multi-label classification task, including VGG16 and AlexNet [31], but the results are not satisfactory. VGG19

performs relatively better. We suppose for the given task of tactile image classification, AlexNet and VGG16 are not deep enough to extract all the useful features from the limited dataset. At the same time, the complicated architecture of VGG19 also makes it more likely to overfit on the training set.

Unfortunately, the neural network trained on videos (Figure 4-6(b)) does not make a significantly better performance, which was expected. The possible reason is that the networks overfit on the textures of the clothing, and the training set is not large enough to train the neural networks to learn the information from the dynamic change of the GelSight images.

4.4.2 Exploring planning

We experiment on picking effective gripping locations from the Kinect depth image, using the Kinect images from the 6616 exploration iterations. The images are also divided into the training set, the validation set (on the same clothes), and test set (on unseen clothes). On both the validation and test sets, the output of the neural network has a success rate of 0.73 (chance is 0.5). The result indicates the identification of the clothing item has limited influence on the result of gripping location selection. In the training process, the network quickly reaches the point of best performance and starts to overfit. For achieving better results for exploration planning, we plan to develop a more robust grasping system, and collect more data or use online training.

4.4.3 Online robot test

In this experiment, the robot runs the exploration autonomously using the depth CNN and tactile CNN. The exploration is similar to the procedure in the data collection part, that firstly a Kinect sensor takes the picture of the clothing on the table, and then the robot chooses a set of points on the wrinkles. Instead of gripping on the robot directly, in the online robot test, we feed the candidate points into the depth CNN for choosing a good gripping position. The robot follows the prediction of depth CNN and gripping on points with a high probability of success, and collect a set of

Table 4.3: Property perception on unseen clothes in online robot test

Properity	Chance	Without Re-trial	With Re-trial	With Re-trial, on Easy Clothes
Thickness	0.2	0.59	0.65	0.72
Smoothness	0.2	0.71	0.74	0.82
Fuzziness	0.25	0.67	0.74	0.82
Softness	0.5	0.60	0.66	0.72
Stretchiness	0.5	0.74	0.81	0.88
Durability	0.5	0.86	0.86	0.91
Woolen	0.5	0.92	0.91	0.93
Wind-proof	0.5	0.83	0.82	0.86
Season	0.25	0.57	0.64	0.71
Textile type	0.05	0.37	0.50	0.59
Wash Method	0.17	0.50	0.60	0.71

tactile images on the clothing. We use the single-image-input neural network, as shown in Figure 4-6(a), to estimate the properties of the target clothing.

We also try the re-trail strategy when the tactile data is 'not good', that the tactile CNN could not effectively estimate the material properties from the obtained tactile data. This is usually caused by an undesirable gripping position. The method is described in Section 4.3.4.

We experiment on the test clothes(30 items), and each clothes is explored 5 times. The result is shown in Table 4.3. Here we compared the result of 'without re-trial' which means the system would not judge the data quality, and 'with re-trial'. Note that the 'without re-trial' results are worse than the results in Table 4.2 because the tactile data here is all the raw data generated by the robot, while Table 4.2 is only from good tactile data. Another reason is that the gel sensor in this experiment is a different one, and not seen in the training set before, so that there is some slight difference in the lighting distribution. The results also showed that with the re-trials, the precision of property classification increases largely. On average, the robot makes 1.71 trials for each exploration, but 77.42% of the clothing is 'easy' for the robot, that it takes less than 2 grasps to get a confident result, and it turns out the property estimation is more precise. The rest clothes are more 'confusing', that the robot needs

to explore them for multiple times, but the properties are still not well recognized.

4.5 Discussion

This work proposes a system for a robot to autonomously explore the comprehensive property set of common clothing through touch. The breakthrough of the work is that it makes a system that can be generalized to a big set of different common clothing, and can be applied to the natural environment. However, the current challenge for the system is that, the property recognition of the clothing overfits the training set, and lack the ability of generalization. Especially, the networks largely rely on the material texture, but not other useful information, like the overall shape of the contact surface. In other words, what the current neural network does is mostly modeling the mapping between the texture and the property classes. For the same reason, the neural network that takes video as the input does not make a much better performance, while the videos certainly contain more information about the clothing.

A common method to address this challenge in deep learning, is to build a much larger dataset. As far as the dataset contains enough variety about the clothing types and contact mode, the trained neural network should be able to recognize the general situations much better. But enlarging the dataset, in this work, requires large cost and effort for collecting clothing and conduct experiments.

Another concern comes from the architecture design of the neural networks. The current network is designed for computer vision, but not exactly for GelSight images. The comparison of the performance of multiple network architectures also shows that the VGG19, which has a much more complicated architecture, makes the best performance. The data does not necessarily need a complicated architecture to recognize the properties, but the network has been inefficient. The situation with the gripping point evaluation from the Kinect images is the same. In a preliminary experiment, we designed a much smaller neural network architecture that has only 3 layers, and tried it on the Kinect images from scratch, and the performance was better than the current one. So, a future direction for improving the performance is to explore more

efficient network architectures for the specific tasks, like understanding the tactile images and explore on the Kinect data.

Another question that is raised in this research is how to more effectively tell whether a tactile data is 'good', and could produce reliable property recognition results. In this work, we apply a hack way, that is to predict whether the data is good or not from the 'confidence' of the network output. This is easy, yet not necessarily the best solution. A possible way to improve on this is to train a specific network to judge 'whether the input data is good', or using online learning to train the two networks recursively, thus to achieve the goal of unsupervised learning.

Considering the real-world scenarios of clothing recognition, humans not only use touch to evaluate clothing or other objects. The rule should be the same for robots: in the future, if we could combine multi-modal sensory input to the property recognition task, such as vision, motion sensing, the robot should be able to better evaluate the target objects.

4.6 Conclusion

In this Chapter, I introduce a collaborative project on recognizing the comprehensive properties of common clothing using autonomous active touches of robots. We divide the task into two parts: planning a gripping motion of the robot from the external depth image captured by Kinect, in order to obtain tactile data through the contact with the clothing; then recognizing the material properties from the high-resolution GelSight images. We use neural networks to get the useful information from the original 2D images, either the depth images or the tactile images. In the robot experiment, we combine the 2 networks to make the robot actively explore the clothing and recognize their properties. And when the property prediction output from the neural network is not of high confidence, we also make the robot to retry on the clothing, which in turn makes the exploration result more reliable.

Chapter 5

Cross-modal material perception with vision and touch

In this chapter, we further focus on the problem of finding a comprehensive description of common objects that related to their intrinsic material properties. The previous chapter focused on making a robot exploring common clothing and find the description of a wide set of properties, that are pre-set by humans. However, those properties do not necessarily well describe the clothing. Humans are good at perceiving and evaluating objects according to their properties, but it is hard for us to explicitly name and value those relevant properties. So, there raises a question, is it also possible for robots to find a latent description of the material properties?

We take the common fabrics as the target object, since they have a rich set of mechanical properties. They could have significant different configurations, while humans' evaluation of them are not influenced. At the same time, humans have perception through different modalities, especially vision and touch. We see the fabrics, and we touch the fabrics, through seeing or touching we get the 'feeling' of the fabrics. We consider the similar motivation here: a robot should be able to recognize the latent properties of the material, and those properties should be consistent, regardless of the perception modality or the exact configuration for one observation.

