博士論文

Understanding Hand Manipulation from First-Person View Videos

(一人称視点映像からの手操作解析に関する研究)

東京大学大学院
情報理工学系研究科
電子情報学専攻

48-137405　蔡 敏捷

指導教員　佐藤 洋一 教授

平成 27 年 12 月

Abstract

Understanding the ways how human hands interact with objects (hand manipulation) automatically from daily tasks is important for domains such as robotics, human grasp understanding, and motor skill analysis. To promote the study of daily hand manipulation, I present a recognition framework for hand manipulation under first-person vision paradigm with a wearable camera, which overcomes the constraints of tactile sensors and calibrated cameras used in traditional approaches. However, the tasks of recognizing different types of hand manipulation from first-person view video are challenging due to rapidly changing background, ambiguous hand appearance and mutual hand-object occlusions. To tackle the challenges, I propose approaches to reason about semantic information of hands and objects which are considered critical in understanding hand manipulation.

The thesis work is composed by three components which address different aspects of understanding hand manipulation from first-person view videos: (1) An image-based approach for hand grasp analysis from image appearance is presented, which plays a central role in understanding hand manipulation; (2) A sequence-based method is proposed for hand grasp analysis from a different perspective of hand dynamics rather than static appearance; (3) An unified framework for recognizing grasp types, object attributes and manipulation actions is proposed, in which semantic relationship between hands, objects, and actions is modeled.

The study of hand grasp plays a central role in understanding hand manipulation since hand grasp characterizes the ways how hand hold an object and implies attribute information of the manipulated object. Therefore, an appearance-based approach for hand grasp analysis under first-person vision (FPV) paradigm is first presented. The proposed approach recognizes the types of hand grasp from image appearance and analyzes visual similarity among different grasp types (visual structures of hand grasp). Experiment results demonstrate the potential of automatic grasp recognition in unstructured environments. Analysis of real-world video shows that it is possible to automatically learn intuitive visual grasp structures that are consistent with expert-designed grasp taxonomies.

Appearance-based method is insufficient to discriminate between different grasp types which are ambiguous from a single image, and is sensitive to unreliable hand detection. To address this problem, I propose a sequence-based method to study hand grasp from perspective of hand dynamics. In particular, a feature representation which encodes dynamical information of hand appearance and motion is proposed based on hand-guided feature tracking from image sequences. In addition, I propose a metric for comparing hierarchical clusters in order to quantitatively evaluate the consistency between different visual structures of hand grasp. Through extensive experiments, effectiveness of the proposed method is verified that hand dynamics can help improve grasp recognition and learn more consistent grasp structures.

Building on the work of hand grasp analysis, a further step is taken to study hand manipulation in a broader scale. I believe that grasp types together with object attributes provide complementary information for characterizing different manipulation actions. Thus, I propose an unified model for recognizing hand grasp types, object attributes and manipulation actions from a single image. Experiments strongly support the hypothesis that: (1) Attribute information of the manipulated object can be extracted without any specific object detectors by

exploring spatial hand-object configuration; (2) Contextual information between grasp types and object attributes is important in dealing with mutual hand-object occlusions; (3) Action models that address the semantic relationship with grasp types and object attributes outperform traditional appearance-based models which are not designed to take into account semantic constraints and are overfit to image appearance.

Acknowledgements

I could never have completed this work without the support and assistance of many people. First and foremost, I would like to express deepest gratitude to my adviser, Prof. Yoichi Sato, for his kind advising, valuable suggestions, and good-hearted encouragement in academic. Without his tolerant and open-minded style, I could not have a chance to start my PhD pursuit in computer vision without any experience. With his help, I learned how to read an academic paper; how to formalize a research problem; and how to write a paper and present the work. His wide curiosity inspires me to think what a researcher really is. I also would like to express grateful thanks to Dr. Kris M. Kitani from Carnegie Mellon University (CMU). With his help, I learned how to tackle research problems, how to compose and express an idea more logically, and most importantly how to think like a scientist rather than an engineer. His insight and passion for the research encourage me to face and overcome difficulties and setbacks with perseverance and optimism. Once again, I thank them from the bottom of my heart for their consistent and generous support which helps me grow as a researcher.

I would like to express my deepest love and gratitude to my wife, Han Zhang, for her great support in my life. It is for her unconditional love that she quited her job and came to Japan with me after our marriage in the end of my PhD first year. It is her tender care and delicious meal every day that make me full of energy and concentrate on my thesis work.

I also would like to thank all members of Sato Laboratory, for their kindness and assistance. I would appreciate the final defense committee members: Prof. Shin'ichi Satoh, Prof. Kiyoharu Aizawa, Prof. Toshihiko Yamasaki, and Prof. Takeshi Oishi, for their valuable comments and suggestions to my work.

This thesis would not have been possible without generous financial support from Japanese Government (MEXT) Scholarship program. MEXT Scholarship program also provides a chance for me to make good friendship with other recipient, which makes my life in Japan happy and memorable. These supports are gratefully acknowledged.

Finally, I would like to express my feelings of loves and gratitude to my parents. Even though they know little about foreign countries or academic, it is their perpetual support and unconditional love that make me overcome all the difficulties through all the years of my study and living in Japan.

December 2015

Contents

List of Figures

List of Tables

Chapter 1

Introduction

This thesis aims to automate the understanding of hand-object interactions (hand manipulation) in daily tasks using a wearable monocular camera. In particular, I focus on recognizing (1) hand grasp types and (2) manipulation actions from first-person view videos. *Hand grasp types* are a discrete set of canonical hand poses often used in robotics to describe various grasping strategies for objects. For example, the use of all fingers around a curved object like a cup is called "medium wrap". Figure 1.1 shows examples of different grasp types. *Manipulation actions* in this work refer to different patterns of hand-object interactions such as "open" or "pour". Figure 1.2 shows examples of different manipulation actions.

1.1 Motivations

The ability to understand daily hand-object interactions automatically from visual sensing is important for domains such as robotic manipulation [Cut89] [YLFA15b], human grasp understanding [FBD14], and motor control analysis [CSPA+92]. In robotic manipulation, the study of human hand function

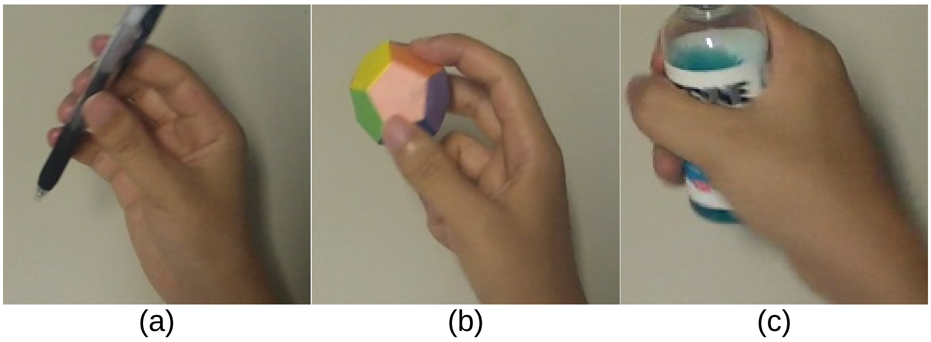


Figure 1.1: Examples of hand grasp types. Images come from a self-collected dataset. (a) Thumb-n Finger (b) Tripod (c) Medium Wrap

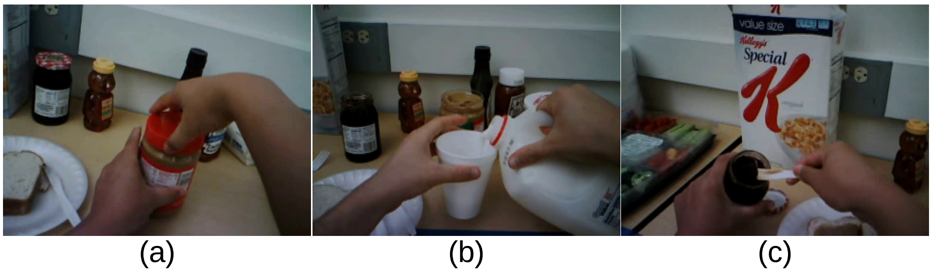


Figure 1.2: Examples of manipulation actions. Images come from a public dataset [FLR12].(a) Open (b) Pour (c) Scoop

provides critical information about robotic hand design and action planning. In human grasps understanding, the recognition of hand-object manipulations enables automatic analysis of human manipulation behavior, making it more scalable than traditional manual observation used in previous studies [ZDLRD11].

Existing approaches on studying hand-object interactions have been developed primarily in the controlled laboratory settings which often include hand-contact sensors or calibrated cameras as shown in Figure 1.3. However, there are many limitations in these settings. Intrusive sensors often inhibit free hand-object interactions; calibrated camera system requires hand manipulation to be recorded in limited workspace. As a result, hand-object interactions in everyday manipulation tasks have seldom been studied.

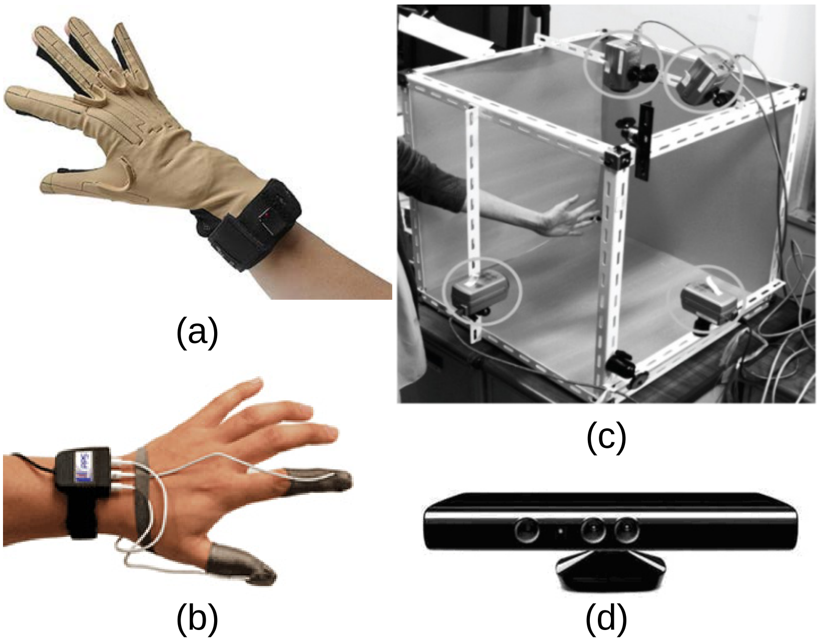


Figure 1.3: Examples of sensors used for capturing hand motion and interactions in controlled laboratory settings. (a) CyberGlove [CYB] used for measuring hand articulation (b) isoTOUCH [ISO] used for measuring finger touch pressure (c) Camera arrays (d) Kinect [KIN] RGB-D sensor

To promote the study of natural hand-object interactions, I propose first-person vision-based approaches for understanding hand-object interactions using a wearable camera in this thesis. A wearable camera (as shown in Figure 1.4) overcomes the constraints of other modes of direct sensing by allowing for continuous recording of natural hand interactions at a large scale, both in time and space. Furthermore, it provides an ideal first-person viewing perspective under which hands and objects are visible up-close in the visual field.

However, understanding manipulations with first-person vision is also

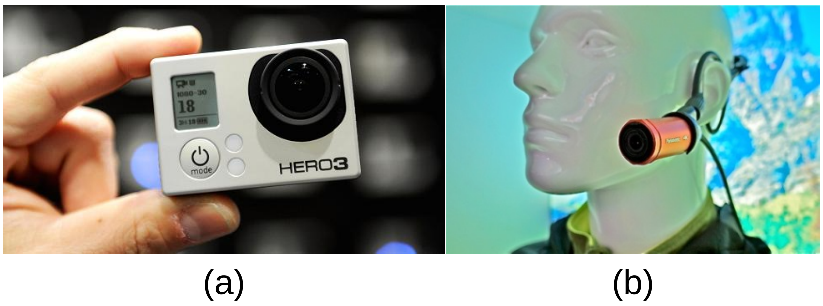


Figure 1.4: Examples of wearable cameras which can record first-person view videos. (a) GoPro HERO3 [GOP] (b) Panasonic HX-A1 [HX-]

very challenging. There are many occlusions of the hand, especially the fingers, during hand-object interactions. It is also challenging to reliably detect the manipulated object since the object is also often occluded by the hand. Furthermore, cluttered background with rapidly changing appearance is a common situation in first-person view videos, which makes it unreliable to directly model hand manipulation from image appearance. This suggests that semantic information about the hands and objects need to be reasoned about.

I believe the ability of recognizing different hand grasp types is of great importance in understanding hand manipulation, since hand grasp characterizes the way how hand holds an object during manipulation. Hand grasp also implies attribute information of the manipulated object as the object attributes, such as shape and mass, affect the selection of different grasp types. Furthermore, hand grasp helps describe the functionality of an action, whether it requires more power, or more flexible finger coordination. Thus, in this thesis I propose approaches for hand grasp analysis with first-person vision which play a central role for understanding hand manipulation.

1.2 Overview

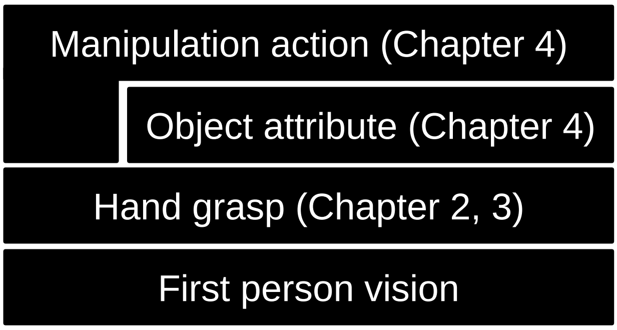

Manipulation action (Chapter 4)
Object attribute (Chapter 4)
Hand grasp (Chapter 2, 3)
First person vision

Figure 1.5: Structure of the thesis work. Hand grasp is studied under the first-person vision paradigm, which plays a central role in this thesis. Object attributes are extracted by exploring spatial hand-object configuration, and manipulation action is modeled by its semantic relations with hand grasp and object attributes.

In this thesis, I propose approaches for understanding hand manipulation under the first-person vision paradigm, in which semantic information about hand grasp, object attribute, and manipulation action are studied based on their intrinsic logical structure. The hierarchical structure of this thesis is illustrated in Figure 1.5.

The thesis work is composed by three components which address different aspects of understanding hand manipulation: (1) An image-based approach for hand grasp analysis in unstructured environments is presented in Chapter 2, which recognizes hand grasp types from a single image and analyzes visual similarity between different grasp types; (2) A sequence-based method for hand grasp analysis from dynamical hand information is proposed in Chapter 3; (3) An unified framework for recognizing grasp types, object attributes and manipulation actions is presented in Chapter 4, in which semantic relationship between hands, objects, and actions is modeled. The overview of this thesis is given as follows:

1.2.1 Hand grasp analysis with static appearance features

Grasp is commonly defined as every hand postures used for holding an object stably during hand manipulation tasks. The study of hand grasp plays a central role in understanding hand manipulation, since for most manipulation tasks objects are first required to be grasped by hands and then the following manipulation can be performed. Traditional approaches to grasp analysis have been developed primarily in controlled laboratory settings which pose limitations on the recording and study of free hand-object interactions. As a result, hand grasp in everyday manipulation tasks has seldom been studied.

To enable hand grasp analysis in natural working/living scenes, an appearance based approach for hand grasp analysis is presented which can recognize different hand grasp types in unstructured environments using a wearable monocular camera. A wearable camera allows for continuous recording of natural manipulation tasks and enables the study of hand grasp at a large scale. It also provides an ideal first-person viewing perspective for grasp analysis. The proposed approach incorporates advances of computer vision techniques. In particular, egocentric hand detection techniques are adopted to segment hand regions, and popular appearance-based features are extracted for training discriminative grasp classifiers. Building on the output of grasp classifiers, visual similarity among different grasp types are analyzed and visual structures of hand grasp are automatically learned. Experiments show the potential of automatic grasp recognition in unstructured environments.

1.2.2 Hand grasp analysis with dynamic appearance features

Visual grasp recognition is a challenging task and the appearance-based method is unreliable in real world scenario. Different hand grasp types are ambiguous from a single image as they share similar hand shape/appearance, thus hand appearance alone is insufficient to discriminate between different grasp types. It is also challenging to reliably detect the hand and the appearance-based method is sensitive to hand detection noises.

In this chapter, hand grasp is studied from a different perspective of hand dynamics rather than static appearance. In particular, I propose a new feature representation based on hand-guided feature tracking to encode dynamical information of hand appearance and motion from image sequences. The hand-guided feature tracking is called "Dense Hand Trajectories" (DHT). Dense hand trajectories are obtained by densely sampling and tracking feature points in a short interval of images which are guided by hand detection. Feature descriptors are computed for each trajectory to encode the information of both hand motion and hand appearance. The feature representation based on dense hand trajectories has several advantages over appearance-based features. First, trajectory itself contains motion information of the hand during interaction which is useful for identifying different grasp types. Second, hand appearance at multiple adjacent images along the hand trajectory can be computed as more compact representation for single grasp type than image-based features. Moreover, features based on hand tracking are more robust to hand detection noises than hand appearance-based features. Extensive experiments verified effectiveness of the proposed method.

1.2.3 Understanding manipulation actions with grasp types and object attributes

The ability to understand actions of hand-object manipulation automatically from images is important for domains such as robotic manipulation, human grasp understanding, and motor control analysis. However, the recognition task for understanding manipulations from monocular images is also very challenging. There are many occlusions of the hands and the manipulated objects during interactions.

In this work, I propose a novel method to extract object attribute information from the manipulated object. Furthermore, the recognition of grasp types and object attributes is enhanced by exploring their mutual context information (contextual relationship between two components that by knowing one component facilitates the recognition of the other). Finally, a semantic action model based on grasp types and object attributes is provided. Specifically, discriminative classifiers for different actions are trained based on the recognition output (belief distribution) of grasp types and object attributes.

There are several advantages for jointly modeling actions in this way: (1) Grasp type helps describe the functionality of an action, whether it requires more power, or more flexible finger coordination; (2) Object attributes provide a general description about the manipulated object and indicates possible interaction patterns; (3) High-level semantic labels of grasp types and object attributes enable the model to encode high-level constraints (*e.g.*, medium wrap can only be used for cylindrical objects) and as a result, results of the learned model are immediately interpretable. Experiments strongly support our hypothesis.

Chapter 2

Hand grasp analysis with static appearance features

2.1 Background

Grasp is commonly defined as every hand postures used for holding an object stably during hand manipulation tasks. Understanding the way how humans grasp object is important in different domains ranging from robotics [Cut89], prosthesis [Kel47], hand rehabilitation [WCE$^+$01], to motor control analysis [CSPA$^+$92] and many others. In robotics, the study of hand function provides critical information regarding design of robotic hands [Cut89]. In rehabilitation, statistical information about daily usage of grasp types is an important factor in evaluation criterion for injured hand recovery [WCE$^+$01]. I believe the ability of automatic hand grasp analysis is of great importance in understanding hand manipulation, since for most manipulation tasks objects are first required to be grasped by hands and then the following manipulation can be performed. Thus the study of hand grasp plays a central role in this thesis.

Traditional approaches to grasp analysis often use tactile sensors which can provide precise measurement of hand articulation and finger touch pressure. However, intrusive hand sensors are required to be worn and often inhibit free hand interactions. As a result, hand grasp analysis is mainly conducted in controlled laboratory settings. In recent years, although some researchers have studied daily hand usage based on manual annotation of egocentric video recording everyday tasks, the annotation process required many hours of visual inspection by skilled annotators and such manual approaches can not scale to larger datasets.

The goal of this chapter is to develop a fully automatic recognition system for studying hand grasp in natural hand-object interactions. In particular, I propose an image-based approach for hand grasp analysis under first-person vision paradigm using a wearable camera. A wearable camera is qualified for its portability and allows for continuous recording of daily activities at a large scale. It also provides an ideal egocentric viewing perspective for grasp analysis with hand-object interactions naturally recorded in the center of the visual field.

The proposed approach incorporates advances of computer vision techniques that can be used as a tool to advance studies in prehensile analysis. In particular, state-of-the-art egocentric hand detection techniques are adopted in order to deal with the new challenges of first-person vision such as unconstrained hand movements and rapidly changing imaging conditions (i.e., illumination and background) due to extreme camera motion. Based on detected hand regions, popular appearance-based features are examined and extracted as feature representation for hand grasp. Grasp classifiers are trained for discriminating between different grasp types. Finally, the grasp classifiers are used to learn the visual similarities between grasps in order to automatically build an appearance based grasp hierarchy, which we call *the visual structures of hand grasp*. In the experiments, the analysis of real-world video shows that it is possible to automatically learn intuitive visual grasp structures that are consistent with expert-designed grasp taxonomies.

The contributions of this chapter are as follows: 1) An appearance-based approach is proposed for hand grasp recognition from a single image recorded by a wearable camera. 2) An iterative clustering method is proposed for learning visual structures of hand grasps using a visual clustering approach which enables the system to automatically learn task-based grasp taxonomies.

This chapter is organized as follows: Section 2.2 gives a brief review of the related works about hand grasp taxonomy and hand detection in first-person view video. Section 2.3 describes the architecture and main components of our first-person vision-based system. Performance evaluation of the system is shown in Section 2.4, and conclusions are made in Section 2.5.

2.2 Related works

2.2.1 Human grasp taxonomy

Grasp taxonomies have been studied for decades to better understand the use of human hands [Sch19, Kel47, Nap56, IBA86, Cut89, KI93, FPS+09]. Early work by Schlesinger [Sch19] classified hand grasps into 6 major categories based on hand shape and object properties. In 1956, Napier proposed a scheme [Nap56] that divides grasps into power and precision grasps based on requirements of the manipulation task. The categorizations of power and precision grasps was widely adopted by researchers in the medical, biomechanical and robotic fields. In studying grasps in manufacturing tasks, Cutkosky provided a comprehensive hand grasp taxonomy [Cut89] which played an important role in guiding robotic hand design. In the early 1990's, Kang and Ikeuchi [KI93] presented a computational framework for grasp identification, allowing automatic grasp planning of a robotic system from a demonstrated human grasp. Recently, Huang et al. [HMMK15] proposed an unsupervised method to discover appearance-based grasp taxonomies. In their method, hand images with similar appearance are clustered together as distinct grasp types.

The human grasp taxonomy proposed by Feix et al. [FPS+09] is the most complete to date as argued and has been widely used in grasp analysis in recent years [RFKK10, BZR+13, DB14]. Considerable efforts have been devoted in obtaining the statistics of human hand use [BZR+13, BFD13, FBD14]. The created statistics is based on manual annotation of egocentric video recording everyday tasks. However, the annotation process required many hours of visual inspection by skilled annotators. As it becomes easier to acquire large amounts of visual data, it is clear that manual approaches will not scale to larger datasets. In this work, however, the aim is to propose an automatic first-person vision-based framework that will help to support next generation research in the area of prehensile analysis using a large amount of video data.

2.2.2 Automated grasp analysis

Approaches for automatic hand grasp analysis have been developed primarily in structured environment. Hand tracking devices such as data gloves or inertial sensors have been used to obtain detailed measurements of joint angles and positions of the hand [SFS98, FGE+99, BOID05, EK05]. Santello et al. [SFS98] used Principle Component Analysis (PCA) to analyze finger coordination of hand grasp using joint angle data from a data glove. However, the main limitation of hand tracking devices is that they must be worn on the hand and inhibit free hand interactions.

Vision tracking of hand grasping an object [KRK08, HSKMVG09, OKA11, RKEK13] allows a completely non-contact markerless form of hand interactions. Romero et al. [RKEK13] proposed a non-parametric estimation method to track hand poses interacting with objects by performing a nearest neighbor search in a large synthetic dataset. However, most visual tracking systems require that hand interactions are recorded in a structured environment. In this work, a first-person vision-based approach is proposed which can handle large scale video data in real-life manipulation tasks.

2.2.3 Hand detection in first-person vision

Wearable camera allows for continuous recording of hand interactions in real world environments at a large scale and provide an ideal first-person viewing perspective for studying hand interactions. Recognition from egocentric video has become a popular topic in computer vision community. Li and Kitani [LK13] first addressed hand detection problem in the context of egocentric video. They proposed a pixel-level hand detection method which can adapt to changing illuminations. Li et al. [LFR13] studied the eye-hand coordination in egocentric video and used mid-level information from hand detection to predict where the eyes look. Baraldi et al. [BPS+14] proposed to use dense trajectories with hand segmentation for hand gesture recognition in ego-vision scenarios. Dense trajectories which is often used in action recognition is proved to work well in egocentric paradigm. Rogez et al. [RIR15] recently presented promising results on discrete hand pose recognition from a chest-mounted RGB-D camera. However, these discrete poses have no di-

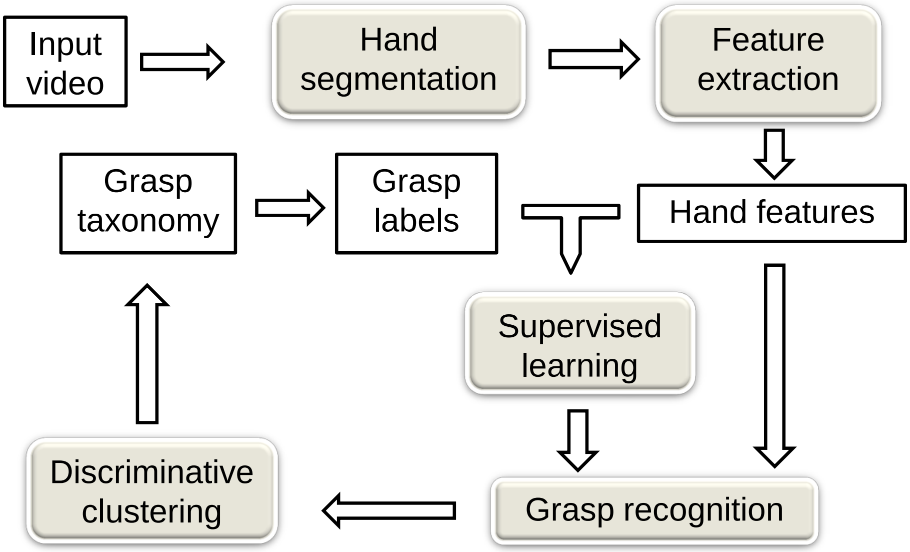


Figure 2.1: Outline of the proposed framework.

rect semantic correspondence to hand grasp types. This work is the first to develop computer vision-based techniques for grasp recognition under first-person vision.

2.3 Grasp analysis framework

A scalable grasp analysis framework is desired which can recognize different hand grasp types in daily manipulation tasks and learn visual structures of hand grasps from large scale of data. To achieve this goal, a first-person vision-based approach is developed which can learn discriminative classifiers and visual structures of hand grasp automatically with a single wearable camera. The outline of the framework is illustrated in Figure 2.1. The input to the system is egocentric video recording daily manipulation tasks. Based on state-of-the-art hand detection techniques hand regions are segmented from egocentric videos. Then grasp-related features are extracted

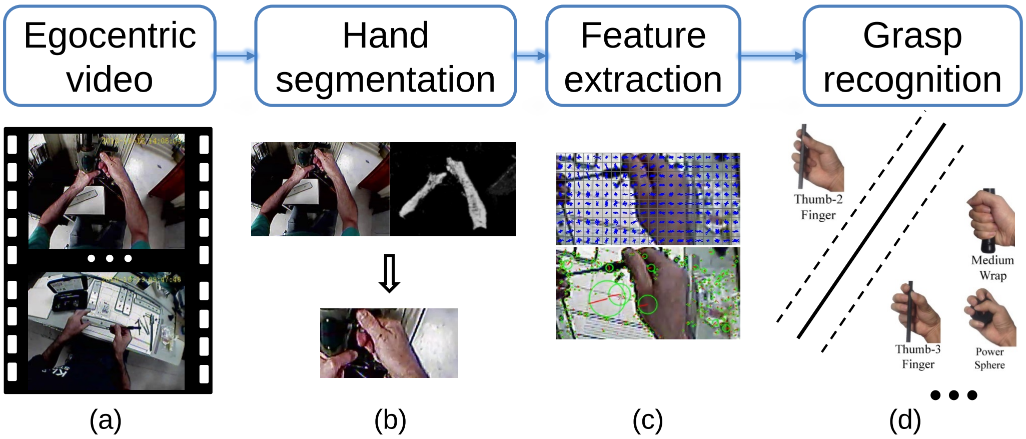


Figure 2.2: Pipeline of the grasp recognition system. (a) Manipulation task recorded with a wearable camera (b) Hand segmentation from pixel-level hand detection (c) Appearance-based features (d) Multi-class classification

from hand regions for training discriminative grasp classifiers. Finally, an iterative discriminative clustering method is used to learn visual structures of hand grasp.

More specifically, the procedure of grasp recognition (after excluding the clustering part of the proposed framework) is similar to ordinary visual recognition system, and its pipeline is demonstrated in Figure 2.2. Different components of grasp recognition as well as the discriminative clustering will be described in details in the following subsections.

2.3.1 Hand segmentation

Robustly identifying hand regions with a wearable camera is a challenging yet essential pre-processing needed to automate hand grasp analysis. As the camera is mobile, the background is rapidly changing, hands are moving without constraint and the camera can move with extreme ego-motion.

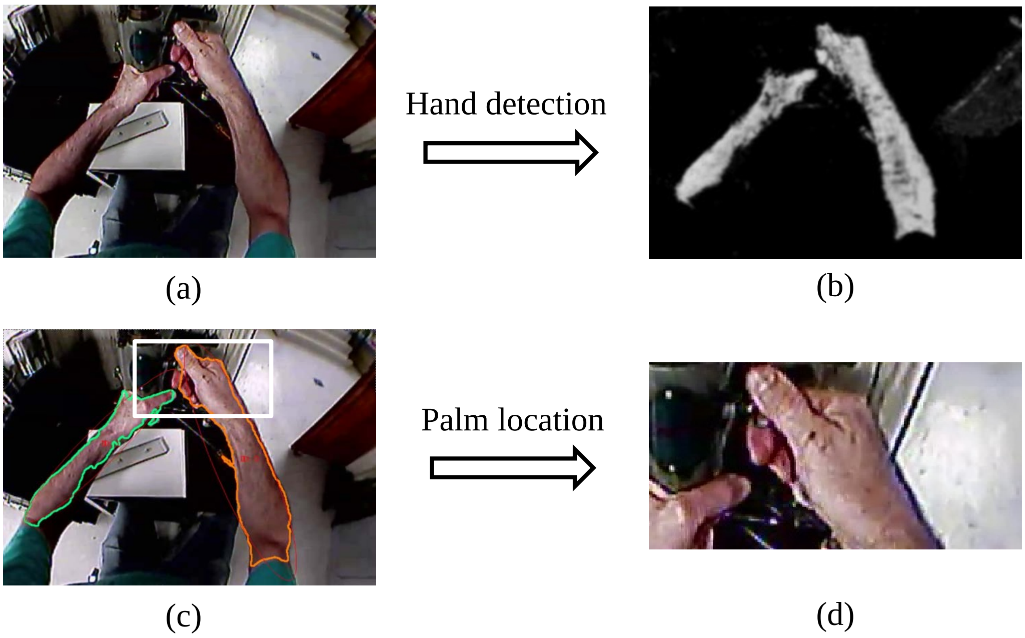


Figure 2.3: Example of hand segmentation. (a) Image from egocentric video (b) Hand probability map (c) Candidate hand regions (d) Hand region within a bounding box

Recent work on detecting hand regions using a wearable camera has shown that robust hand detection performance can be achieved if the hand model is rapidly adapted to changes in imaging conditions [LK13]. Following [LK13], a multi-model hand detector is trained which is composed by a collection of hand pixel classifiers indexed by global appearance models. Given a test image, the global appearance modeled by a color histogram is computed as a visual probe, for every frame, in order to recommend the n-best hand pixel classifiers. Based on the multi-model hand detector, a probability map is generated for each image as illustrated in Figure 2.3(b). The value of each pixel represents the likelihood of being a hand pixel in the original image.

Once the hand probability map has been detected, hand region, which contains most of the grasp information, is then segmented with a bounding box. Candidate hand regions with arms are first selected by binarizing the probability map with a threshold. Regions under a certain area proportion are discarded and at most two regions are retained. Fig. 2.3(c) shows two candidate hand regions painted with green and orange contours. In this work I only consider the right hand grasp. The left hand is suppressed by simply selecting the candidate hand region which is right-most. If no hand region is detected, that is when no hands are visible, the image is discarded. Each hand region is extracted with a fixed size bounding box which is shown as the white rectangle in Fig. 2.3(c). In detail, ellipse parameters (length of long/short axis, angle) are fitted to the original hand region. The arm part is approximately removed by shortening the length of long axis to 1.5 times of the length of short axis. A fixed size bounding box is drawn by fixing the top-center of the bounding box to the top-center of the arm-removed hand region. The size of the bounding box is determined heuristically for each video and takes advantage of the fact that the distance between the hands from the head-mounted camera is consistent across various manipulation tasks.

2.3.2 Feature representation

In expert-defined grasp taxonomies, different grasp types are often identified by hand postures, object properties and types of hand-object interactions. Therefore, grasp-related features for palm regions are examined which encodes the shape of different hand postures and visual context of manipulated objects.