We apply Convolutional Neural Networks (CNN) to extract an embedding vector

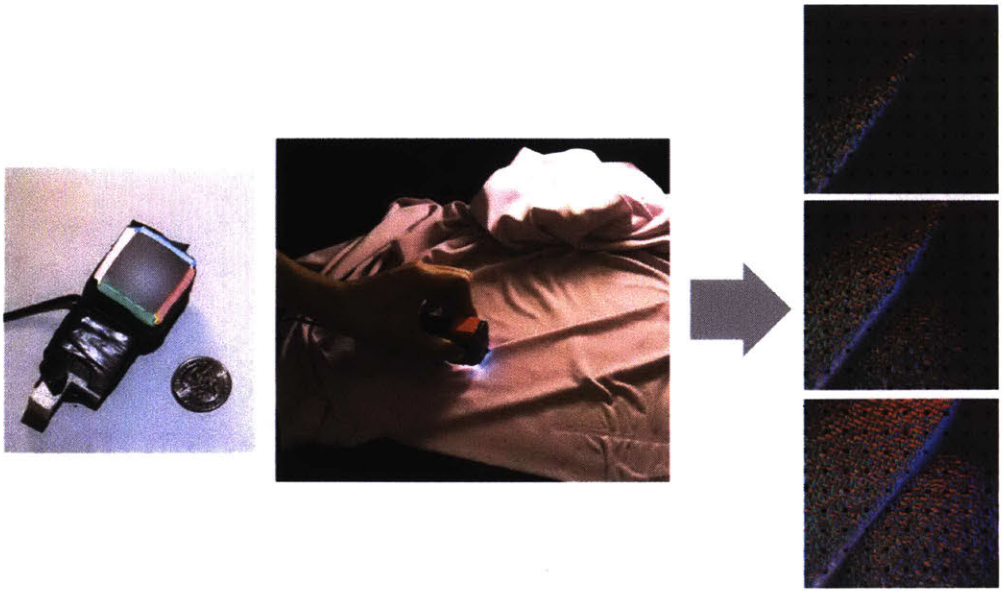


Figure 5-1: A human presses Fingertip GelSight sensor on the fold of a fabric, and gets a sequence of tactile images.

to describe the latent properties, from the high-dimensional observation, either touch or vision. To train the neural network, we jointly train the three networks for the 3 modalities: color vision, depth vision, and touch. The goal is to make the property vector very close to each other when they are from the observation of the same piece of fabric. To demonstrate the networks' performance, we design the task of 'picking the image of the fabrics that match the one you touched', by comparing the embedding vectors from the inputs. The task can be compared to the human perception part, that when we see some objects, we can naturally imagine how it feels like through touch, and vice versa.

The content of this chapter is published in [73].

5.1 Background

5.1.1 Fabric perception

There have been works studying the perception of fabrics. Works like [1, 16] showed humans use visual cues to infer different material properties. Specifically, Xiao et al.

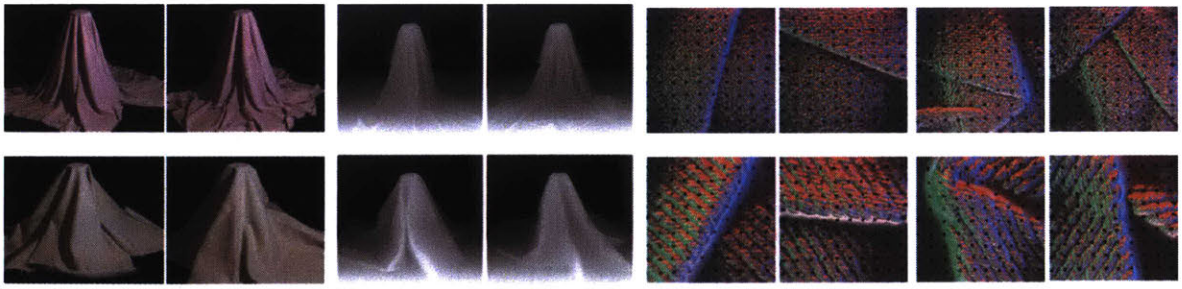


Figure 5-2: Three modalities of the fabric data. For the visual information, the fabrics are draped from a cylinder in natural state; for the tactile information, a human holds the GelSight sensor and presses on folds on the fabric.

[65] studied human perception of fabrics and the influencing factors by using tactile sensing as ground truth to measure visual material perception. They showed that humans made a high matching accuracy, while the color and 3D folds of the fabrics are the most important to the human visual perception.

Researchers in computer vision and graphics have been trying to track or represent fabrics or clothes, but their visual representation is difficult to obtain compared with that of rigid objects', and the uncertainty and complexity of the shapes and motion make the fabrics or clothes more difficult to predict. To track the exact shape of the clothes, White et al. [61] made dense patterns on fabrics or clothes, and used multiple cameras to track their motion thus to reconstruct the 3D shapes of the clothes. Han et al. [20] represented the cloth shape with a 2-layer model: one represents the general shape, and the other one represents fold shapes, which are measured by shape-from-shading methods. Some other researches tried to represent the fabrics by physical parameters, and estimated the parameters from the visual appearance. Baht et al. [3] used a model made of physical properties, including density, bending stiffness, stretch stiffness, damping resistance and friction, to describe and simulate clothes. They estimate the properties by comparing the real clothes' motion video with the simulated videos. Bouman et al [4] measured fabric properties (stiffness and density) directly from the video of fabric motion using hand-crafted features, when the fabrics were hung and exposed to different winds.

5.1.2 Joint neural networks

The joint neural network is the network architecture that joins two or more separate networks for different inputs. Chopra et al. [5] first proposed a Siamese Neural Network (SNN), that learned low dimensional embedding vectors from a single-modal input. The SNN has two identical neural networks with shared weights, and outputs the distance of embedding vectors from the two inputs. In the training, the network uses energy-based contrastive loss [19] to minimize the distance of the embeddings from similar input pairs while making the distance of dissimilar input pairs' embeddings larger than margin. SNN has been applied in face verification [5] and sentence embedding [43].

For single-modality recognition, Chopra et al. [5] proposed Siamese Neural Network (SNN) to learn low dimensional embedding vectors. The SNN has two identical neural networks with shared weights. Each time it takes in a pair of data inputs, and outputs the distance of their embedding vectors. In the training, the network uses energy-based contrastive loss [19] to minimize the distance of the embeddings from similar input pairs while making the distance of dissimilar input pairs' embeddings larger than margin. SNN has been applied in face verification[5] and sentence embedding[43].

In recent years, people have been using joint neural networks for cross-modality learning – mostly two modalities. A traditional method is to extract features from one modality and project the other modality to this feature space. Frome et al. [17] proposed hinge rank loss to transform visual data to text. Li et al. [38] learned the joint embedding by associating generated images to the trained embeddings from shape images of objects. Owens et al. [46] combined CNN and LSTM to predict objects' hitting sound from videos. They extracted the sound features first, and then regress the features from images by neural networks. Their other work [47] presented a CNN that learn visual representation self-supervised by features extracted from ambient sound.

5.2 Method

In this work, we use Convolutional Neural Networks (CNN) to extract an embedding vector from the high-dimensional input of the fabrics. The embedding vector, although could hardly be manually interpreted, is expected to be a comprehensive description of the important material properties of the fabrics, that decide how the fabrics 'look like' and 'feel like'. The input is from either touch image with GelSight, or color image of the fabric, or the depth image of the fabric. In each case, the fabric may have different configuration under the same method of observation. But the embedding vector of the fabrics, which describes the intrinsic properties of the fabrics, should be consistent regardless of the observation bias or input modal. At the same time, the quantitative description of the embedding vector should also be able to estimate the 'similarity' of two fabrics.

Thus, we build joint neural network models to associate visual and tactile information of the fabrics in order to calculate the embedding vector. The input data is of three different modalities: the depth image, the color images, and the tactile images from GelSight. All in the image form. The input data from each modality goes through an independent CNN to form an embedding vector $\mathbf{E}$, as a low-dimension representation of the fabrics. We use the sum of Euclidean distance $D = ||\mathbf{E}_1 - \mathbf{E}_2||$ to measure the differences between two $\mathbf{E}$s, regardless of the input's modality. Ideally, all the input data on the same fabric will make the same $\mathbf{E}$ through the networks, while two fabrics, when they are similar, will have a small distance D between the embedding vectors $\mathbf{E}$, and two very different fabrics will have large D. We trained a joint CNN of the three modalities and compared the performance of different architectures. Figure 5-3 shows the network architecture.

5.2.1 Neural network architectures

We designed and experiment with different neural network architectures in order to obtain the embedding vector.