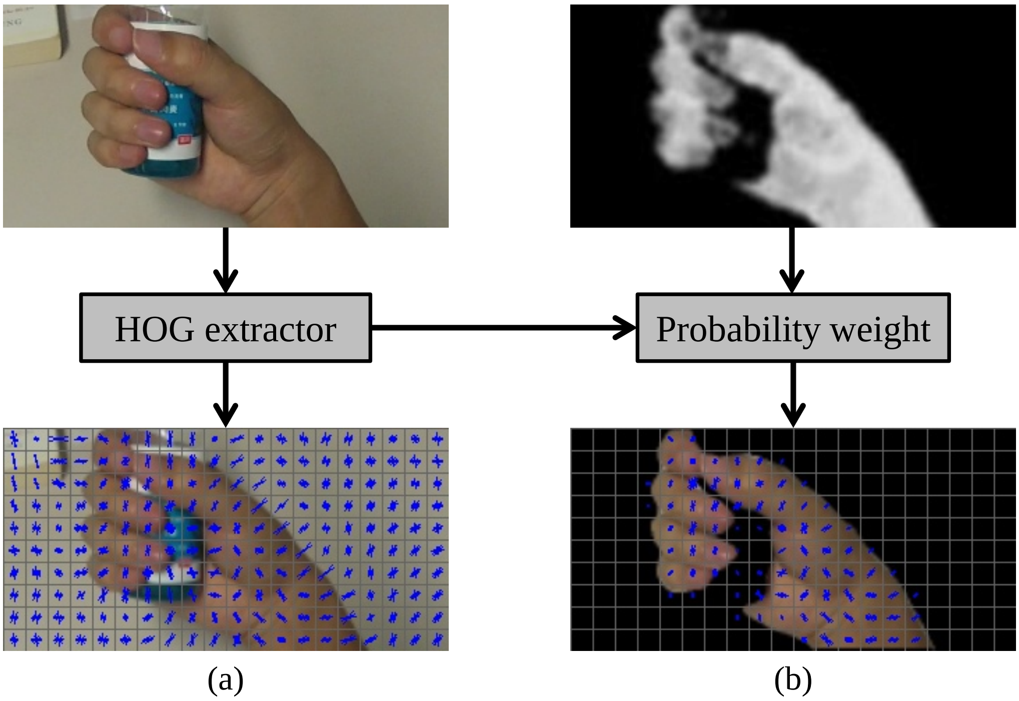


Figure 2.4: Visualization of HOG features. (a) HOG (b) HandHOG

Hand shape

Hand shapes are represented with Histogram of Oriented Gradient (HOG) [DT05] computed from a palm region. HOG features are an image descriptor based on collected local distributions of intensity gradients and have been widely used in object detection. The HOG features are computed by first dividing a palm region into a grid of smaller regions (cells) and then computing histogram of gradient orientations in each cell. Cell histograms within a larger region (blocks) are then accumulated and normalized to make the block descriptor less sensitive to varying illumination. Finally, the resulting block histograms are concatenated to form a HOG feature descriptor. A cell size of 8 × 8 pixels, block size of 16 × 16 pixels, and window size of 160 × 80 pixels with 9 orientation bins are used. A visualization of example HOG features is shown in the bottom-left of Fig. 2.4.

In the experiments, three variants of the HOG feature descriptor are examined. The first is the global HOG feature described above. The second is a dimension-reduced version of HOG using Principle Component Analysis (HOG-PCA) to reduce the dimension of feature descriptor from 6156 to 100. The third is HOG features weighted by a skin probability map (HandHOG). HandHOG effectively suppresses gradients due to object being manipulated or background regions. As shown in Fig. 2.4, HOG features corresponding to non-hand regions are removed by weighting each block histogram by squared hand probability at the center of the block.

Object context

Features based on local keypoints are also examined in order to capture the visual context of the object and hand-object interaction. In particular, the following two local gradient descriptors are extracted.

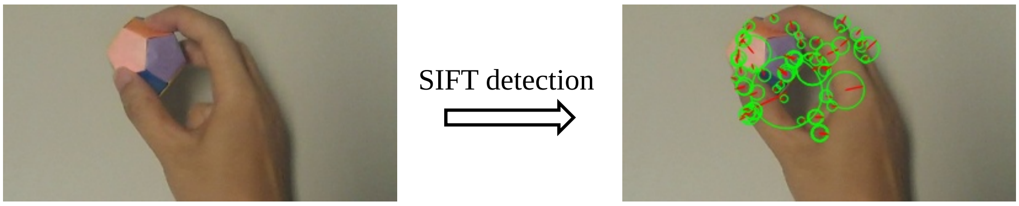


Figure 2.5: Visualization of SIFT keypoints.

SIFT features [Low04] are extracted as a representation of the visual context of manipulated objects. Example keypoints are visualized in Fig. 2.5 where the scale and orientation of each keypoint are illustrated with a circle and a red radius. Histogram of gradients around each keypoint is computed as a keypoint descriptor. Note that keypoints are detected around the object and the part of the hand in contact with the object. A bag-of-words (BOW) approach for obtaining an image descriptor is used which contains the frequency of keypoint patterns. A total of 100 keypoint patterns are generated using k-means clustering over all keypoint descriptors.

In addition to the SIFT BOW, the same approach is used to obtain a 100-dimensional image descriptor counting frequency of block-based HOG features which are generated using k-means clustering over all block HOG descriptors. The two 100-dimensional feature vectors are then concatenated together to generate a new feature (BlockHOG-SIFT).

2.3.3 Grasp recognition

One-versus-all multi-class grasp classifiers are trained for the grasp types defined in Feix's taxonomy [FPS+09]. This taxonomy is preferred since it is the most complete one in existence and has previously been applied to grasp analysis in [BFD13][FBD14]. Probability calibration [Pla99] is performed for each classifier in order to produce comparable scores. During testing, each frame is classified to the grasp type of the classifier with the highest score.

A correlation index is also defined for evaluating the visual similarity between different grasp types based on classification results. The correlation index $C_{i,j}$ between grasp type i and grasp type j is defined as:

$$C_{i,j} = \frac{m_{i,j} + m_{j,i}}{2}\left(\frac{1}{n_i} + \frac{1}{n_j}\right) \tag{2.1}$$

where $m_{i,j}$, $m_{j,i}$ denotes the number of samples from grasp type i misclassified as grasp type j and vice versa. n_i, n_j are the number of samples from grasp type i and grasp type j, respectively.

2.3.4 Discovering visual structures of hand grasp

The visual similarity between different grasp types poses big challenges in training discriminative grasp classifiers based on visual features. Some visually similar grasp types are extremely difficult to differentiate, even for human annotators. Taking *Thumb-2 Finger* and *Thumb-3 Finger* for example, it is hard to tell how many fingers are used in holding the tool only from visual perception.

Algorithm 1 Iterative Grasp Clustering

```
Initialize: $N \Leftarrow$ the number of grasp types, consider each grasp type as a single-member grasp cluster
while $N > 1$ do
    Step1: Train grasp classifiers for each grasp cluster
    Step2: Perform grasp classification, compute correlation index for each pair of grasp clusters
    Step3: Merge two grasp clusters with biggest correlation index into one grasp cluster, $N \Leftarrow N - 1$
end while
```

To address this challenge, I take another direction to explore the visual structures of hand grasps based on the correlation between visually trained grasp classifiers. As introduced in Section 2.3.3, a correlation index is defined for evaluating the visual similarity between different grasp types based on classification results. Based on the correlation index, An iterative grasp clustering algorithm was implemented by iteratively clustering two most similar grasp types. The algorithm is described in Algorithm 1. This procedure defines a visual structure between grasp types – a grasp dendrogram.

By incorporating the iterative clustering procedure into the grasp recog-

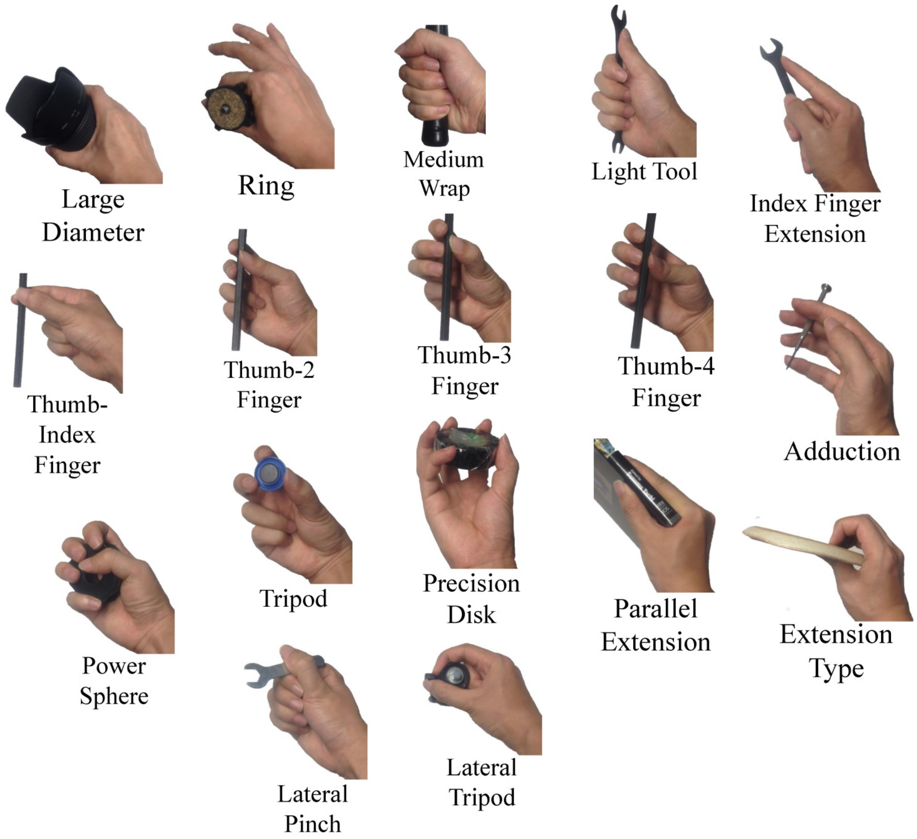


Figure 2.6: 17 different grasp types from the Feix's taxonomy[FPS+09]. Grasp types are selected based on the study of grasp usage in [BFD13].

nition procedure, an first-person vision system for hand grasp analysis is composed as illustrated in Fig. 2.1. With input of first-person view video recording daily manipulation tasks, this system can recognize different hand grasp types and learn visual structures of hand grasp automatically.

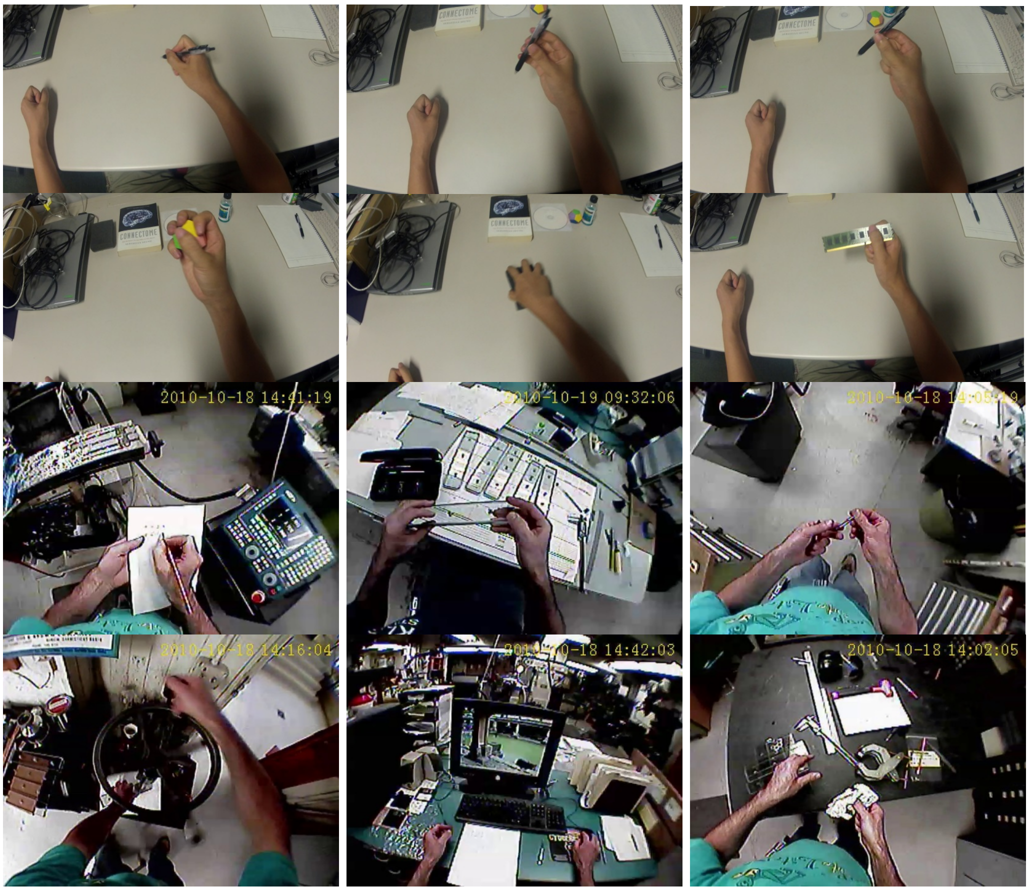


Figure 2.7: Images samples from UT Grasp Dataset (top 2 rows) [CKS15] and Machinist Grasp Dataset (bottom 2 rows) [BFD14].

2.4 Evaluation

2.4.1 Experimental setting

To explore the effectiveness of the examined visual features for recognizing grasp types, a new dataset was collected under controlled environment ("UT Grasp Dataset"). Only a subset of grasp types in Feix's taxonomy are considered in our dataset, since not all the grasp types are commonly used in everyday activities. Seventeen grasp types were selected as shown in Fig. 2.6 based on the statistical result of grasp prevalence provided by Bullock et al. [BFD13]. Four subjects were asked to grasp a set of objects placed on a desktop after brief demonstration of how to perform each type of grasps. Each subject performed hand grasps with a unique set of objects (*e.g.*, different objects with a cylindrical shape are used by different subjects in the *medium wrap* grasp type). Video was recoded by a HD head mounted camera (GoPro Hero2) at 30 fps while subjects performed each grasp type with varying hand poses. The recorded video was then downsized to 960×540 pixels. Fig. 2.7 (top 2 rows) shows some images from UT Grasp Dataset.

To examine the proposed system in more natural environments, a real-world grasp dataset [BFD14] is used, which is composed of 20 video sequences recording a machinist's daily work ("Machinist Grasp Dataset"). The Machinist Grasp Dataset is part of a larger human grasping dataset provided by Yale University and is manually labeled with grasp types. The video quality of the Machinist Grasp Dataset is relatively low with the image resolution of 640x480 pixels. Fig. 2.7 (bottom 2 rows) shows some example images. In the experiments on Machinist Grasp Dataset, rare grasp types were removed and seventeen remaining grasp types were selected which at least take place three times through out all sequences. The 17 grasp types in Machinist Grasp Dataset are slightly different from that in UT Grasp Dataset since grasp usage varies in different tasks.

Hand regions are segmented with a bounding box with the size of 320×160 for UT Grasp Dataset and 256×128 for Machinist Grasp Dataset. Then four feature descriptors (HOG, HOG-PCA, HandHOG, and BlockHOG-SIFT) are extracted for each of the segmented hand regions as explained in Section 2.3.2. Finally, three types of classifiers are trained by using the obtained feature descriptors: (1) Linear Support Vector Machine (SVM-linear), (2) SVM with Radial Basis Function kernel (SVM-rbf), and (3) Exemplar SVM (ESVM). The average F1 score computed from a weighted average of the F1 score of each grasp type is used for evaluating the grasp recognition performance. Value ranges from 0 to 1, where 1 represents perfect performance.

2.4.2 Performance of grasp recognition

The proposed approach is applied to UT Grasp Dataset and Machinist Grasp Dataset to see how visual features can discriminate between different grasp types in both controlled and natural environments.

First grasp recognition results are presented for a single user on UT Grasp Dataset. Grasp classifiers are trained and tested for each user using 5-fold cross validation. The average F1 scores of the 17 grasp classifiersare shown in Table 2.1 for different feature descriptors and different machine learning algorithms. From Table 2.1, it can be seen global features (HOG, HOG-PCA, HandHOG) outperform local feature histograms (BlockHOG-SIFT). While different hand grasps may share similar statistics of local gradient patterns, it can be observed that global gradient information is important for robust classification. Although the separation between hand and object in HandHOG seems intuitive and well-motivated, HandHOG performs slightly worse than HOG in nearly all cases. This is in part because of the hand segmentation noises, but also because HOG encodes additional information about the appearance of the object being held. The big performance gap between SVM-linear and SVM-rbf, especially when using HOG-PCA, indicates that hand grasps have wide variance in pose and are therefore not linearly separable. More importantly, the experimental results show that it is possible to construct high performance vision-based task-specific classifiers for a single user.