Cross-modal Net

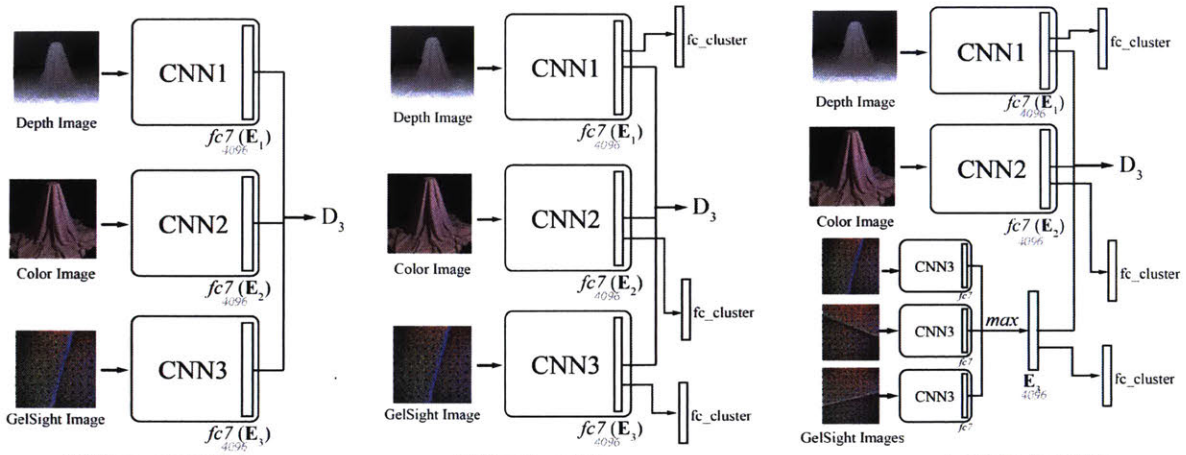


Figure 5-3: The architectures for training the neural network in order to get the embedding vector to describe the fabrics. (a) The Cross-modal Net: data from the three modalities goes through three independent CNNs (AlexNet [31]) in a joint network, and be presented by an embedding **E**, which is the $fc7$ layer of the network. (b) The Auxiliary Network with the subtask of fabric classification. Clusters of the fabrics are made according to human label. (c) The Multi-input Network, that touch embedding is derived from 3 independent GelSight pressing images.

The basic network to join the three modalities is shown in Figure 5-3(a). In this network, the architecture images, color images, and GelSight images go through three separate CNNs in a joint network. The CNN we used in this work is the AlexNet [31], which is pretrained on ImageNet, and we take the $fc7$ in the network as the embedding vector **E** to represent the property set of a fabric.

We use contrastive loss[5] as the objective function. For an input group of depth image X_1, color image X_2 and GelSight image X_3, the embedding vectors coming from the three neural network G_{W1}, G_{W2} and G_{W3} can be denoted as $\mathbf{E}_1 = G_{W1}(X_1)$, $\mathbf{E}_2 = G_{W2}(X_2)$ and $\mathbf{E}_3 = G_{W3}(X_3)$. For each input group, we measure the overall distance between the embedding vectors, denoted as D_3:

$$D_3 = \|\mathbf{E}_1 - \mathbf{E}_2\| + \|\mathbf{E}_2 - \mathbf{E}_3\| + \|\mathbf{E}_3 - \mathbf{E}_1\| \tag{5.1}$$

We make $Y = 0$ if X_1, X_2 and X_3 are sourced from the same fabric, and $Y = 1$ if

they are from different fabrics. The network loss is

$$L(W1, W2, Y, X_1, X_2) = \frac{1}{2}(1 - Y) \times D_3^2 + \frac{1}{2} Y \times \max(0, m - D_3)^2 \qquad (5.2)$$

where $m > 0$ is a margin (we used $m = 2$ in our experiments). Dissimilar pairs contribute to the loss function only if D_3 is smaller than the margin radius m. The existence of dissimilar pairs is meaningful to prevent the D_3 and the loss L being zero by setting G_Ws to a constant.

Auxiliary Net

The auxiliary net is the network architecture that is based on the basic cross-modal net, but with an auxiliary task of fabric classification on the embedding vector **E**, as shown in Figure 5-3(b). The motivation is to make similar fabrics have close embedding vectors by adding some supervision. The classification label of the fabrics is made based on human labeling, as described in Section 5.3.3. Examples of the cluster are shown in Figure 5-4. The three cross-entropy losses of cluster classification are combined with the contrastive loss(5.2) in addition for a total loss.

Multi-input Net

The multi-input network is designed based on the auxiliary net, but the input from touch is of 3 different GelSight images, instead of on only one. The three GelSight images go through the same network G_{W3} respectively, making 3 $fc7$ vectors, and we make the final embedding **E** of the inputs as element-wise maximum of them. The network is shown in Figure 5-3(c). The motivation for this design is that humans are likely to touch an object for multiple times before obtaining a confident perception of it, and similarly, we design the multi-input architecture to exploit more information from the multiple presses. In practice, we find that element-wise maximum makes the training much easier compared with concatenating the three embeddings and also it's symmetric unlike concatenating.

5.2.2 Training and testing

In the training, we use the Adam [30] optimizer and fix learning rate as 0.001 throughout the experiment. Parameters of AlexNet before $fc7$ will be fixed during training. We train the network for 25,000 iterations with a batch size of 128.

In the test, we used the trained CNNs G_{W1}, G_{W2} and G_{W3}. Each input image, either a depth image, color image or GelSight image, go through the corresponding network to produce an embedding $\mathbf{E}$, as a representation of the fabric. For different inputs, either from the same or different modalities, we calculate the $\mathbf{E}$s from the input, and compare the distance D between the two $\mathbf{E}$ to decide the likeliness that the two inputs are from the same fabric.

5.3 Dataset

We collect a dataset for fabric perception that consists of visual images (color and depth), GelSight videos, and human labeling of the properties. The dataset contains 118 fabrics, including the apparel fabrics like broadcloth, polyester, knit, satin; bedding fabrics like terry, fleece; and functional fabrics like burlap, curtain cloth, oilcloth (examples shown in Figure 5-2). About 60% of the fabrics are of single but different colors, others have random color patterns. Each fabric piece is of the approximate size 1m×1m. Some of the fabrics are kindly provided by researchers working on [65] and [4].

5.3.1 Vision data

We drape the fabrics from a cylindrical post(30.7cm height, 6.3cm diameter) in natural states and take both the color images and depth images of them. The color images are taken by a Canon T2i SLR camera, and depth images are taken by a Kinect 2. For each fabric, we take pictures of 10 different drapes. Those drapes make different appearances, but humans could easily tell the similarity between the images.

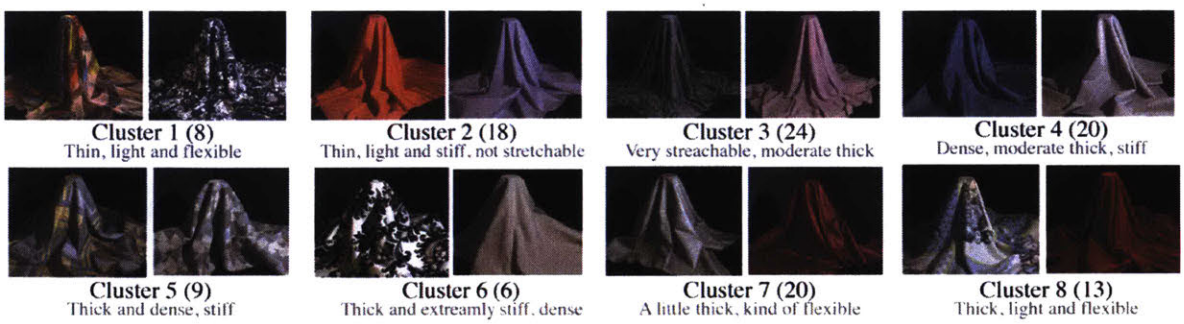


Figure 5-4: Clustering of the fabrics based on human labeling on properties. Numbers in the bracket denote the fabric number in the cluster.

5.3.2 Tactile data

We press the tactile sensor, GelSight, on the fabrics when they lay on a hard flat surface, thus obtaining a sequence of GelSight tactile images for the press process. The sensor we used is the fingertip GelSight device [37]. We collected three forms of tactile data: one is the 'flat data', when the GelSight is pressed on the single-layer of the flat fabrics; another one is 'fold data', when the GelSight is pressed on the fold of the fabrics, as shown in Figure 5-1; and the last one is 'random data', that the GelSight is pressed on some randomly shaped foldings of the fabrics, as demoed in the rightmost columns in Figure 5-2. For each fabric, we collect 10 pressing samples of the flat data and 15 samples of the fold data. Note that in the current stage of this research, we only experiment with the 'flat data' and 'fold data', which have relatively more constraints in the data form.

5.3.3 Attribute labels

We label each fabric with the estimation of the physical parameters that we believe are the most important determining the fabric draping and contact process: *thickness, stiffness, stretchiness* and *density*. The human label, although hardly precise due to the interpersonal difference, offers a reference to the evaluation by the neural network. The thickness and density are measured by a ruler and a scale; stretchiness is roughly estimated at the level of 'non-stretchable', 'stretchable', and 'extremely stretchable';

the stiffness is estimated by humans: we ask 5 human subjects to score the fabric stiffness in the range of 0 to 5 (with the permission of excess for extra stiffness), and take the mean value. Note that the label does not necessarily cover all the true properties that influence the drape, and the values contain human bias, but they can provide a convenient and reasonable reference.

In this work, we cluster the fabrics into 8 clusters by using k-means on the fabrics' physical parameters, as shown in Figure 5-4. For humans, fabrics in the same cluster will have relatively similar properties. We describe the human intuitive description of each cluster in Figure 5-4.

5.4 Experiment

We train the neural networks that extract property-representative embedding vectors from the high-dimension input. To test how well the embedding vector represents the fabrics, we design a following task: given an input from a target fabric, either on touch, color image or depth image modality, and 10 candidate inputs from different fabrics, we ask the neural network to pick the candidate input that is most likely from the same fabric source.

The task can be achieved by comparing the Euler distance between the embedding vectors from the inputs. When the two embedding vectors are close, it is likely the source fabrics are very similar, or are the same piece of fabric. In the experiment, we train the three networks jointly with input from three modalities, and test the match between either 2 different modalities, or a single modality. We report the result of top 1 precision and top 3 precision, which means the correct match ranks top 1 or top 3 among all 10 candidates. Note that the task is even challenging for humans. Considering the large fabric datasets, some fabrics may appear very similar to humans, and humans can hardly make a very high top 1 precision in the test.

We divide the 118 fabrics in the dataset as a training set (100 fabrics) and test set (18 fabrics). The 18 test fabrics are selected evenly from the 8 clusters in Figure 5-4, so they are considered well representative of the entire dataset.

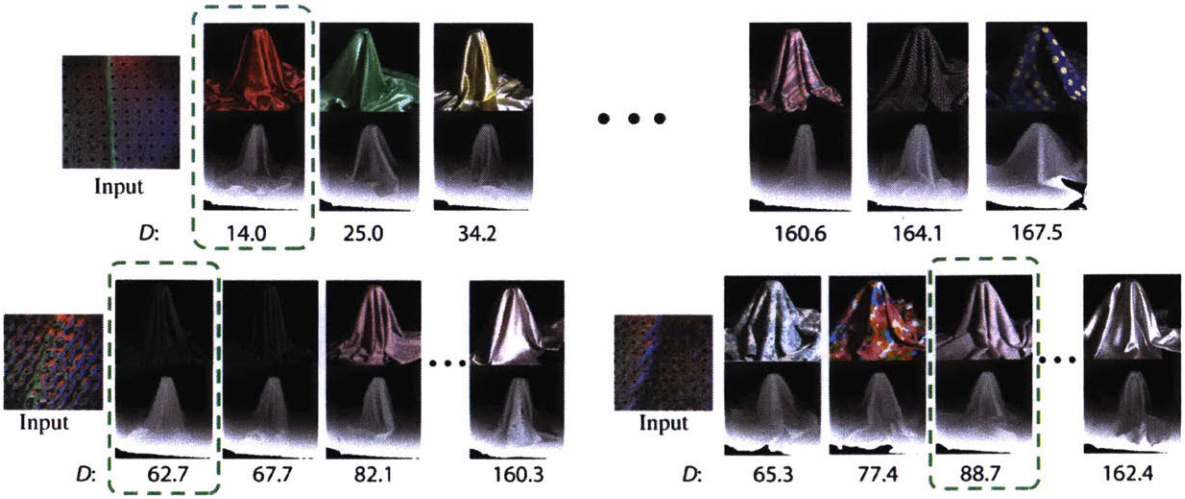


Figure 5-5: Examples of picking the corresponding depth image to the GelSight input, according to the distance D between their embeddings. Trained on the Auxiliary Net. Green frames mark the ground truth.

5.4.1 Inferring touch from vision

The first experiment is picking the depth or color images that best match the GelSight input. The match is according to the D between the **E**s from the given GelSight image and the candidate images. In the experiment, the candidate depth or color images are 10 images from 9 random selected fabrics and the ground-truth fabric from the test set. The network will calculate the embeddings of the input GelSight images and each candidate depth or color images, and compute the distance D between them. The candidate with smallest D is considered most likely to be the correct correspondence. The selecting procedure is shown in Figure 5-5. For each network that is experimented, we test each 15 different GelSight input images on each fabric for 10 times, and calculate the average precisions.

We test the performance of 4 networks: 1. the cross-modal network (Figure 5-3(a)), when the GelSight input is the pressing image on flat fabrics without folds; 2. the cross-modal network, when the GelSight input is one pressing image on the folded fabrics; 3. the auxiliary network (Figure 5-3(b)) that compares depth images and GelSight on single folds, but with the auxiliary task of clustering the embeddings; 4. the auxiliary network that takes three GelSight images as the input (Figure 5-3(c)).

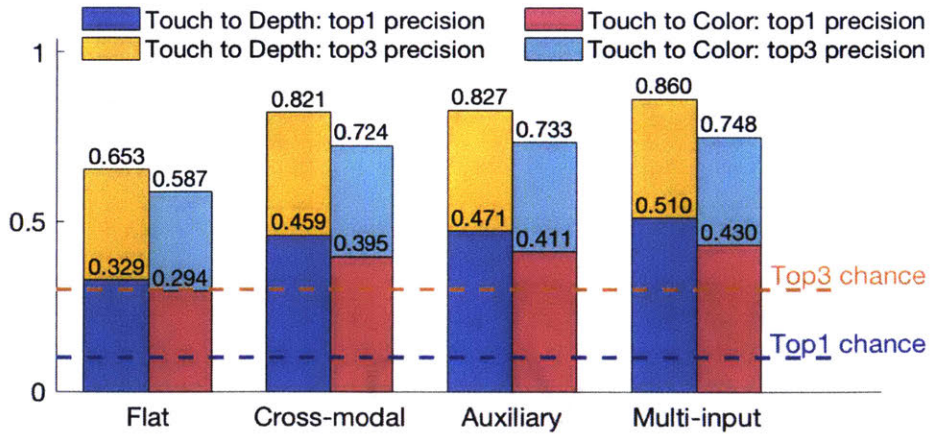


Figure 5-6: Test result: the top 1 and top 3 precision on matching the depth or color image candidates to a given GelSight input, using different network structures or tactile data input. The first row is an example on training set, second row shows examples on the test set.

Model	Flat	Cross-modal	Auxiliary	Multi-input
Depth2Gel	0.3063	0.4292	0.4318	0.4576
Color2Gel	0.2681	0.3742	0.4022	0.4124
Depth2Color	0.4133	0.4329	0.4141	0.4417
Color2Depth	0.4050	0.4240	0.4070	0.4306

Table 5.1: Result on the test set: average top 1 precision on test set for the "pick 1 from 10" experiment of matching 2 modalities.

The results of the top 1 precision and top 3 precision on the test set is shown in Figure 5-6. We also test the precisions of matching other modalities, and the results are shown in Table 5.1. In comparison, the precisions on matching the data from a single modality are much higher, as shown in Table 5.2.