Table 2.1: Performance of single user on UT Grasp Dataset

	SVM-linear	SVM-rbf	ESVM
HOG	0.85	0.86	**0.89**
HOG-PCA	0.79	0.88	**0.89**
HandHOG	0.8	0.85	**0.88**
BlockHOG-SIFT	0.79	**0.8**	0.79

Table 2.2: Performance on Machinist Grasp Dataset

	SVM-linear	SVM-rbf	ESVM
HOG	0.31	0.37	**0.39**
HOG-PCA	0.18	**0.42**	0.38
HandHOG	0.32	**0.38**	0.34
BlockHOG-SIFT	0.29	**0.39**	0.37

The grasp recognition performance on Machinist Grasp Dataset using 5-fold cross validation is shown in Table 2.2. Note that the dataset contains nearly eight hours of video data recording a machinist's daily work, thus it provides a good platform to evaluate how our vision-based approach works under real-world conditions. The combination of HOG-PCA and SVM-rbf achieves the best average F1 of 0.42, the average F1 for classification of 17 classes is 0.06 at the chance level. Although the absolute performance is still low, I believe that the result demonstrates the potential of automatic visual classification of grasp types in a realistic setting.

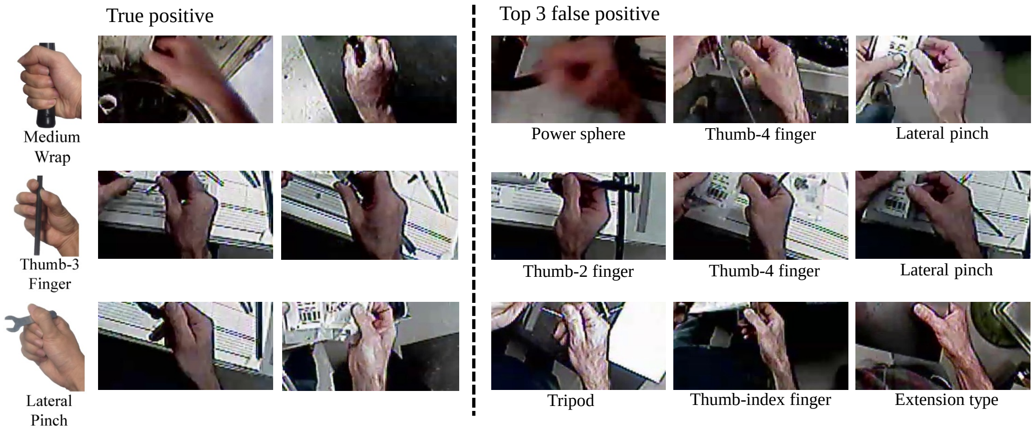
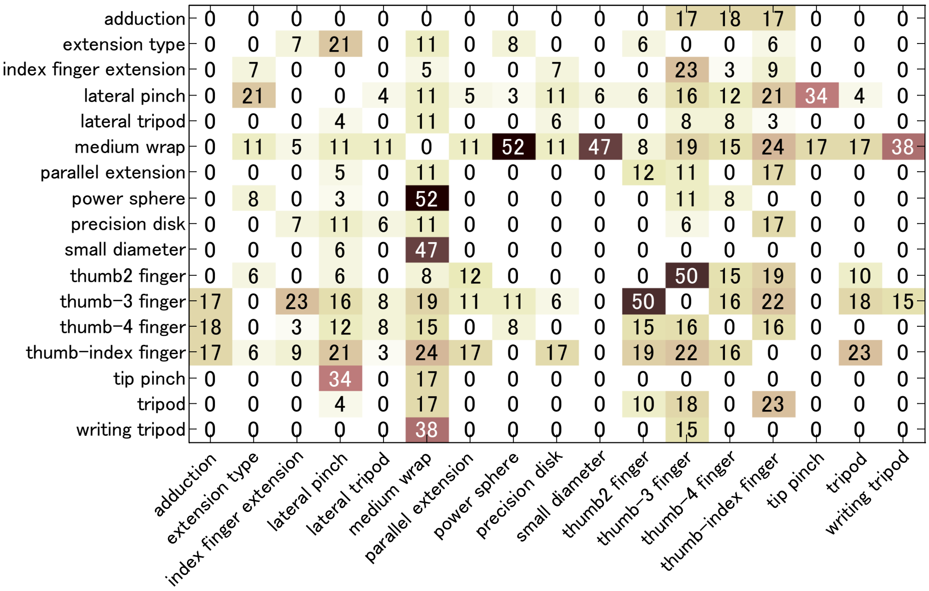


Figure 2.8: Examples of true positives and false positives on Machinist Grasp Dataset.

Some examples of true positives and false positives are shown in Fig. 2.8. Two columns to the left of the dashed line show true positives of a grasp type of which the prototype is illustrated in the left-most column. The false positives are shown in the right side of Fig. 2.8. From these examples, it can be seen that some grasp types are extremely difficult to differentiate, even for human annotators. Taking *Thumb-3 Finger* for example, both of the first true positive and the first false positive show the machinist's hand holding a tool. It is hard to tell how many fingers are used in holding the tool only from visual perception.

The visual similarity between some pairs of grasp types (*e.g.*, *Thumb-2 Finger* and *Thumb-3 Finger*) poses big challenges in training discriminative grasp classifiers based on visual features. Differentiating between fine-grained categories such as these will require more advanced vision-techniques for extracting exact finger positions. This is left to my future work.

2.4.3 Appearance-based grasp structures



	adduction	extension type	index finger extension	lateral pinch	lateral tripod	medium wrap	parallel extension	power sphere	precision disk	small diameter	thumb2 finger	thumb−3 finger	thumb−4 finger	thumb−index finger	tip pinch	tripod	writing tripod
adduction	0	0	0	0	0	0	0	0	0	0	0	17	18	17	0	0	0
extension type	0	0	7	21	0	11	0	8	0	0	6	0	0	6	0	0	0
index finger extension	0	7	0	0	0	5	0	0	7	0	0	23	3	9	0	0	0
lateral pinch	0	21	0	0	4	11	5	3	11	6	6	16	12	21	34	4	0
lateral tripod	0	0	0	4	0	11	0	0	6	0	0	8	8	3	0	0	0
medium wrap	0	11	5	11	11	0	11	52	11	47	8	19	15	24	17	17	38
parallel extension	0	0	0	5	0	11	0	0	0	0	12	11	0	17	0	0	0
power sphere	0	8	0	3	0	52	0	0	0	0	0	11	8	0	0	0	0
precision disk	0	0	7	11	6	11	0	0	0	0	0	6	0	17	0	0	0
small diameter	0	0	0	6	0	47	0	0	0	0	0	0	0	0	0	0	0
thumb2 finger	0	6	0	6	0	8	12	0	0	0	0	50	15	19	0	10	0
thumb−3 finger	17	0	23	16	8	19	11	11	6	0	50	0	16	22	0	18	15
thumb−4 finger	18	0	3	12	8	15	0	8	0	0	15	16	0	16	0	0	0
thumb−index finger	17	6	9	21	3	24	17	0	17	0	19	22	16	0	0	23	0
tip pinch	0	0	0	34	0	17	0	0	0	0	0	0	0	0	0	0	0
tripod	0	0	0	4	0	17	0	0	0	0	10	18	0	23	0	0	0
writing tripod	0	0	0	0	0	38	0	0	0	0	0	15	0	0	0	0	0

Figure 2.9: Correlation matrix of 17 grasp types on Machinist Grasp Dataset.

Here it is shown how the correlation between visually trained grasp classifiers can be used to discover the visual structures of hand grasps. The correlation index between all pairs of grasp types for Machinist Grasp Dataset is computed based on classification results using combination of HOG-PCA and SVM-rbf. The correlation matrix of 17 grasp types is shown in Fig. 2.9, where each element indicates the correlation index (scaled by 100 for visualization) between a pair of grasp types indexed by rows and columns. Top 5 pairs of grasp types with highest correlation index are shown in Fig. 2.10.

Following the iterative grasp clustering algorithm described in Algorithm 1, a dendrogram of grasp types was constructed by iteratively clustering two most correlated grasp types after each iteration of supervised learning. A dendrogram is a binary tree which gives a complete graphical description of the hierarchical clustering. The final constructed grasp dendrogram is

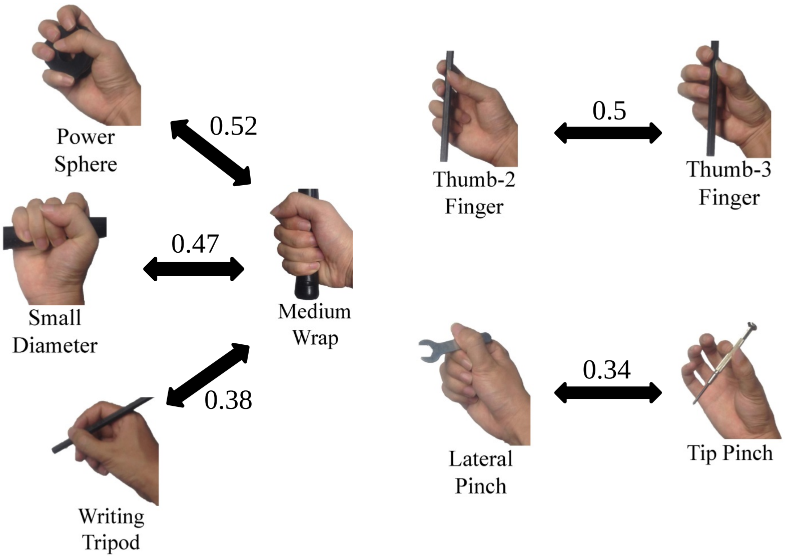


Figure 2.10: Top 5 grasp correlations on Machinist Grasp Dataset.

shown in Fig. 2.11. Grasp types with the highest classifier correlation are clustered first at lower level nodes, while those dissimilar with each other are clustered later at higher levels in the tree. The original grasp types from Feix's taxonomy are located at the leaf nodes (level-0). It can be observed that for the first six iterations, grasps are clustered in a manner consistent with known divisions of power and precision grasps in expert-designed grasp taxonomies[Cut89][FPS+09]. With the exception of *Writing Tripod* and *Extension Type*, the division between power and precision grasps are preserved until level-12 (the 12-th iteration) of the grasp hierarchy.

The more important observation however is that the visual structures of hand grasp for the machinist has been learned automatically in a data-driven manner. While classical grasp taxonomies have been created through deep introspection, the shared uncertainty between visual classifiers can also be used to learn intuitive hierarchies over human grasps.

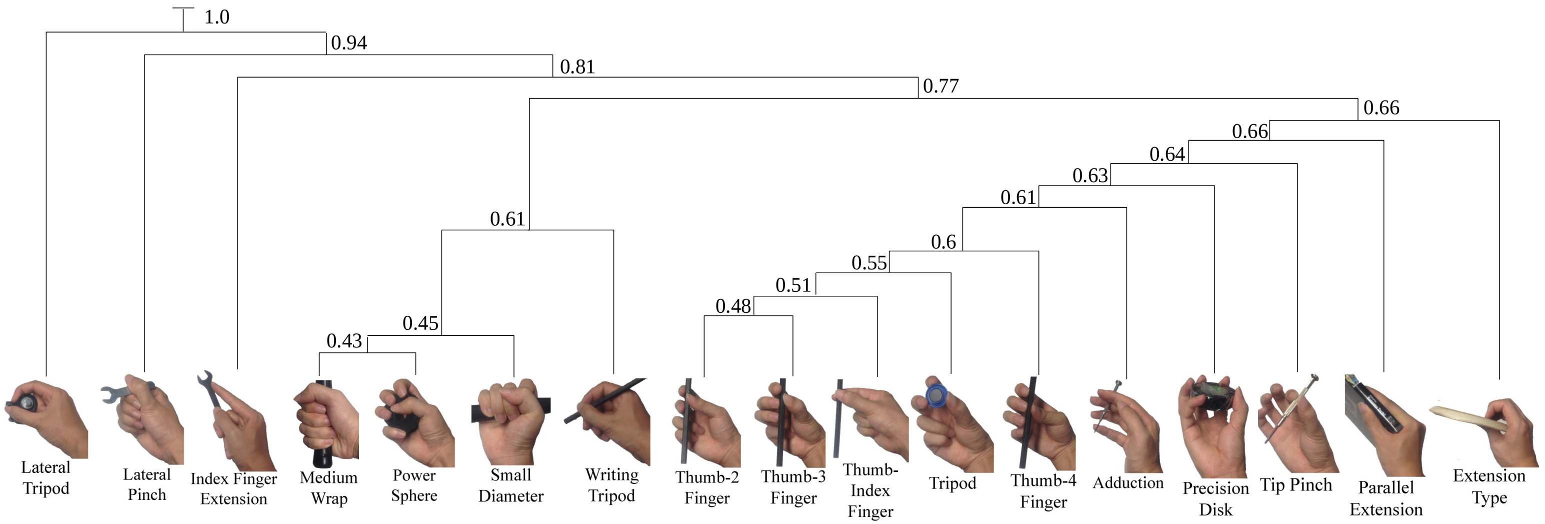


Figure 2.11: Dendrogram of hand grasp types based on hand appearance. Average F1 scores computed at different abstraction levels are added near each clustering node.

2.4.4 Recognition using grasp abstractions

Based on the dendrogram in Fig. 2.11 it is possible to 'cut' the tree at different levels to obtain different set of grasp clusters. Furthermore, each slice (abstraction) level can be interpreted as a new grasp taxonomy. By learning new grasp classifiers for each category of the new taxonomies, a trade-off between more detailed classification and more robust classification can be achieved. Average F1 scores are computed for grasp recognition at each level of grasp abstractions in Fig. 2.11. If we utilize a higher level of the tree to define grasp categories, more reliable grasp classification can be obtained. For example, at level-12 of the tree, it is able to differentiate between 5 grasps with an average F1 score of 0.66. On the other hand, choosing level-5 will allows us to differentiate between 12 grasps with an average F1 score of 0.55.

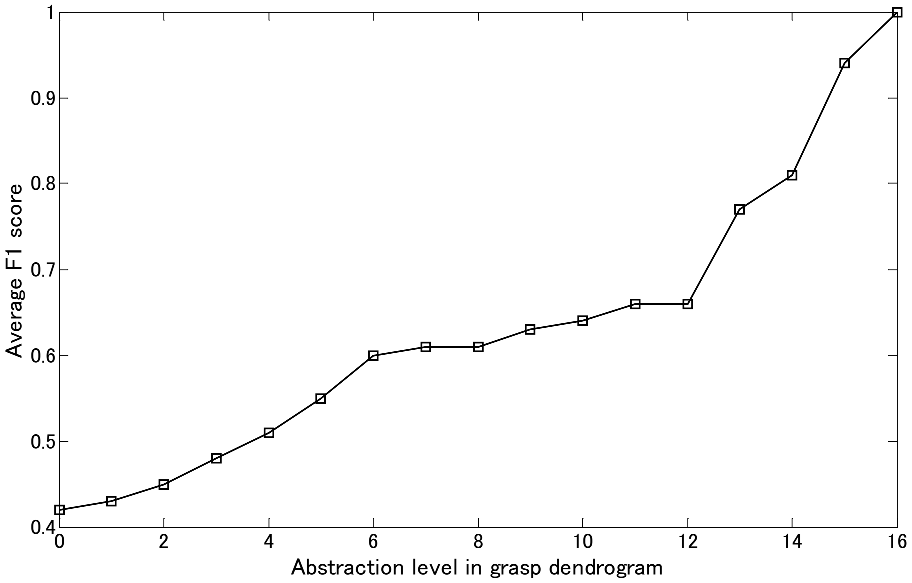


Figure 2.12: Grasp recognition performance at different levels of grasp abstractions. Performance at different abstraction levels shows a trade-off between more detailed classification and more robust classification.

The changes of grasp recognition performance at different levels of the grasp dendrogram is shown in Fig. 2.12. The average F1 grows up steadily until level-6 since at initial six iterations similar grasp types are being clustered together. From level-7 to level-12, average F1 increases relatively slowly compared to previous steps. For example, average F1 of level-11 and level-12 are almost the same (0.66). This can be explained as newly clustered grasp types become more dissimilar and thus only limited improvement of recognition performance is achieved. Average F1 increases dramatically from level-13 since big grasp clusters are merged together and chance of misclassification is low.

This learned visual structure gives researchers the flexibility of finding a good balance between better performance and more detailed grasps analysis.

2.5 Conclusion

In this chapter, I propose a first-person vision-based approach for automatic grasp analysis from image appearance. In the approach, discriminative classifiers are trained to recognize different grasp types based on computer vision techniques, and visual structures of hand grasps are learned by a supervised grasp clustering method. This work shows the potential for using computer vision techniques for analyzing hand grasps with large scale of data in real-life settings.

There still exists a lot of work to do to improve grasp recognition performance. The temporal aspect of grasping is obviated in this paper and it would be helpful to impose temporal coherence to improve classification performance. Moreover, explicit object attributes such as weight, shape and size are important factors affecting human grasp selection. I believe a reliable detection framework of object attributes would be very useful in inferring grasp usage. These problems will be addressed in next chapters.

Chapter 3

Hand grasp analysis with dynamic appearance features

3.1 Introduction

In the previous chapter, I proposed an image-based method for hand grasp analysis from static image appearance within a first-person vision framework. Although the proposed method can recognize different grasp types when only an image is given, the recognition performance is not accurate enough, especially in real world scenario. Different grasp types which share similar hand shape/appearance are ambiguous to be differentiated only from a single image. Even hand appearance of one grasp type might be dynamically changing during interactions, making image appearance alone insufficient for accurate grasp recognition. Furthermore, it is sometimes challenging to reliably detect the hand and the appearance-based method is sensitive to hand detection noises. To address these problems, a more compact and richer feature representation which encode dynamical information of hand interactions is desired.

Instead of the case when only an image is used as in Chapter 2, I consider another case when an image sequence with consecutive frames are available. This enables us to study hand grasp from a different perspective, that is, from hand dynamics by which it means dynamical information of hand appearance and motion during interactions. In particular, a feature representation based on hand-guided feature tracking is proposed and called as "Dense Hand Trajectories" (DHT). Dense hand trajectories are obtained by densely sampling feature points and tracking them within a short video interval and is guided by hand detection. What makes it different from traditional dense trajectories is that each tracked trajectory is given a weight based on its spatial relations with detected hand regions. Trajectories with low weight are discarded, and feature descriptors are computed for each trajectory to encode the information of both hand motion and hand appearance. Features based on dense hand trajectories have several advantages over appearance-based features. First, trajectory itself contains motion information of the hand during interaction which is useful for identifying different grasp types. Second, hand appearance at multiple adjacent images along the hand trajectory can be computed as more compact representation for single grasp type. Moreover, grasp classifiers trained on trajectory-based features are more robust to hand detection noises.

In addition, to better evaluate the visual grasp structures automatically learned from data, I propose a new metric to quantitatively compare different hierarchical grasp structures. Quantitative evaluation with qualitative comparison demonstrate the consistency of automatically learned grasp structures with expert-designed grasp taxonomies.

Contributions of this chapter are summarized as follows: (1) A new feature representation for grasp recognition from image sequences is proposed which achieves best classification accuracy and is robust to unreliable hand detection. (2) A new metric is proposed to quantitatively evaluate the consistence of the automatically learned grasp structures with expert-designed grasp taxonomies. (3) The performance of the grasp recognition system is extensively evaluated by examining state-of-the-art feature representation used in object and action recognition.

The rest of this chapter is organized as follows. Section 3.2 presents related work. Section 3.3 introduces the proposed feature representation based on dense hand trajectories. Performance evaluation of the system is shown in Section 3.4. Section 3.5 discusses the advantages of proposed method. Conclusions of the work is made in Section 3.6.

3.2 Related works

3.2.1 Vision-based grasp recognition

There exist few previous studies on vision-based grasp recognition. Work from Cai et al. [CKS15] first developed techniques to recognize a complete set of hand grasp types in everyday hand manipulation tasks recorded with a wearable RGB camera and provided promising results with appearance-based features. Yang et al. [YLFA15a] utilized a convolutional neural network to classify hand grasp types on unstructured public dataset and presented the usefulness of grasp recognition for action understanding. However, it only considers a small number of grasp types trained on static scene hand images. Saran et al. [STK15] used detected hand parts as intermediate representation to recognize fine-grained grasp types. The intermediate representation outperforms low-level appearance-based representation when hand parts can be well detected. In this work hand grasp types are recognized from perspective of hand dynamics in order to tackle the challenges of unreliable hand detection.

3.2.2 Dense trajectories

Dense trajectories proposed by [WKSL11] have become one of predominant feature representation for video recognition. The main idea is to densely sample feature points at each frame and track them for an amount of time in the video using optical flow. Multiple descriptors encoding appearance and motion information are computed along the trajectories of feature points. Several approaches are proposed to improve dense trajectories. Vig et al. [VDC12] employed saliency-mapping algorithms to address the descriptors corresponding to informative regions. This space-variant method improves action recognition accuracy with a more compact video representation. Wang and Schmid [WS13] improved dense trajectories by removing trajectories consistent with camera motion and cancel the camera motion from optical flow for motion-based descriptors. In this work, when estimating the camera motion, only feature points between frames which are beyond the hand region are matched since hand motion is in general different from camera motion in first-person videos.

Baraldi et al. [BPS$^+$14] proposed to use dense trajectories with hand segmentation for hand gesture recognition in ego-vision scenarios. Dense trajectories which is often used in action recognition is proved to work well in egocentric paradigm. The proposed dense hand trajectories is similar to the work of [BPS$^+$14] but with the differences as follows: First, hand detection is utilized to weight the tracked feature points in order to give flexible evaluation of the trajectories' relatedness to hand interactions. Second, only feature descriptors from trajectories which have high relatedness to hand interactions are extracted.

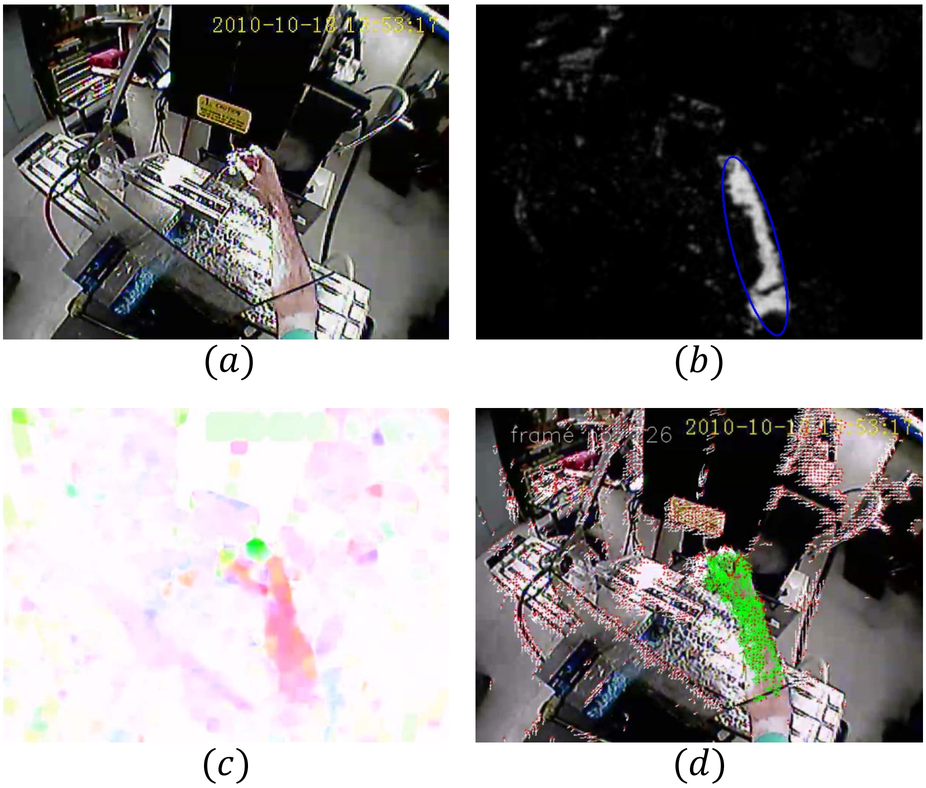
3.3 Proposed method

3.3.1 Dense hand trajectories

Dense trajectories proposed by Wang et al. [WKSL11] have been widely used as video representation for action recognition, and proven to achieve state-of-the-art results on many video datasets of third person view. To apply it to grasp recognition in first person video, it is important to focus on the region where hand interaction occurs and remove irrelevant features from background. Motion-based background subtraction doesn't work well in first person video since the background is moving and is hard to reliably estimate the camera motion as illustrated in Figure 3.1(c). In this work, I propose a feature representation of "Dense Hand Trajectories (DHT)" which uses hand detection as a spatial prior to extract dense trajectories most related to hand interactions.

First the extraction procedure of traditional dense trajectories [WKSL11] is described on which dense hand trajectories is based. At each frame, feature points are densely sampled on a grid spaced by 5 pixels at multiple spacial scales. Points in homogeneous area are removed since it is impossible to track them without any structure. Feature points at each spacial scale are tracked separately using a dense optical flow algorithm [Far03]. Each trajectory is composed by feature points tracked for consecutive frames with trajectory length set to $L = 15$ frames.

The main difference of the proposed DHT from [WKSL11] is that the detected hand regions are used as spatial prior to weight trajectories which pass through the hand regions. Specifically, a variable H is used to count



(a) (b) (c) (d)

Figure 3.1: Example of dense hand trajectories. (a) Image from egocentric video (b) Hand probability map (c) Visualization of optical flow (d) Visualization of dense hand trajectories in green color

the times of being tracked within the hand regions for each trajectory as illustrated in Figure 3.2. At each frame t, a trajectory with a starting feature point sampled within the hand region is initialized with $H = 1$ as indicated by the trajectory (a), otherwise it is initialized with $H = 0$ as indicated by the trajectory (b). At each subsequent frame during the tracking procedure, H is increased by 1 for all trajectories of which the feature points being tracked are within the hand regions. At the end of tracking, trajectories with H less than a certain threshold T_h are considered as non-hand trajectories and thus removed. In the experiments, I set $T_h = L/2$ based on empirical results.

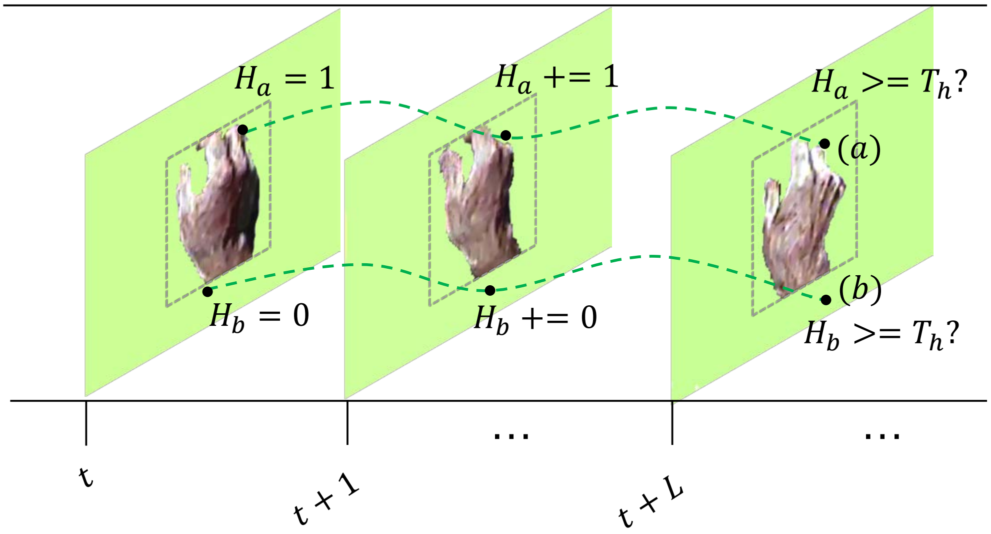


Figure 3.2: Illustration of my approach to extract dense hand trajectories.

3.3.2 Feature extraction

There are two stages of feature extraction based on dense hand trajectories. At the first stage, descriptors are computed for each trajectory. At the second stage, descriptors of trajectories are pooled together and further encoded for each frame.

At the first stage, four descriptors (Displacement, HOG, HOF, MBH) are computed the same as in [WS13]. Dimensions of these descriptors are 30 for Displacement, 96 for HOG, 108 for HOF and 192 for MBH. These descriptors contains information of both hand motion and hand appearance in the space-time volume along the trajectory. The Displacement descriptor captures pixel displacement along the trajectory, HOG are based on the orientation of image gradient and encode the static appearance of the region surrounding the trajectory, HOF and MBH are based on optical flow and capture motion information. Homography estimation between consecutive frames is also used to remove global camera motion as in [WS13]. The difference is that hand segmentation mask is used to discard the feature matches within hand regions since hand motion is not consistent with camera motion in first-person view videos.

At the second stage, Fisher vector is used to encode pooled trajectory descriptors for each frame. Fisher vector has shown performance improvement over bag-of-features for image/video classification in recent researches. For details of Fisher vector encoding, one can refer to [PSM10]. Principal Component Analysis (PCA) is first used to reduce the dimension of each descriptor type to $D = 16$, and then a randomly sampled subset of $300,000$ features are used to estimate the Gaussian Mixture Model (GMM) with number of Gaussians set to $K = 256$, as in [PSM10]. The dimension of each descriptor type after Fisher vector encoding is $2DK$. Each frame is represented by concatenation of Fisher vectors of different descriptor types. The procedure of feature aggregation based on Fisher vector is summarized in Algorithm 2.

3.4 Evaluation

Like in previous chapter, system performance is evaluated on two datasets: UT Grasp Dataset and Machinist Grasp Dataset. Recognition performance of six different features is examined in the system. Four features (HoG, HHoG, SIFT, CNN) rely on hand patches of fixed size. In the experiments, hand patches are segmented with a bounding box with the size of 160×160 for UT Grasp Dataset and 128×128 for Machinist Grasp Dataset. HoG and HHoG are computed on hand patches resized to 80×80 and the feature dimension is 2916. The feature dimension of SIFT is 100 since it is encoded using BoW with 100 dictionary entries. Features based on CNN are

Algorithm 2 Feature extraction (the second stage)

```
Initialize: $D \Leftarrow$ feature dimension after PCA, $K \Leftarrow$ Number of Gaussians for GMM, $T \Leftarrow$ Number of consecutive frames for trajectory pooling
Training: Estimate PCA with $D$ retained components for each descriptor type (Displacement, HOG, HOF, MBH) from trajectory descriptors in training data; then estimate GMM with $K$ Gaussians for dimension-reduced descriptors after PCA
for all frame $t$ do
    Step1: Pool together descriptors of all trajectories ended within $[t, t+T]$
    Step2: Perform Fisher vector encoding for each descriptor type separately
    Step3: Concatenate Fisher vector of different descriptor types as feature descriptor for current frame
end for
```

extracted from each hand patch using the Caffe implementation [JSD+14] of the CNN model proposed by Krizhevsky et al. [KSH12]. Each hand patch is forward propagated through five convolutional layers and a fully connected layer and the feature dimension is 4096. Another two features are based on dense trajectories. Improved Dense Trajectories (IDT) proposed by Wang and Schmid [WS13] improves dense trajectories by removing camera motion between two consecutive frames. Dense Hand Trajectories (DHT) is the proposed feature. Both IDT and DHT are encoded using Fisher vector with same parameter settings and the feature dimension is 32768. Note that for fair comparison, appearance-based features are aggregated from adjacent ($L = 15$) frames using the same aggregation scheme as in the computation of trajectory descriptors [WKSL11].

Linear SVMs are trained for each hand grasp type using the obtained features mentioned above. The implementation of LIBSVM [CL11] is used for training. At test time, each frame with detected hand region is assigned to a grasp type of which the classifier obtains the highest score. The classification accuracy is used for evaluating the grasp recognition performance.

3.4.1 Performance comparison

The proposed approach is applied to UT Grasp Dataset and Machinist Grasp Dataset to see how visual features can discriminate between different grasp types in both controlled and natural environments.

First grasp recognition results are presented for a single user on UT Grasp Dataset. Grasp classifiers are trained and tested for each user using 5-fold cross validation. Recognition performance of 17 grasp types are shown in Table 3.1 for different feature representations. Precision and recall for each grasp type and accuracy for overall performance are shown in the table.

Here only nine most frequent grasp types according to sample proportion are shown due to the space limit. From Table 3.1, it can be seen CNN-based feature and proposed DHT achieve best accuracy of 94%. As for the four features (HoG, HHoG, SIFT, CNN) which rely on exact hand patches, the best performance achieved by CNN indicates the importance of high level features in robust classification. The lowest performance from SIFT indicates local appearance-based feature alone is less discriminative than global features. Although the separation between hand and object in HHoG seems intuitive and well-motivated, HHoG performs worse than HoG. This is in part because of the hand segmentation noises, but also because HoG encodes additional information about the appearance of the object being held. As for the two trajectory-based features, better performance of the proposed DHT over IDT demonstrates the effect of removing unrelated information from the background. Although DHT doesnot outperform CNN as expected, I believe this is because the motion information contained in DHT doesn't help in the controlled environment. More importantly, the experimental results show that it is possible to construct high performance vision-based task-specific classifiers for a single user.

Table 3.1: Performance comparison of proposed features and appearance-based features on UT Grasp Dataset. Precision (P) and Recall (R) are performance metrics for each grasp type. Accuracy is for overall performance. Number within parentheses aside each grasp type indicates sample proportion.