From the results, we can see that all the networks can predict the matching images better than average chance. As for the architectures, the auxiliary net with 3-frame input performs the best, the auxiliary net with 1-frame input places the second, and the basic model with the plain GelSight press comes the last. The match between touch images and depth images is better than the match with color images. At the same time, matching between inputs from the same modality is very easy.

The positive results in the matching experiments show that the neural networks are able to automatically extract the features related to fabric intrinsic properties

Model	Flat	Cross-modal	Auxiliary	Multi-input
Depth2Depth	0.6030	0.6265	0.6224	0.6459
Color2Color	0.7941	0.7831	0.7968	0.8247
Gel2Gel	0.8025	0.7672	0.8090	0.9351

Table 5.2: Result on the test set: average top 1 precision on test set for the "pick 1 from 10" experiment of matching single modality.

from either visual or tactile information. The properties from the three modalities are correlated, so that the networks can match one modal input with the other by comparing the embedding vectors. But in the given dataset with the limited size, the neural networks extract the physical properties better from the depth images than from the color images, because the former has less information and the fabric shape is more directly related to the physical properties. The results also show that, the additional information helps the network to better recognize the materials: the comparison between model 1 and 2 shows that the folds on the fabric reveal more properties; comparison between model 2 and 3 shows on this small dataset, the human label help to improve the network performance; the comparison between model 3 and 4 shows that providing more touch information, the network will extract the relevant information better, and makes the matching more robust.

5.4.2 Data augmentation

To compensate the error caused by the color of the fabrics, which does not influence the intrinsic material properties of the fabrics, we augment the dataset on the color images, by changing the hue and exposure of the images during the training. To be specific, we perform Gamma Correction (range 0.5-2.0) to each image, and change the order of the RGB channels. The matching tests with the color images involved make a better result, as shown in Table 5.3. But the results of other matching tests between GelSight images and depth images do not change.

We also tried other data augmentation on the GelSight images and the depth images, including adding noise to the input, and cropping the images randomly, but the results make little difference.

Model	Cross-modal	Cross-modal (with aug)	Multi-input	Multi-input (with aug)
Gel2Color	0.3954	0.4359	0.4303	0.4937
Color2Gel	0.3742	0.4088	0.4124	0.4264
Depth2Color	0.4329	0.4674	0.4417	0.4924
Color2Depth	0.4240	0.4607	0.4306	0.4624

Table 5.3: Comparison of the top 1 precision before and after data augmentation on the color images.

5.5 Analysis

In this section, we try to find some insight on how well the embedding vector **E** represents the fabrics. Especially, whether **E**s from the same or similar fabrics are closer than those of distinct fabrics.

we continue with the experiment of 'picking the possible depth image given a GelSight image' as an example. To denote the possibility that the two **E**s are sourced from the same fabric, we build a function P to describe the distance between two **E**s in the exponential scale:

$$\mathrm{P}(\mathbf{E}_1, \mathbf{E}_2) = A \exp\left(-c \times D(\mathbf{E}_1, \mathbf{E}_2)^2\right) \tag{5.3}$$

Where c is a positive coefficient (I set it as 8.5×10^{-2}), and A is a coefficient that can be set according to each fabric. For a given input with embedding $\mathbf{E}_{tar}$, and a set of candidates with embedding vectors $\{\mathbf{E}_i\}$, we normalized P by adjusting A for each fabric so that

$$\sum_i \mathrm{P}(\mathbf{E}_{tar}, \mathbf{E}_i) = 1 \tag{5.4}$$

Here we make $\{\mathbf{E}_i\}$ from all the depth images in the candidate fabric set. For each test fabric, we calculate P over all the available GelSight input image and take their average, so that we got a possibility of 'mismatching the touch data from the current fabrics to the other fabrics'. We draw confusion matrices of the mean P between the fabrics in Figure 5-7. In the figure, we re-order the fabrics numbers to put the fabrics adjacent when human subjects consider them similar, so that the bright spots near

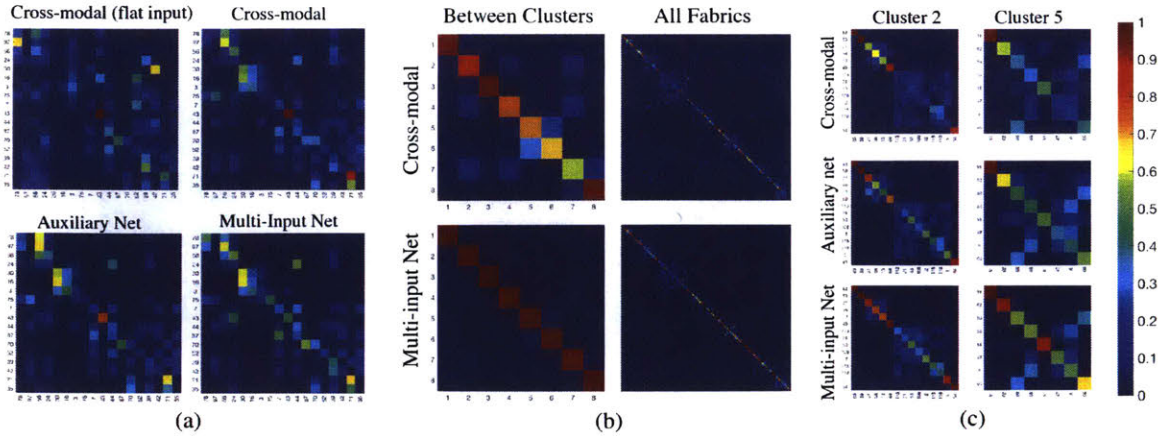


Figure 5-7: Confusion matrices based on the distance between the **E**s between fabrics, on "picking the possible depth image to a given GelSight input". The fabrics are placed in the order according to human subjects, so that similar fabrics are close. (a) Test results for different networks. (b) Training set for the Cross-modal net and Mutli-input Net, either between clusters, or on the individual fabrics. (c) Confusion matrices on fabrics in the training set within Cluster 2 and Cluster 5.

the diagonal line means the neural network gets confused with the fabrics that are likely to confuse humans too.

Figure 5-7(a) shows the confusion matrix on the test dataset, and it indicates that most of the possible confusion occurs between the similar fabrics. The order of the fabrics on the axes is according to the fabric clusters, so that the adjacent two fabrics are likely to be similar according to human labels. So the bright spots near the diagonal line are more likely to represent the confusion when the two fabrics are very similar. Figure 5-8 shows examples of the fabrics that are easily confused with each other, while they also appear similar to human. In general, the Multi-input Net performs the best on the confusion distribution, while the Cross-modal Net with only plain input performs the worst.

Figure 5-7(b) shows the probability in matching the GelSight data and depth image in the training set (100 fabrics). Here we compared the matching probability of all the independent fabrics, and also between different clusters. The figures indicate that both networks well distinguish the fabrics in different clusters. Even the Cross-modal net does well, while it does not know the cluster in the training. But within