	T (.11)		LP		LT		IFE		T4F		PS		MW		A (.05)		ET		Total
			(.09)		(.09)		(.07)		(.07)		(.06)		(.06)				(.05)		
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	Accu.
HoG	.84	.88	.85	1.0	.98	1.0	1.0	1.0	.66	.96	.61	.65	.85	.71	1.0	.96	.96	.92	.87
HHoG	.69	.79	.93	.98	1.0	1.0	.94	.97	.89	.86	.68	.58	.72	.75	.81	.92	.86	.75	.82
SIFT	.58	.85	.63	.80	.76	.65	.93	.93	.61	.79	.47	.35	.83	.79	1.0	.92	.44	.22	.70
CNN	1.0	1.0	.95	.98	1.0	1.0	.97	1.0	.74	1.0	1.0	.96	1.0	1.0	.96	1.0	1.0	1.0	.94
IDT	.96	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.73	.96	1.0	.85	.91	.83	1.0	.92	.92	1.0	.92
DHT	.94	1.0	1.0	1.0	1.0	1.0	1.0	1.0	.85	1.0	.95	.73	.82	1.0	1.0	1.0	1.0	1.0	.94

Table 3.2: Performance comparison of proposed features and appearance-based features on Machinist Grasp Dataset. Precision (P) and Recall (R) are performance metrics for each grasp type. Accuracy is for overall performance. Number within parentheses aside each grasp type indicates sample proportion.

	MW (.20)		LP (.19)		LT (.12)		T3F (.11)		TIF (.11)		T4F (.07)		T2F (.05)		PS (.04)		IFE (.03)		Total
	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	P	R	Accu.
HoG	.35	.37	.48	.67	.38	.53	.17	.14	.17	.23	.29	.06	.15	.09	.08	.06	.86	.69	.34
HHoG	.32	.37	.39	.49	.38	.58	.18	.14	.20	.20	.09	.02	.06	.02	.06	.06	.33	.42	.29
SIFT	.19	.21	.43	.62	.26	.41	0.0	0.0	.04	.01	0.0	0.0	0.0	0.0	.20	.06	.27	.69	.24
CNN	.59	.56	.59	.74	.64	.77	.26	.23	.31	.35	.26	.25	.21	.21	.41	.28	.70	.73	.49
IDT	.63	.60	.68	.84	.80	.95	.19	.16	.39	.46	.33	.28	.20	.19	.76	.69	.83	.73	.54
DHT	.69	.71	.65	.86	.88	.94	.24	.22	.46	.49	.40	.38	.32	.40	.69	.63	.95	.77	.59

Grasp recognition performance of different features on Machinist Grasp Dataset using 5-fold cross validation is shown in Table 3.2. The proposed DHT achieves highest classification accuracy of 59% compared to other baseline features. It is reasonable the proposed DHT works better than IDT since irrelevant trajectory information from background has been removed using hand detection. CNN-based feature improves the performance by over 15% compared to HoG, which verifies the superiority of high-level features over hand-crafted features. Also it is clear that trajectory-based features (DHT, IDT) outperform appearance-based features (CNN, HoG), partly because hand motion information is also captured in trajectory-based features which can help discriminate different grasp types.

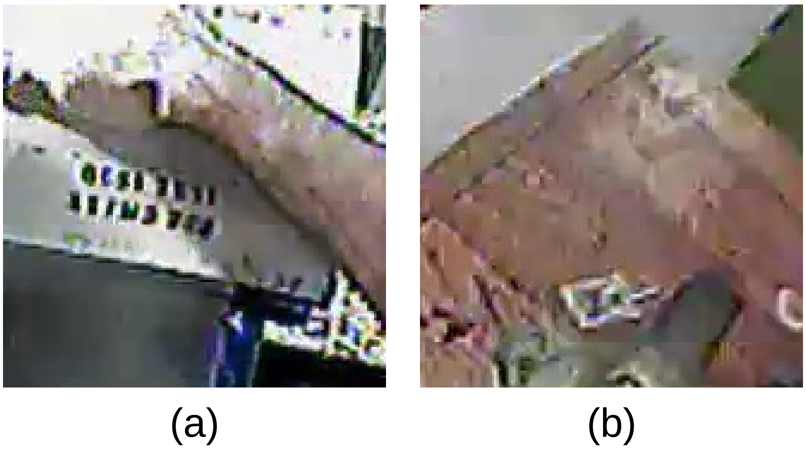


Figure 3.3: Examples of unreliable hand detection. (a) Incomplete hand detection with fingers missing due to extreme lighting condition (b) False detection from background with similar skin color

I believe the robustness to unreliable hand detection of trajectory-based features is another important reason why they outperform appearance-based features. Hand detection in real-world first-person video is sometimes unreliable due to extreme imaging conditions such as changing background and extreme hand motion. Fig. 3.3 shows some examples of bad detection. Grasp recognition relying on appearance-based features might be heavily influenced by unreliable hand detection. To evaluate the influence of hand detection, classification accuracy under different hand detection conditions are also compared. For ideal detection, image samples are manually selected in which automatic hand detection results are acceptable. For real detection, all image samples are used. The results are shown in Table 3.3. There is a performance drop from ideal detection to real detection for HoG and CNN, which indicates appearance-based features are sensitive to hand detection. However, IDT and DHT are robust to hand detection with even slight performance improvement under real detection. I believe the reason resides on the feature tracking procedure through which IDT and DHT are extracted since feature tracking is independent on hand detection. And more training data under real detection results in further performance improvement.

Table 3.3: Performance influences by hand detection

	Ideal detection	Real detection
HoG	40.8%	33.9%
HHoG	32.5%	29.4%
SIFT	27.1%	23.8%
CNN	52.4%	48.5%
IDT	52.3%	54.3%
DHT	57.9%	59.2%

3.4.2 Learning and comparing grasp structures

Here shows the visual structures of hand grasp learned based on the proposed DHT. The correlation index between all pairs of grasp types is computed for Machinist Grasp Dataset based on classification results of grasp classifiers trained on DHT. Bad hand detection samples are removed from training data in order to make the correlation between classifiers more likely reflect the visual similarity of hand grasps.

Following the iterative supervised clustering algorithm described in Algorithm 1, a dendrogram of grasp types based on DHT is constructed and is shown in Fig. 3.4. Grasp types with the highest classifier correlation are clustered first at lower level nodes, while those dissimilar with each other are clustered later at higher levels in the tree. The original grasp types from Feix's taxonomy are located at the leaf nodes (level-0). It can be observed that grasps are clustered in a manner consistent with known divisions of power and precision grasps in expert-designed grasp taxonomies[Cut89][FPS+09]. With the exception of *Precision Disk* and *Extension Type*, the division between power and precision grasps are preserved until level-12 (the 12-th iteration) of the grasp hierarchy. There are five groups of grasp types remained at level-12. One group ranging from *Medium Wrap* to *Power Sphere* represents the power grasps characterized by stably holding an object with palm and five fingers. In contrast, the group ranging from *Thumb-4 Finger* to *Adduction* represents the precision grasps which can be used to flexibly manipulate an object with dexterous finger articulation. Another interesting group represented by *Lateral Pinch* and *Writing Tripod* stands intermediately between power and precision grasps where both stability and dexterity are addressed. These qualitative examples show that the proposed DHT can also discover grasp relationships consistent with parts of expert-designed taxonomy.

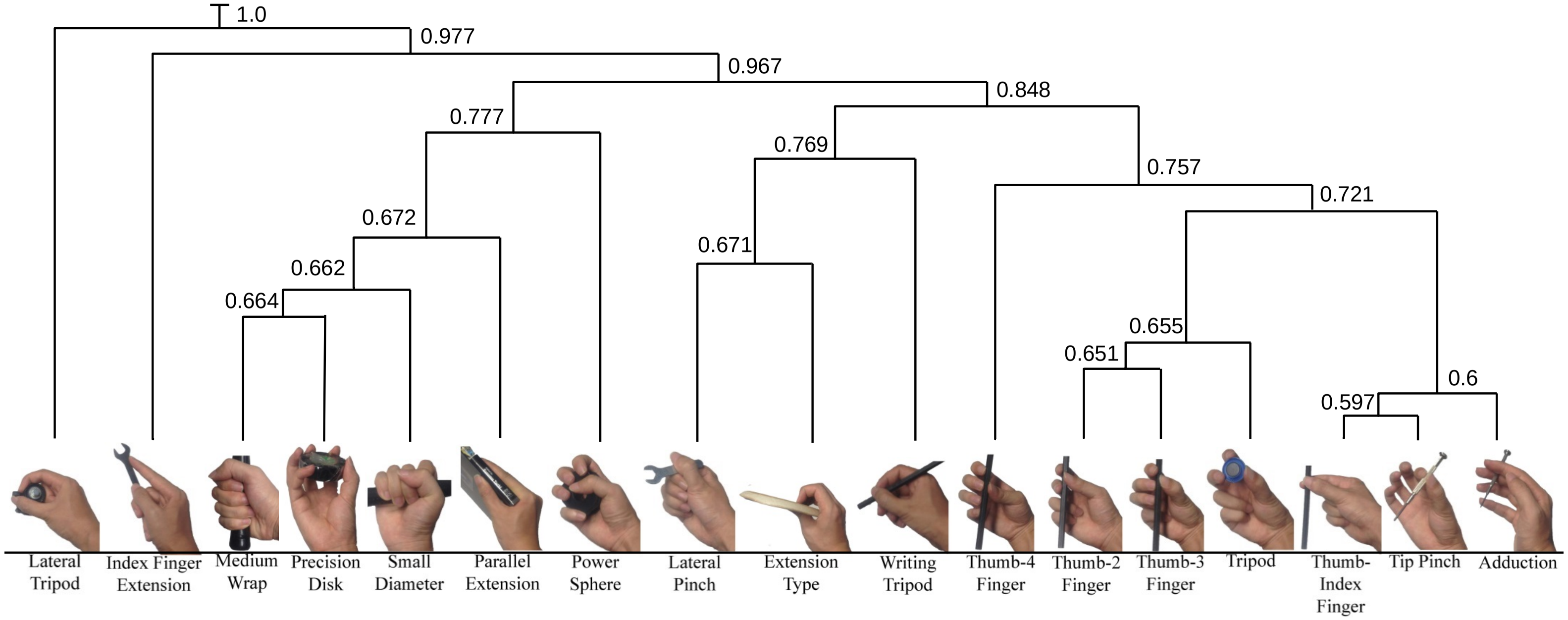


Figure 3.4: Dendrogram of hand grasp types learned based on the proposed DHT. Accuracy computed at different abstraction levels are added near each clustering node.

Although it has been shown from qualitative evaluation that intuitive visual structures of hand grasp can be learned automatically in a data-driven manner, there is no quantitative evaluation on these automatically learned grasp structures. To have a quantitative comparison of different hierarchical grasp taxonomies, I propose a mew metric called Normalized Common Distance (NCD) score. The NCD is composed as:

$$NCD(T_a, T_b) = \sum_{l_A, l_B \in T_a, T_b} |\frac{d_a(l_A, l_B)}{H_a} - \frac{d_b(l_A, l_B)}{H_b}| \tag{3.1}$$

where l_A and l_B are leaf nodes with labels of A and B respectively, H_a and H_b are maximum depth of tree T_a and T_b, and $d(*, *)$ is the Lowest Common Ancestor [DPZ91] distance between two nodes. In our case, trees are hierarchical grasp taxonomies and labels A and B are grasp labels from the taxonomy. The proposed NCD metric is necessary for comparing tree structures with different branches and depth and uncommon terminal nodes.

Table 3.4: Distance between Cutkosky's taxonomy and the automatically learned grasp structures based on three features (HoG, CNN, DHT).

Tree pair	NCD score
(T_{ref},T_{hog})	16.1
(T_{ref},T_{cnn})	18.8
(T_{ref},T_{dht})	15.9
(T_{hog},T_{cnn})	9
(T_{cnn},T_{dht})	14.6
(T_{dht},T_{hog})	13.7

Taxonomy trees are built automatically based on three different features (HoG, CNN, DHT) and compared to the reference tree based on Cutkosky's taxonomy tree. The automatically learned trees are also compared between themselves. The NCD scores are shown in Table 3.4. The tree based on DHT has the smallest NCD score and is most similar to Cutkosky's taxonomy tree. More important observation is that the tree based on HoG has slightly bigger NCD score to the reference tree than the tree based on DHT, which means low-level appearance-based feature can also learn meaningful grasp relationships. The NCD scores of comparing between the trees based on three features indicate the automatically built trees are actually very similar to each other.

3.4.3 Performance comparison at different abstraction levels

As stated in previous chapter, the learned grasp structures give researchers the flexibility of finding a good balance between better performance and more fine-grained grasp classification. Here the grasp classification accuracy of different features at different abstraction levels is shown to give a better glance of trade-off between categorization and robustness.

The changes of grasp recognition performance for HoG, CNN and DHT at different levels of the grasp dendrogram is shown in Fig. 3.5. As expected, the classification accuracy for all three features grows up steadily as the abstraction level increases. From level-12 the accuracy increases dramatically since big grasp clusters are merged together and chance of misclassification is low. Moreover, the big performance gap among the three features at lowest level (fine-grained classification) becomes smaller as abstraction level increases and inter-class ambiguity diminishes.

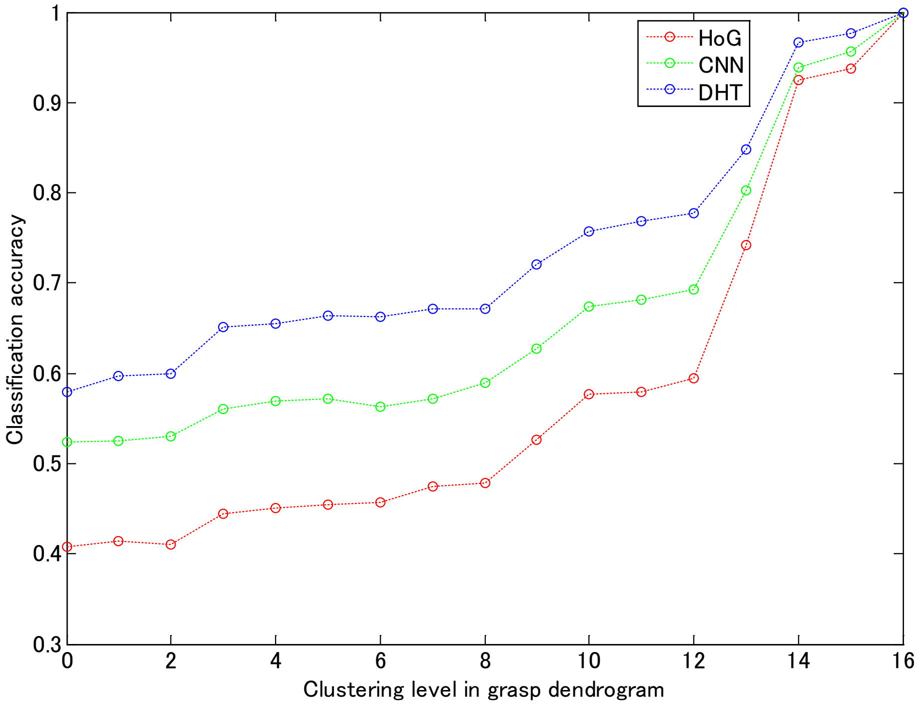


Figure 3.5: Grasp classification accuracy of different features at different levels of grasp abstractions.

3.5 Discussion

In this work, a new feature representation based on dense hand trajectories is used to improve grasp recognition. The advantages of hand trajectories-based features over appearance-based features are: (1) Hand trajectories capture motion information of the hand and thus encode richer information than appearance only. (2) Hand trajectories-based features are more robust to segmentation error. In this section, I will discuss in more details on how dense hand trajectories improve the grasp recognition on the two aspects.

To demonstrate how motion information helps improve grasp recognition, recognition performance based on different components of dense hand trajectories are compared in Table 3.5. Experimental setting is the same

Table 3.5: Recognition performance based on different components of dense hand trajectories.

Component	Accuracy
HoG	55.1%
Disp.	35.1%
HoF	38.4%
MBH	44.7%
Disp.&HoF&MBH	49.5%
All	59.2%

as in Chapter 3. The HoG component which represents the appearance part achieves an accuracy of 55.1%, while the combined components of Disp., HoF and MBH which represent the motion part achieves an accuracy of 49.5%. By combining appearance part and motion part together, the recognition performance is improved by 4.1% compared to using appearance part only. Note that the performance of using HoG component alone is still much better than using CNN-based feature (48.5% according to Section 3.4) despite both features are appearance-based. Two reasons can explain this. First, the appearance extracted along hand trajectories encodes intrinsic appearance variation within one grasp class than frame-based appearance. Second, feature extracted in the context of hand tracking is more robust to segmentation errors which is described following.

To demonstrate the robustness of dense hand trajectories to hand segmentation error, test images are divided into good and bad samples according to hand segmentation and the percentage of correct prediction on these good/bad samples are computed to measure the robustness to segmentation

Table 3.6: Recognition performance on test samples with good and bad hand segmentation. Test samples are divided into good samples and bad samples based on hand segmentation, and 18% of test samples are counted as bad samples.

	Good samples	Bad samples
HoG	39.6%	30.2%
CNN	50.8%	38.5%
DHT	60.2%	55.0%

error. Table 3.6 compares the robustness of HoG feature, CNN-based feature and dense hand trajectories (DHT)-based feature. It can be seen that the DHT-based feature has a much smaller performance drop on bad samples than HoG feature and CNN-based feature, thus the robustness of DHT to hand detection noises is verified.

3.6 Conclusion

In this chapter, a new feature representation of dense hand trajectories which encodes the hand dynamics is proposed to improve grasp recognition from consecutive image frames. Feature descriptors based on dense hand trajectories encode dynamical information of hand appearance and motion during hand interactions. Experiments show that the proposed method achieves best recognition performance and is robust to hand detection noises in real world environments.

While the recognition performance is not accurate enough in real-world scenario, this work shows the potential for using computer vision techniques for analyzing hand grasps with large scale of data in real-life settings. The proposed method achieved an classification accuracy of 59%, and the learned visual structure of hand grasps gives researchers the flexibility of finding a good balance between better performance and more detailed grasps analysis.

Chapter 4

Understanding manipulation actions with grasp types and object attributes

4.1 Background

Building on the prior work of hand grasp recognition introduced in previous chapters, this work takes a further step to study the hand manipulation in a broader scale. In particular, this work aims to recognize (1) grasp types, (2) object attributes and (3) actions from a single image within a unified model. These terms are defined as follows: *Grasp types* are a discrete set of canonical hand poses often used in robotics to describe various grasping strategies for objects. For example, the use of all fingers around a curved object like a cup is called a medium wrap. *Object attributes* characterize physical properties of the objects such as rigidity or shape. And *actions* in this work refer to different patterns of hand-object interactions such as open or pour.

The ability to understand egocentric activities (manipulation actions) automatically from images is important for domains such as robotic manipulation [Cut89, YLFA15b], human grasp understanding [FBD14], and motor control analysis [CSPA+92]. In robotic manipulation, the study of human hand function provides critical information about robotic hand design and action planning. In human grasps understanding, the recognition of hand-object manipulations enables automatic analysis of human manipulation behavior, making it more scalable than traditional manual observation used for previous studies [ZDLRD11]. Wearable cameras enable recording of hand-object manipulations at a large scale, both in time and space, and provides an ideal first-person point-of-view under which hands and objects are visible up-close in the visual field.

The recognition task for understanding manipulations from monocular images is also very challenging. There are many occlusions of the hand, especially the fingers, during hand-object interactions making it hard to observe and recognize hand grasps. It is also challenging to reliably detect the manipulated object and infer attributes since the object is also often occluded by the hand. This suggests that visual information about the hands and objects need to be reasoned about jointly by taking into account this mutual context.

In this chapter, I propose a novel method to extract object attribute information from the manipulated object without using specific object detectors by instead exploring spatial hand-object configurations. Furthermore, recognition of grasp types and object attributes is enhanced by their mutual context (contextual relationship between two components that by knowing one component facilitates the recognition of the other). Object attributes (*e.g.*, thick or long shape of a bottle) have strong constraints on the selection of hand grasp types (*e.g.*, *Large Wrap*). Thus, with the knowledge

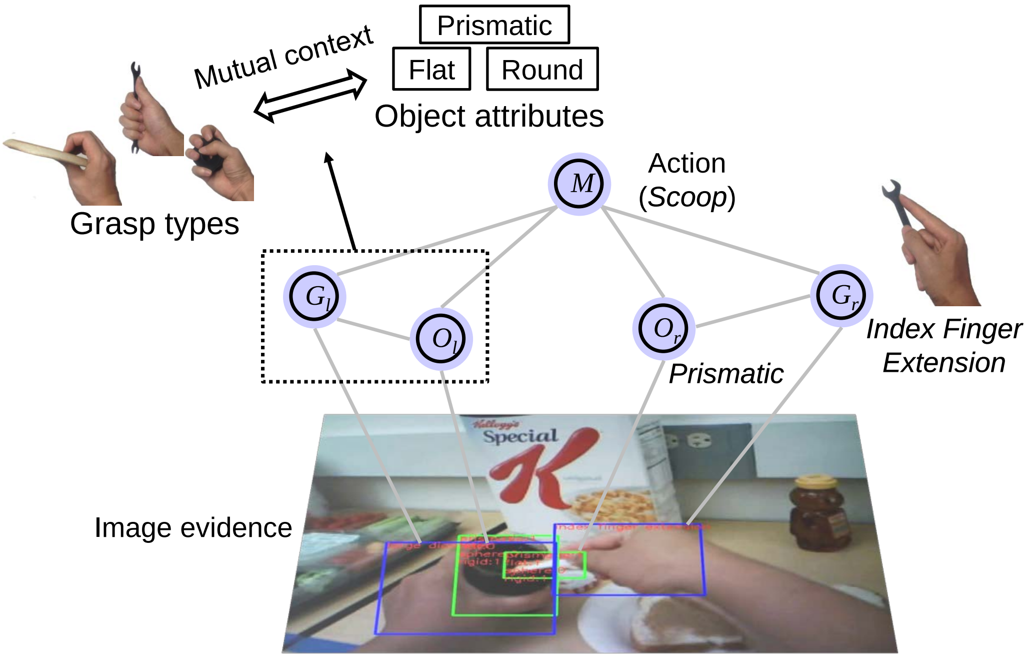


Figure 4.1: Relationship between grasp types, object attributes, and manipulation actions. Grasp types and object attributes at both hands are learned from image evidence. Mutual context between grasp types and object attributes is explored. Manipulation actions are modeled based on grasp types and object attributes.

of object attributes, it is able to predict a large percentage of grasp types. On the other hand, humans use the same or similar grasp types for certain types of objects, thus the grasp type used reveals attributes of the object being grasped. In the end, I propose a Bayesian model to encode the mutual context between grasp types and object attributes in which recognizing one facilitates the recognition of the other.

Based on the visual recognition of grasp types and object attributes, a semantic action model is provided as illustrated in Figure 4.1. Specifically, discriminative classifiers are trained for different actions based on the recognition output (belief distribution) of grasp types and object attributes.

There are several advantages for jointly modeling actions in this way: (1) Grasp type helps describe the functionality of an action, whether it requires more power, or more flexible finger coordination; (2) Object attributes provide a general description about the manipulated object and indicates possible interaction patterns; (3) High-level semantic labels of grasp type of object attributes enable the model encode high-level constraints (*e.g.*, medium wrap can only be used for cylindrical objects) and as a result, is results of the learned model are immediately interpretable.

The contributions of this work are as follows: (1) A novel method is proposed for extracting attributes of the manipulated objects without any specific object detection models; (2) The mutual context of grasp types and object attributes is explored to boost the recognition of both; (3) Semantic action model is proposed based on grasp types and object attributes which achieves state-of-the-art recognition performance.

4.2 Related works

4.2.1 Hand grasp

Hand grasps have been studied for decades to better understand the use of human hands [Nap56, SFS98, BZR+13, HMMK15]. Grasp taxonomies have also been proposed to facilitate hand grasp analysis [Cut89, KI93, FPS+09]. Approaches for vision-based hand grasp analysis were developed primarily in structured environment. Vision tracking of hand grasping an object [KRK08, HSKMVG09, OKA11, RKEK13] allows a completely non-contact markerless form of hand interactions. However, most hand tracking systems require that hand interactions are recorded in a structured environment. Rogez et al. [RIR15] recently presented promising results on discrete hand pose recognition from a chest-mounted RGB-D camera. However, these discrete poses have no direct semantic correspondence to human grasp types commonly used.

Cai et al. [CKS15] first developed techniques to recognize hand grasp types in everyday hand manipulation tasks recorded with a wearable RGB camera and provided promising results with appearance-based features. Yang et al. [YLFA15a] utilized a convolutional neural network to classify hand grasp types on unstructured public dataset and presented the usefulness of grasp types for predicting action intention. Saran et al. [STK15] used detected hand parts as intermediate representation to recognize fine-grained grasp types. However, the recognition performance is far from practical usage in real-world environment. In this work object contextual information is explored to improve the grasp recognition performance.

4.2.2 Attribute classification

Visual attributes (physical properties inferred from image appearance) are often used as intermediate representation for many applications, such as object recognition [FEHF09, LNH09, VMT$^+$14], facial verification [KBBN09], image retrieval and tagging [SFD11, PG11, ZPR$^+$14]. Lampert et al. [LNH09] performs object detection based on a human-specified high-level description of the target classes for which no training examples are available. The description consists of attributes like shape, color or even geographic information. Parikh and Graumn [PG11] explored the relative strength of attributes by learning a rank function for each attribute which can be used to generate richer textual descriptions. In this work, visual attribute information from the manipulated object are extracted as semantic information for modeling manipulation actions.

The relations between object attributes and hand grasps are widely studied for decades. It has been shown that humans use the same or similar grasp types for certain types of objects, and the shape of the object has a large influence on the applied grasp [KMD+87, GHD12]. Recently, Feix et al. [FBD14] investigated the relationship between grasp types and object attributes in a large real-world human grasping dateset. However, behavioral studies in previous work do not scale to massive dataset. In this work, a Bayesian network is used to model the relations between grasp types and object attributes to boost the recognition of both.

4.2.3 Manipulation action

Past researches on recognizing actions of hand manipulation focused on using first-person vision since it provides an ideal viewing perspective for recording and analyzing hand-object interactions. In [FFR11, FLR12], Fathi et al. used appearance around the manipulation region to recognize egocentric actions. The work in [PR12] has shown that recognizing handled objects helps to infer daily hand activities. In [IKM+15], hand appearance is combined with dense trajectories to recognize hand-object interactions. However, most of previous work are learning actions directly from image appearance, thus the action models learned are easily overfit to image appearance. There are small number of works which aim to reason beyond appearance models [YFA13, JLSZ14, YLFA15a]. In [JLSZ14] a hierarchical model is built to identify persuasive intent of images based on syntactical attributes, such as “smiling” and “waving hand”. The work of [YLFA15a] is most related to our work which seeks to infer action intent from hand grasp types. However, the action model in [YLFA15a] is relatively simple with only three categories to be learned. This work aims to model manipulation actions by jointly considering grasp types together with object attributes.

4.3 Approach

In this work, I propose an unified model to recognize grasp types, object attributes and actions from a single image. The approach is mainly composed by three components: 1) A visual recognition layer which recognizes hand grasp types and attributes of the manipulated objects. 2) A Bayesian network which models the mutual context of grasp types and object attributes to boost the recognition of both. 3) An action modeling layer which learns actions based on the belief distribution of grasp types and object attributes (output of the visual recognition layer).

4.3.1 Visual recognition of grasp types and object attributes

The visual recognition layer consists of two recognition modules, one for grasp types and the other for object attributes. Grasp types and object attributes are important for understanding hand manipulation. Grasp types determine the patterns of how a hand grasps an object, while object attributes indicate the possible functionality of the manipulation. Furthermore, grasp types together with object attributes provide consistent characterization of the manipulation actions.

Grasp types

Hand grasp is important for understanding hand-object manipulations since they characterize how hands hold the objects during manipulation. A number of work have investigated the categorization of grasps into a discrete set of types [Cut89][FPS+09] to facilitate the study of hand grasps. Classifiers are trained for recognizing nine different grasp types selected from a widely used grasp taxonomy proposed by Feix et al. [FPS+09]. The grasp types as shown in Figure 4.2 are selected to cover different standard classification criterion based on functionality [Nap56], object shape, and finger articulation. Some grasp types in original taxonomy which are too similar in appearance are also abstracted into single grasp type (e.g. Thumb-n Finger). Furthermore, all the nine grasp types have a high frequency of daily usage based on the work of [BZR+13]. Thus the grasp types can be applied to larger manipulation tasks.

Hand patches are needed to train grasp classifiers. Following [LK13], a multi-model hand detector composed by a collection of skin pixel classifiers is trained which can adapt to different imaging conditions often faced by a wearable camera. For each test image, a pixel-level hand probability map is generated from the hand detector, and hand patches are then segmented with a bounding box. In detail, candidate hand regions are first selected by binarizing the probability map with a threshold. Regions under a certain area proportion are discarded and at most two regions are retained. Ellipse parameters (length of long/short axis, angle) are fitted to the hand region and the arm part is approximately removed by shortening the length of long axis to 1.5 times of the length of short axis. Then the remaining region is cropped with a bounding box. Linear SVM classifiers are trained for each grasp type using feature vectors extracted from hand patches. As the

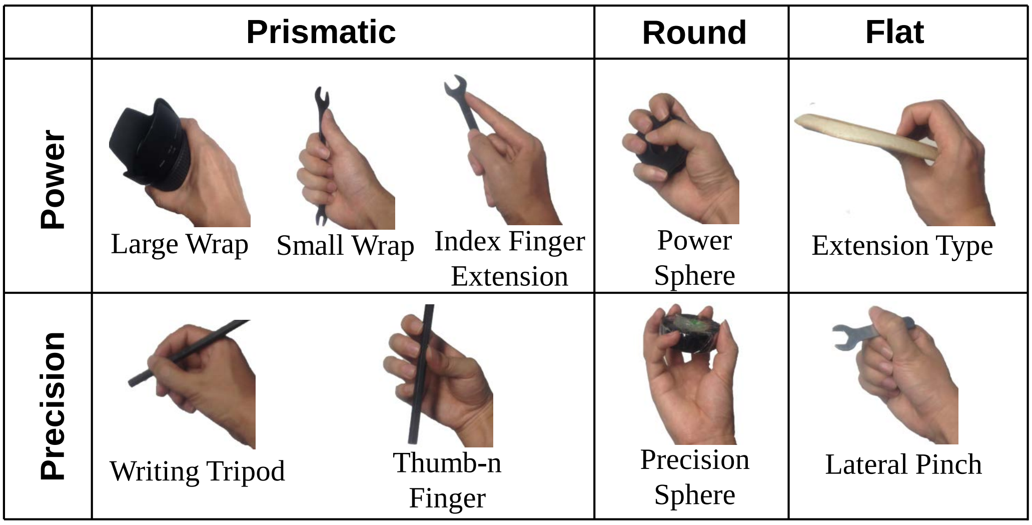


Figure 4.2: The list of nine grasp types selected from [FPS+09], grouped by functionality (*Power* and *Precision*) and object shape (*Prismatic*, *Round* and *Flat*).

recognition output, belief distribution of grasp types (denoted as $P(G|f_G)$) as well as the predicted grasp type with highest probabilistic score are obtained.

Recognition of grasp types provide information about how the hands are holding the objects during manipulation. However, The grasp type alone is not enough to identify fine-grained actions without information from the manipulated objects. In the next section, the method for recognizing object attributes will be presented.

Object attributes

Object attributes are important for understanding hand manipulation since they indicate possible functionality in manipulation. For example, a thick and long object is probably used as a container while a thin and long object is

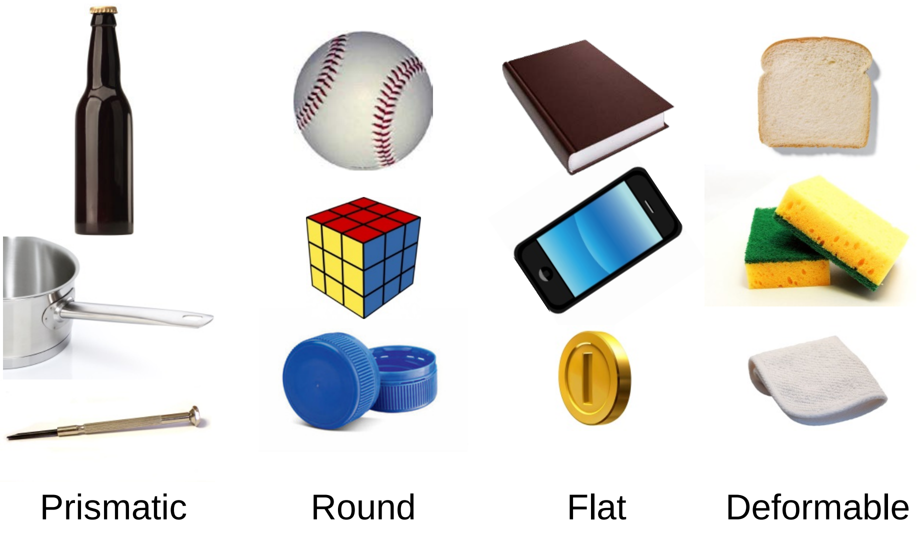


Figure 4.3: Object examples with four different attributes.

probably used as a tool for drawing or stirring. While objects can be assessed by a wide range of attributes, only attributes that are relevant to grasping are focused based on the study of [FBD14]. Here four binary attributes are considered which are important for grasping and can be possibly learned using computer vision techniques. Figure 4.3 illustrates the attributes, three of which are related to object shape and the fourth is related to object rigidity. Three different shape classes are identified based on the criterion in Table 4.1. The fourth attribute of *Deformable* identifies the object that deforms under normal grasping forces. Examples are a sponge or a rag.

Similar to grasp type recognition, object patches are needed to train classifiers for object attributes. However, object detection is a challenging task in computer vision, particularly unreliable when there are occlusions during manipulation. It is observed that hand appearance provides important hint about the relative location and size of the grasped object. As illustrated in Figure 4.4, relative location (d_x, d_y) from the center of hand to the center of

Table 4.1: Classification criterion of three shape classes. Length of object along three object dimensions (major axes of the object) are denoted as A, B, and C, where $A \geq B \geq C$.