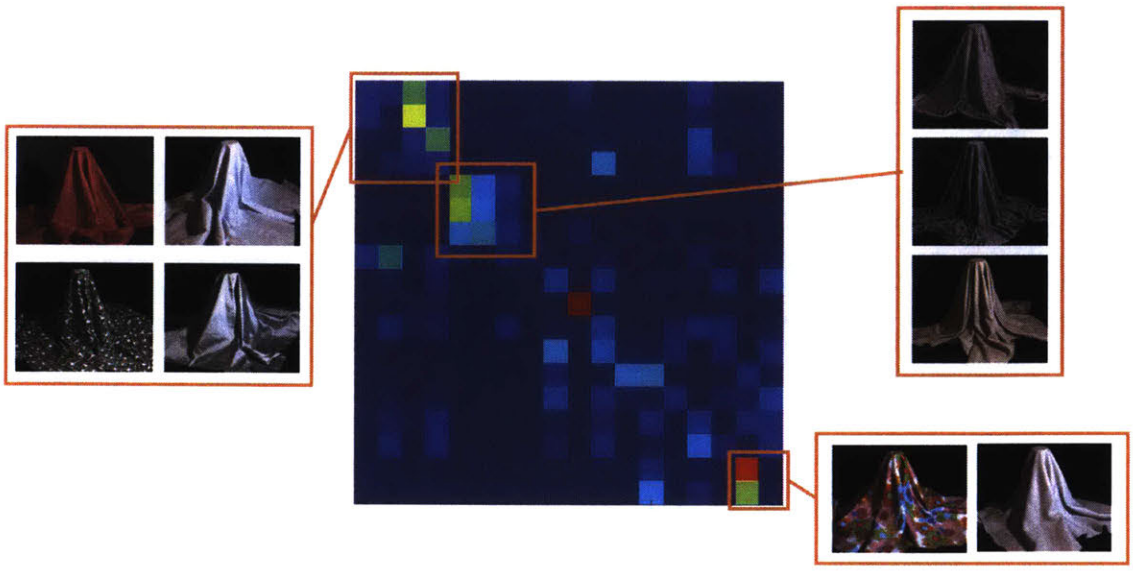


Figure 5-8: The confusion matrix on the test dataset, using the cross-modal network.

the clusters, the network can be confused between fabrics. We also compare the correspondence of the embeddings on the training set (100 fabrics), and the confusion matrices for the Cross-modal Model and Multi-input Net in Figure 5-7(b). The figures indicate that in both cases, the fabrics in the same cluster tend to have more close embeddings than those between clusters, and the majority of confusion comes within the cluster.

Figure 5-7(c) shows the confusion matrices of fabrics within Cluster 2 and 5. Cluster 2 denotes fabrics that are 'thin, light and stiff', and contains many broadcloths. They appear very similar to human; similarly, the Cross-modal Net and Auxiliary Net make their embedding vectors close, and display a blurred area in the bottom left in the matrices. the two upper right fabrics, however, are unique in that they are translucent gauze, so that even human labeling of physical properties suggests they are similar to other fabrics in the cluster, they show special textures or optical characteristics for both GelSight and depth images, and are distinguished from the other fabrics. But for the Multi-input Net, as there is more input information, the network is able to represent the more subtle differences between the fabrics, so that

the confusion matrix concentrated. Cluster 5 contains fabrics that are thick and stiff. Similarly, the Multi-input Net reduced the confusion between different fabrics the best (although not totally), and the embedding vectors would better represent the fabrics.

The results in this section prove that all those factors will improve the network's ability to represent the fabrics: touching the folds instead of the plain fabric; multiple presses that contain less biased information. The clustering information made according to human label also help the network to narrow down the fabric range to represent the properties.

5.6 Comparison of cross-modal learning and single-modal learning

In the research, we also find that training with the cross-modal network boosts the performance. Taking the single-modal match as an example: "picking a depth image of draped fabrics that best matches a given depth image". We can use the same CNN to extract the embedding vector, but trained in different ways. One is using a Siamese Neural Network (SNN) [5] trained on only depth images, and the other is a joint network similar to Figure 5-3(a), but have only two branches for depth images and GelSight images. The two architectures are the same other than they take in different modalities as branches.

In this test, we select 80% of the data on the 100 training fabrics as the training set, and the rest 20% data, as well as the data from 18 test fabrics as the test set. The test results are shown in Table 5.4.

As shown in the results, on this size-limited dataset, the joint model on both touch and depth images have much better performance than single-modal SNN model. We assume this means the extra information from one modality will help the training in the other modality to reduce overfit and find a better local minimum. The situation is similar to human learning: we learn the materials, especially fabrics, with both vision

Model	Seen Fabrics		Novel Fabrics	
	Top1	Top3	Top1	Top3
SNN (only depth)	0.482	0.660	0.554	0.729
Cross-mdl (depth&Gel)	0.608	0.786	0.606	0.786

Table 5.4: Test results on the depth-to-depth match on two networks: a Siamese Neural Network (SNN) [5] trained only on depth images, and a Cross-modal Net trained on depth and GelSight images.

and touch. Our brain automatically combined the two modalities when we explore a fabric, and built a connection between the two. This also helps us better understand the material even when we observe a material with single-modality input.

5.7 Conclusion

This chapter introduces the work of obtaining the latent vector that describes the comprehensive properties of the common fabrics, through three modalities respectively: color vision, depth vision, and high-resolution touch. The properties are highly related to the perception and evaluation of the fabrics. We use a CNN to achieve the goal, and jointly train the three CNNs that take data from different modalities. In the test set, we compare the similarity of the source fabric of the input data, whichever modality it is from, by comparing the distance between the CNN's embedding vector. We use the experiment of picking 1 from 10 candidate input that matches the inquiry data using the other modality to prove the effectiveness of the network model.

Chapter 6

Conclusion

This thesis introduces the robotic application of a high-resolution tactile sensor, GelSight. The sensor has a soft contact interface, and use vision-based methods to measure the geometry and traction field of the contact surface with a resolution as high as 20 microns. The applications of the sensor introduced in the thesis can be divided into two groups: manipulation, where tactile feedback helps robots to better perform manipulation tasks; exploration, where robots use tactile sensing to obtain a better understanding for the surrounding physical environment. On the manipulation side, the thesis introduced how to use the GelSight sensor to estimate the hardness of deformable materials, how to estimate a broad set of general properties of common clothes in an autonomous robot exploration loops, and how to learn the comprehensive set of fabric's material properties using a cross-modal deep learning set. The works in this thesis provide new possibilities for applying tactile sensing for a more intelligent robotic system.

The high-resolution tactile sensing, which was hardly accessible for robots before the invention of GelSight, opens more possibilities for how could tactile sensing help robots. Much more information is provided through a simple touch, but the challenges lie in deciding the goal of using the information, and exploring effective inference to apply the information. Existing works showed that the high resolution of GelSight helps it to perform extraordinarily on classifying material textures, and localizing the textures on objects. But the sensor is able to provide more information. In this thesis,

I show that by analyzing the dynamic change of the tactile signal during the physical contact, the robots can get insight into the material properties of the objects, such as hardness. The state of contact area tells other information too, such as the slip state during grasping. The research exploring multiple material properties of fabrics and clothing, also indicates that the rich tactile information at the local contact area is correlated with a broad set of material properties. The correlation could be explicit or implicit.

The thesis work also introduces the method of applying deep neural network models on high-resolution tactile data. The neural networks are designed for images, and proved effective in processing the high-dimensional data. The GelSight data, which comes in the form of images, is a typical kind of the high-dimensional data, and the neural network models for computer vision was proved useful on extracting the useful representation of the physical world from the high-dimensional tactile data, while the tasks were very challenging or even impossible using traditional methods. The experimental results also revealed some limitation of the deep learning methods, such as requiring large training datasets, and the incompetence in generalization. It remains a future research topic to exploring better neural network architectures or training methods to prevent those limitations.

The thesis also starts the trial of integrating high-resolution tactile sensing into the closed-loop robot exploration. In the clothing perception project, the robot uses tactile information to infer the clothing's material properties, but it uses visual information to plan touch movement, and conduct the touch motion with the hardware parts; at the same time, the feedback from the tactile sensing guides the robot to decide whether to conduct an extra touch exploration. In the long run, instead of only using tactile input, or any other single-modal sensing, building frameworks that well integrate multi-modal sensing and the motion system of the robots, will enable the robots to better understand and interact with the physical world.

Bibliography

[1] Edward H Adelson. On seeing stuff: the perception of materials by humans and machines. In *Photonics West 2001-electronic imaging*, pages 1–12. International Society for Optics and Photonics, 2001.

[2] Stefan Begej. Planar and finger-shaped optical tactile sensors for robotic applications. *IEEE Journal on Robotics and Automation*, 4(5):472–484, 1988.

[3] Kiran S Bhat, Christopher D Twigg, Jessica K Hodgins, Pradeep K Khosla, Zoran Popović, and Steven M Seitz. Estimating cloth simulation parameters from video. In *Proceedings of the 2003 ACM SIGGRAPH/Eurographics symposium on Computer animation*, pages 37–51. Eurographics Association, 2003.

[4] Katherine L Bouman, Bei Xiao, Peter Battaglia, and William T Freeman. Estimating the material properties of fabric from video. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 1984–1991, 2013.

[5] Sumit Chopra, Raia Hadsell, and Yann LeCun. Learning a similarity metric discriminatively, with application to face verification. In *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05)*, volume 1, pages 539–546. IEEE, 2005.

[6] Craig Chorley, Chris Melhuish, Tony Pipe, and Jonathan Rossiter. Development of a tactile sensor based on biologically inspired edge encoding. In *Advanced Robotics, 2009. ICAR 2009. International Conference on*, pages 1–6. IEEE, 2009.

[7] Virginia Chu, Ian McMahon, Lorenzo Riano, Craig G McDonald, Qin He, Jorge Martinez Perez-Tejada, Michael Arrigo, Naomi Fitter, John C Nappo, Trevor Darrell, et al. Using robotic exploratory procedures to learn the meaning of haptic adjectives. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 3048–3055. IEEE, 2013.

[8] Luke Cramphorn, Benjamin Ward-Cherrier, and Nathan F Lepora. Addition of a biomimetic fingerprint on an artificial fingertip enhances tactile spatial acuity. *IEEE Robotics and Automation Letters*, 2(3):1336–1343, 2017.

[9] Mark R Cutkosky and William Provancher. Force and tactile sensing. In *Springer Handbook of Robotics*, pages 717–736. Springer, 2016.

[10] Ravinder S Dahiya, Giorgio Metta, Maurizio Valle, and Giulio Sandini. Tactile sensingâĂŤfrom humans to humanoids. *IEEE Transactions on Robotics*, 26(1):1–20, 2010.

[11] Jia Deng, Wei Dong, Richard Socher, Li-Jia Li, Kai Li, and Li Fei-Fei. Imagenet: A large-scale hierarchical image database. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 248–255. IEEE, 2009.

[12] Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE conference on computer vision and pattern recognition (CVPR)*, pages 2625–2634, 2015.

[13] Siyuan Dong, Wenzhen Yuan, and Edward Adelson. Improved gelsight tactile sensor for measuring geometry and slip. In *Intelligent Robots and Systems (IROS 2017), 2017 IEEE/RSJ International Conference on*, pages 137–144. IEEE, 2017.

[14] Alin Drimus, Gert Kootstra, Arne Bilberg, and Danica Kragic. Design of a flexible tactile sensor for classification of rigid and deformable objects. *Robotics and Autonomous Systems*, 62(1):3–15, 2014.

[15] Nicola J Ferrier and Roger W Brockett. Reconstructing the shape of a deformable membrane from image data. *The International Journal of Robotics Research*, 19(9):795–816, 2000.

[16] Roland W Fleming. Visual perception of materials and their properties. *Vision research*, 94:62–75, 2014.

[17] Andrea Frome, Greg S Corrado, Jon Shlens, Samy Bengio, Jeff Dean, Tomas Mikolov, et al. Devise: A deep visual-semantic embedding model. In *Advances in neural information processing systems*, pages 2121–2129, 2013.

[18] Antonio Gabas, Enric Corona, Guillem Alenyà, and Carme Torras. Robot-aided cloth classification using depth information and cnns. In *International Conference on Articulated Motion and Deformable Objects*. Springer, 2016.

[19] Raia Hadsell, Sumit Chopra, and Yann LeCun. Dimensionality reduction by learning an invariant mapping. In *2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'06)*, volume 2, pages 1735–1742. IEEE, 2006.

[20] Feng Han and Song-Chun Zhu. A two-level generative model for cloth representation and shape from shading. *IEEE Transactions on pattern analysis and machine intelligence*, 29(7):1230–1243, 2007.

[21] Sepp Hochreiter and Jürgen Schmidhuber. Long short-term memory. *Neural computation*, 9(8):1735–1780, 1997.

[22] Yuji Ito, Youngwoo Kim, Chikara Nagai, and Goro Obinata. Vision-based tactile sensing and shape estimation using a fluid-type touchpad. *IEEE Transactions on Automation Science and Engineering*, 9(4):734–744, 2012.

[23] Yuji Ito, Youngwoo Kim, and Goro Obinata. Robust slippage degree estimation based on reference update of vision-based tactile sensor. *IEEE Sensors Journal*, 11(9):2037–2047, 2011.

[24] Nawid Jamali, Marco Maggiali, Francesco Giovannini, Giorgio Metta, and Lorenzo Natale. A new design of a fingertip for the icub hand. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*, pages 2705–2710. IEEE, 2015.

[25] M K Johnson and E Adelson. Retrographic sensing for the measurement of surface texture and shape. In *Computer Vision and Pattern Recognition (CVPR), 2009 IEEE Conference on*, pages 1070–1077. IEEE, 2009.

[26] Micah K Johnson, Forrester Cole, Alvin Raj, and Edward H Adelson. Microgeometry capture using an elastomeric sensor. In *ACM Transactions on Graphics (TOG)*, volume 30, page 46. ACM, 2011.

[27] Kazuto Kamiyama, Kevin Vlack, Terukazu Mizota, Hiroyuki Kajimoto, K Kawakami, and Susumu Tachi. Vision-based sensor for real-time measuring of surface traction fields. *IEEE Computer Graphics and Applications*, 25(1):68–75, 2005.

[28] Christos Kampouris, Ioannis Mariolis, Georgia Peleka, Evangelos Skartados, Andreas Kargakos, Dimitra Triantafyllou, and Sotiris Malassiotis. Multi-sensorial and explorative recognition of garments and their material properties in unconstrained environment. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016.

[29] Zhanat Kappassov, Juan-Antonio Corrales, and Véronique Perdereau. Tactile sensing in dexterous robot hands. *Robotics and Autonomous Systems*, 74:195–220, 2015.

[30] Diederik Kingma and Jimmy Ba. Adam: A method for stochastic optimization. *arXiv preprint arXiv:1412.6980*, 2014.

[31] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

[32] Robert H LaMotte and Mandayam A Srinivasan. Surface microgeometry: Tactile perception and neural encoding. In *Information processing in the somatosensory system*, pages 49–58. Springer, 1991.

[33] Yann LeCun, Bernhard Boser, John S Denker, Donnie Henderson, Richard E Howard, Wayne Hubbard, and Lawrence D Jackel. Backpropagation applied to handwritten zip code recognition. *Neural computation*, 1(4):541–551, 1989.

[34] Susan J Lederman and Roberta L Klatzky. Extracting object properties through haptic exploration. *Acta psychologica*, 84(1):29–40, 1993.

[35] Nathan F Lepora, Kirsty Aquilina, and Luke Cramphorn. Exploratory tactile servoing with active touch. *IEEE Robotics and Automation Letters*, 2(2):1156–1163, 2017.

[36] Rui Li. *Touching is believing: sensing and analyzing touch information with GelSight.* PhD thesis, Massachusetts Institute of Technology, 2015.

[37] Rui Li, Robert Platt, Wenzhen Yuan, Andreas ten Pas, Nathan Roscup, Mandayam A Srinivasan, and Edward Adelson. Localization and manipulation of small parts using gelsight tactile sensing. In *Intelligent Robots and Systems (IROS 2014), 2014 IEEE/RSJ International Conference on*, pages 3988–3993. IEEE, 2014.

[38] Yangyan Li, Hao Su, Charles Ruizhongtai Qi, Noa Fish, Daniel Cohen-Or, and Leonidas J Guibas. Joint embeddings of shapes and images via cnn image purification. *ACM Trans. Graph*, 5, 2015.

[39] Yinxiao Li, Chih-Fan Chen, and Peter K Allen. Recognition of deformable object category and pose. In *Robotics and Automation (ICRA), 2014 IEEE International Conference on*, pages 5558–5564. IEEE, 2014.

[40] OA Lindahl, Sadao Omata, and Karl-Axel Ängquist. A tactile sensor for detection of physical properties of human skin in vivo. *Journal of medical engineering & technology*, 22(4):147–153, 1998.

[41] Vivek Maheshwari and Ravi F Saraf. High-resolution thin-film device to sense texture by touch. *Science*, 312(5779):1501–1504, 2006.

[42] Philipp Mittendorfer, Eiichi Yoshida, and Gordon Cheng. Realizing whole-body tactile interactions with a self-organizing, multi-modal artificial skin on a humanoid robot. *Advanced Robotics*, 29(1):51–67, 2015.

[43] Jonas Mueller and Aditya Thyagarajan. Siamese recurrent architectures for learning sentence similarity. In *Thirtieth AAAI Conference on Artificial Intelligence*, 2016.

[44] Shogo Okamoto, Masashi Konyo, Yuka Mukaibo, Takashi Maeno, and Satoshi Tadokoro. Real-time estimation of touch feeling factors using human finger mimetic tactile sensors. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3581–3586. IEEE, 2006.

[45] Shogo Okamoto, Masashi Konyo, Yuka Mukaibo, Takashi Maeno, and Satoshi Tadokoro. Real-time estimation of touch feeling factors using human finger mimetic tactile sensors. In *2006 IEEE/RSJ International Conference on Intelligent Robots and Systems*, pages 3581–3586. IEEE, 2006.

[46] Sadao Omata and Yoshikazu Terunuma. New tactile sensor like the human hand and its applications. *Sensors and Actuators A: Physical*, 35(1):9–15, 1992.

[47] Andrew Owens, Phillip Isola, Josh McDermott, Antonio Torralba, Edward H. Adelson, and William T. Freeman. Visually indicated sounds. June 2016.

[48] Andrew Owens, Jiajun Wu, Josh H McDermott, William T Freeman, and Antonio Torralba. Ambient sound provides supervision for visual learning. In *European Conference on Computer Vision*, pages 801–816. Springer, 2016.

[49] Katsunari Sato, Kazuto Kamiyama, Naoki Kawakami, and Susumu Tachi. Finger-shaped gelforce: sensor for measuring surface traction fields for robotic hand. *IEEE Transactions on Haptics*, 3(1):37–47, 2010.

[50] John L Schneiter and Thomas B Sheridan. An optical tactile sensor for manipulators. *Robotics and computer-integrated manufacturing*, 1(1):65–71, 1984.

[51] T Shimizu, M Shikida, K Sato, and K Itoigawa. A new type of tactile sensor detecting contact force and hardness of an object. In *Micro Electro Mechanical Systems, 2002. The Fifteenth IEEE International Conference on*, pages 344–347. IEEE, 2002.

[52] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[53] MA Srinivasan. Haptic interfaces, in virtual reality: Scientific and technical challenges, 1995.

[54] Mandayam A Srinivasan and Robert H Lamotte. Tactile discrimination of shape: responses of slowly and rapidly adapting mechanoreceptive afferents to a step indented into the monkey fingerpad. *Journal of Neuroscience*, 7(6):1682–1697, 1987.

[55] Mandayam A Srinivasan and Robert H LaMotte. Tactual discrimination of softness. *Journal of Neurophysiology*, 73(1):88–101, 1995.

[56] Zhe Su, Jeremy A Fishel, Tomonori Yamamoto, and Gerald E Loeb. Use of tactile feedback to control exploratory movements to characterize object compliance. *Active Touch Sensing*, page 51, 2014.

[57] Li Sun, Simon Rogers, Gerardo Aragon-Camarasa, and J Paul Siebert. Recognising the clothing categories from free-configuration using gaussian-process-based interactive perception. In *Robotics and Automation (ICRA), 2016 IEEE International Conference on*. IEEE, 2016.

[58] M Takei, H Shiraiwa, S Omata, N Motooka, K Mitamura, T Horie, T Ookubo, and S Sawada. A new tactile skin sensor for measuring skin hardness in patients with systemic sclerosis and autoimmune raynaud's phenomenon. *Journal of international medical research*, 32(2):222–231, 2004.

[59] Hong Z Tan, Nathaniel I Durlach, G Lee Beauregard, and Mandayam A Srinivasan. Manual discrimination of compliance using active pinch grasp: The roles of force and work cues. *Perception & psychophysics*, 57(4):495–510, 1995.

[60] Hong Z Tan, Mandayam A Srinivasan, Brian Eberman, and Belinda Cheng. Human factors for the design of force-reflecting haptic interfaces. *Dynamic Systems and Control*, 55(1):353–359, 1994.

[61] Akos Tar and Gyorgy Cserey. Development of a low cost 3d optical compliant tactile force sensor. In *Advanced Intelligent Mechatronics (AIM), 2011 IEEE/ASME International Conference on*, pages 236–240. IEEE, 2011.

[62] Wouter M Bergmann Tiest. Tactual perception of material properties. *Vision research*, 50(24):2775–2782, 2010.

[63] Nicholas Wettels, Veronica J Santos, Roland S Johansson, and Gerald E Loeb. Biomimetic tactile sensor array. *Advanced Robotics*, 22(8):829–849, 2008.

[64] Ryan White, Keenan Crane, and David A Forsyth. Capturing and animating occluded cloth. In *ACM Transactions on Graphics (TOG)*, volume 26, page 34. ACM, 2007.

[65] Thiemo Wiedemeyer. IAI Kinect2. `https://github.com/code-iai/iai_kinect2`, 2014 – 2015. Accessed June 12, 2015.

[66] Bryan Willimon, Stan Birchfield, and Ian Walker. Classification of clothing using interactive perception. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1862–1868. IEEE, 2011.

[67] Robert J Woodham. Photometric method for determining surface orientation. *Optical engineering*, 1(7):139–144, 1980.

[68] Bei Xiao, Wenyan Bi, Xiaodan Jia, Hanhan Wei, and Edward H Adelson. Can you see what you feel? color and folding properties affect visual–tactile material discrimination of fabrics. *Journal of vision*, 16(3):34–34, 2016.

[69] A. Yamaguchi and C. G. Atkeson. Combining finger vision and optical tactile sensing: Reducing and handling errors while cutting vegetables. In *2016 IEEE-RAS 16th International Conference on Humanoid Robots (Humanoids)*, pages 1045–1051, Nov 2016.

[70] Hanna Yousef, Mehdi Boukallel, and Kaspar Althoefer. Tactile sensing for dexterous in-hand manipulation in robotics – a review. *Sensors and Actuators A: physical*, 167(2):171–187, 2011.

[71] Wenzhen Yuan. Tactile measurement with a GelSight sensor. Master's thesis, Massachusetts Institute of Technology, 2014.

[72] Wenzhen Yuan, Siyuan Dong, and Edward H Adelson. Gelsight: High-resolution robot tactile sensors for estimating geometry and force. *Sensors*, 17(12):2762, 2017.

[73] Wenzhen Yuan, Rui Li, Mandayam A Srinivasan, and Edward H Adelson. Measurement of shear and slip with a GelSight tactile sensor. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 304–311. IEEE, 2015.

[74] Wenzhen Yuan, Yuchen Mo, Shaoxiong Wang, and Edward H Adelson. Active clothing material perception using tactile sensing and deep learning. In *Robotics and Automation (ICRA), 2018 IEEE International Conference on*. IEEE, 2018.

[75] Wenzhen Yuan, Mandayam A Srinivasan, and Edward H Adelson. Estimating object hardness with a GelSight touch sensor. In *Intelligent Robots and Systems (IROS), 2016 IEEE/RSJ International Conference on*, pages 208–215. IEEE, 2016.

[76] Wenzhen Yuan, Shaoxiong Wang, Siyuan Dong, and Edward Adelson. Connecting look and feel: Associating the visual and tactile properties of physical materials. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 5580–5588, 2017.

[77] Wenzhen Yuan, Chenzhuo Zhu, Andrew Owens, Mandayam A Srinivasan, and Edward H Adelson. Shape-independent hardness estimation using deep learning and a GelSight tactile sensor. In *Robotics and Automation (ICRA), 2017 IEEE International Conference on*, pages 951–958. IEEE, 2017.