Shape classes	Object dimensions
Prismatic	$A > 2B$
Round	$B \leq A < 2B,\ C \leq A < 2C$
Flat	$B > 2C$

object is consistent to the hand orientation, and the object scale (W_o, H_o) is related to the size of hand opening. Therefore, a target regressor is trained for predicting the relative location and scale of the grasped object based on hand appearance. Specifically, regression is performed for three quantities: normalized relative location of (N_x, N_y) and relative scale of N_s specified as follows:

$$\begin{cases} N_x = \dfrac{d_x}{W_h} \\ N_y = \dfrac{d_y}{H_h} \\ N_s = \sqrt{\dfrac{W_o \times H_o}{W_h \times H_h}} \end{cases} \tag{4.1}$$

Here are the steps of how to recognize object attributes: First, SVM regressors are pre-trained based on feature vectors extracted from hand patches. Object bounding boxes are annotated in order to calculate training labels. Then, object patches are segmented with bounding boxes calculated based on the regressed quantities defined in Equation 4.1. Finally, Linear SVM classifiers are trained for each object attribute based on the feature vectors extracted from object patches. As recognition output, belief distribution of different attributes (denoted as $P(O|f_O)$) as well as the predicted attributes

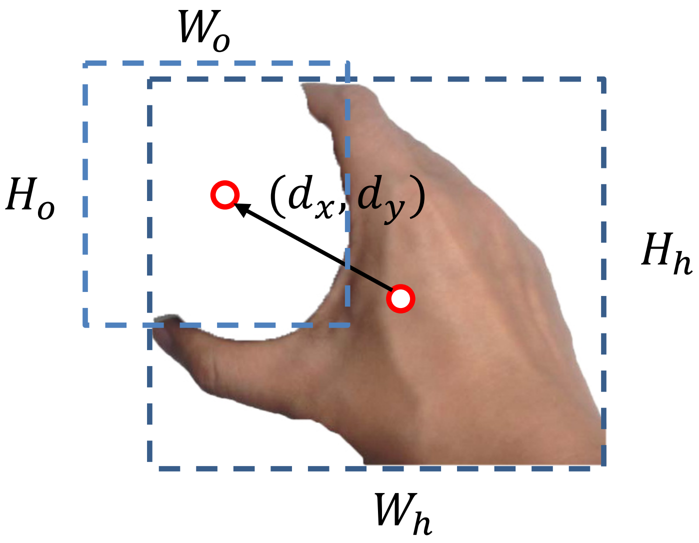


Figure 4.4: Illustration of relative location and scale of the hand and the manipulated object.

are obtained.

Visual recognition of grasp types and object attributes are challenging tasks as there are many occlusions during manipulation. In the next section, the method of how to boost the recognition performance by mutual context will be presented.

4.3.2 Mutual context of grasp types and object attributes

There is strong causal relations between object attributes and grasp types. Object attributes such as geometric shape and rigidity have a large impact on the selection of grasp types. On the other hand, knowing the grasp types used helps to infer the attributes of the grasped object. Thus mutual context between grasp types and object attributes can be explored that knowing the information of one side facilitates the recognition of the other.

In this work, a Bayesian Network is used to model the relations between grasp types and object attributes as illustrated in Figure 4.5. There is a directional connection from object attributes O to grasp types G, encoding the causal relation between object attributes of O and the grasp types of G. f_O and f_G denote the visual features of the corresponding image patches respectively. Based on this model, the posterior probability of object attributes and grasp types given the image evidence can be computed as:

$$\begin{aligned} P(O,G|f_O,f_G) &= \frac{P(O)P(G|O)P(f_O|O)P(f_G|G)}{P(f_O)P(f_G)} \\ &= \frac{P(G|O)P(f_O,O)P(f_G,G)}{P(f_O)P(f_G)P(G)} \\ &\propto P(G|O)P(G|f_G)P(O|f_O) \end{aligned} \tag{4.2}$$

Thus, optimal object attributes O^* and grasp types G^* by maximizing a posterior (MAP) can be jointly inferred as:

$$\begin{aligned} (O^*,G^*) &= \operatorname*{argmax}_{O,G} P(O,G|f_O,f_G) \\ &= \operatorname*{argmax}_{O,G} P(G|O)P(G|f_G)P(O|f_O) \end{aligned} \tag{4.3}$$

where the conditional probability $P(G|O)$ can be estimated by occurrence frequencies of grasp types given certain object attribute, and $P(G|f_G), P(O|f_O)$ are belief distribution of grasp types and object attributes from visual recognition layer.

4.3.3 Action modeling

My hypothesis is that grasp types together with object attributes provide complementary information for characterizing the manipulation action. Previous studies [Nap56] showed that action functionality is an important factor

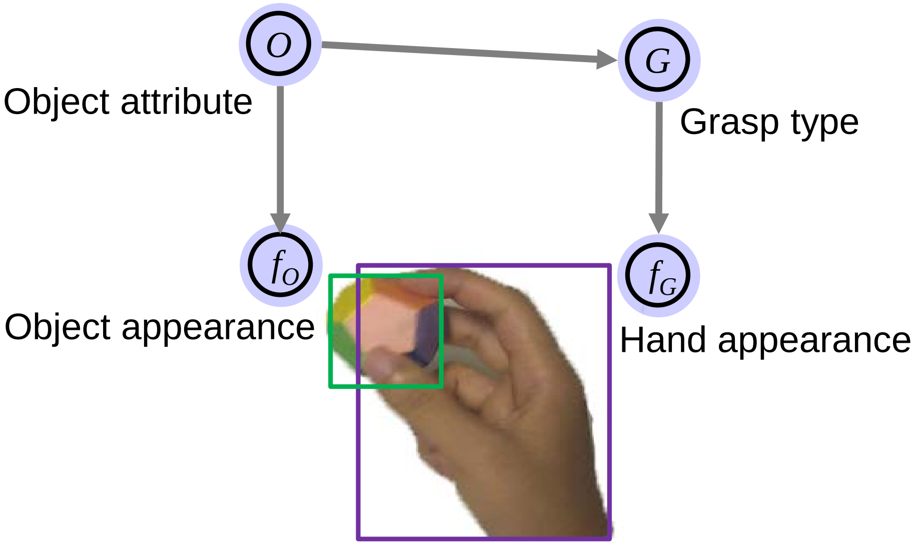


Figure 4.5: A Bayesian network modeling the relationship between object attributes and grasp types.

that affects human grasp selection. Thus it is possible to infer action functionality from grasp types. In this work, a further step is taken to model manipulation actions based on the grasp types of hands as well as the attributes of manipulated objects.

Therefore, I propose a hierarchical semantic action model which builds on visual recognition layer of grasp types and object attributes. The diagram of our approach is shown in Figure 4.6. The hierarchical model separates the action modeling part from the low-level visual recognition part, thus the action learned is independent of image appearance which often changes under different scenes. The visual recognition layer is introduced in Section 4.3.1. At action modeling layer, a linear mapping function is learned for each action based on belief distribution of grasp types and object attributes. More specifically, for each image, the visual recognition layer is applied to extract a 25-dimensional feature vector, of which 17 dimension is composed by belief

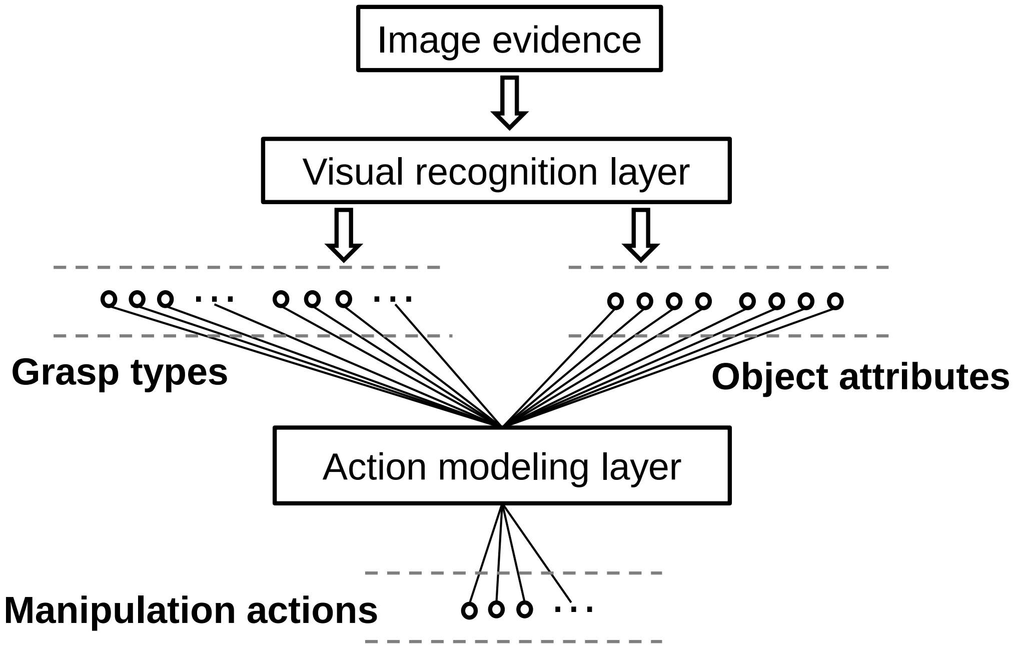


Figure 4.6: Hierarchical semantic action model based on belief distribution of grasp types and object attributes from the visual recognition layer.

distribution of grasp types for two hands (*Writing Tripod* is never used by the left hand) and 8 dimension is composed by belief distribution of object attributes of two grasped objects. Linear SVM classifiers are trained for different actions based on the obtained 25-dimensional feature vectors.

4.4 Evaluation

In this section, four sets of results are presented to validate different components of the proposed approach: (1) grasp type recognition, (2) target regression and object attribute recognition, (3) improved recognition by mutual context of object attributes and grasp types, (4) action recognition.

The approach is evaluated on a public dataset (GTEA Gaze Dataset [FLR12]) of daily activities recorded by wearable cameras. This dataset consists of 17 sequences of cooking activities performed by 14 different subjects. The action verb and object categories with beginning and ending frame are annotated. Additionally, another public dataset (GTEA Gaze+ Dataset [FLR12]) is also used to test the generality of action models. This dataset consists of seven cooking activities, each performed by 10 subjects. Similarly, action labels are provided. The main difference between these two datasets is that in the former dataset activities are performed near a table while in the second dataset activities are performed in a natural setting. The details of evaluation for each component are introduced in following sections.

4.4.1 Grasp type recognition

To train grasp classifiers, grasp types are annotated for 1000 images selected from GTEA Gaze Dataset. Histogram of Oriented Gradient (HoG) is used as baseline feature for grasp type recognition. HoG is compared with other two features based on Convolutional Neural Network (CNN). The two features are extracted from two different layers (*CNN-pool5* and *CNN-fc6*) of the pre-trained CNN model proposed by Krizhevsky et al. [KSH12] using the open source Caffe library [JSD$^+$14]. Compared to *CNN-pool5* which contains five convolutional layers, *CNN-fc6* adds one fully connected layer. Based on these features, linear SVMs are trained for nine grasp types. 5-fold cross-validation is used for evaluation. Note that previous work on visual recognition of grasp types are very few. Only HoG [CKS15] and self-trained CNN [YLFA15a] were used as appearance-based features for grasp type recognition from monocular images. Since there is no sufficient training labels to train a large CNN model, the method in [YLFA15a] is not applied in this work.

Table 4.2: Classification accuracy for nine grasp types on GTEA Gaze Dataset.

	HoG	CNN-pool5	CNN-fc6
Accuracy	50%	61.2%	56.9%

Grasp recognition performance of different features is shown in Table 4.2. Highest classification accuracy of of 61.2% is achieved by *CNN-pool5*. It can be seen that CNN-based feature has advantage over hand-crafted feature HoG, also validated by the work of [YLFA15a]. However, my work shows the feasibility of applying pre-trained CNN model to grasp recognition with scarce training data.

4.4.2 Object attribute recognition

To train target regressors for predicting object location and scale, object bounding boxes are annotated for 1000 images with well detected hand patches from GTEA Gaze Dataset. The bounding box is annotated to include the object part being grasped. To train attribute classifiers, attributes of the grasped objects are also annotated for the same 1000 images. SVM regressors are trained based on features extracted from hand patches. SVM classifiers are trained based on features extracted from within annotated object bounding boxes. The public libSVM library [CL11] is used for implementation. Same features as in Section 4.4.1 are evaluated in 5-fold cross validation. Note that this is the first work on recognizing object attributes for understanding hand-object manipulations and the focus is not on feature design.

Table 4.3 shows quantitative results of target regression. Regressors

Table 4.3: Quantitative results of target regression evaluated by Intersection of Union (IoU) which measures the overlap ratio of ground-truth object bounding box and the predicted object bounding box. The predicted object bounding box with equal width and height are determined based on the regressed quantities defined in Equation 4.1.

	HoG	CNN-pool5	CNN-fc6
IoU	0.471	0.739	0.736

trained by *CNN-pool5* and *CNN-fc6* have similar performance but work much better than HoG. Figure 4.7 shows some qualitative results of the predicted regions of object targets. It can be seen that the predicted regions match well with ground-truth bounding boxes of the manipulated object parts, although the background is cluttered and objects are partially occluded by hands. More importantly, the results indicate that it is possible to detected the manipulated object parts without any specific object detectors.

Table 4.4 shows the classification results for four binary object attributes. Accuracy of over 80% is achieved for all binary attributes and the advantage of CNN-based features over hand-crafted features is verified. For combined attributes, CNN-pool5 achieves best accuracy of 72.4% which means the percentage of cases that all binary features are correctly classified is over 72.4%. The results demonstrate the potential of learning physical properties of the object with monocular images.

4.4.3 Better recognition by mutual context

In this section, evaluation shows the recognition of grasp types and object attributes can be improved by mutual context. The probability of grasp

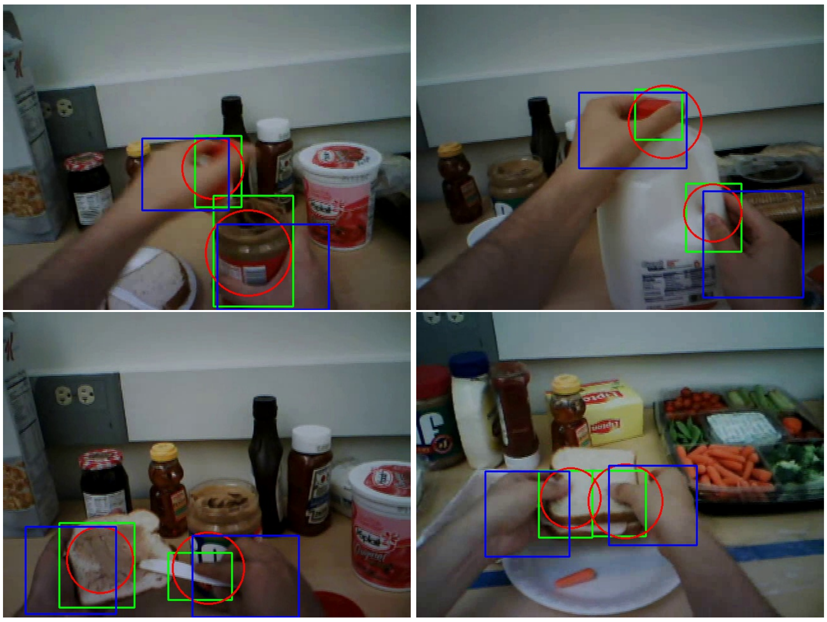


Figure 4.7: Qualitative results of target regression. Blue and green bounding boxes show the detected hand regions and ground-truth object regions respectively. Red circles show the predicted object regions with center of circle indicating object location and radius indicating object scale.

types conditioned on object attributes is estimated as prior information by occurrence frequencies from training data. Figure 4.8 shows the estimated conditional probability. It can be seen that different kinds of objects have very different distribution over grasp types. *Rigid-Prismatic* objects such as a bottle are often held with *Large Wrap* or *Index Finger Extension*, while *Rigid-Round* objects such as a bottle cap are often held with *Precision Sphere*.

The recognition performance of with and without context information are compared. For both two cases, features of *CNN-pool5* are used. The results in Table 4.5 and Table 4.6 show that visual recognition of grasp types and object attributes are significantly improved by using context information. For grasp types, overall classification accuracy is improved by 12.9%. Perfor-

Table 4.4: Performance of attribute classification on GTEA Gaze Dataset. Accuracy is evaluated for four binary attributes separately as well as combined. When evaluating combined attributes, a prediction is considered as accurate if all the attributes are correctly classified.

Object Attribute	HoG	CNN-pool5	CNN-fc6
Prismatic	80.2%	87.9%	84.5%
Round	94.0%	94.0%	95.7%
Flat	81.0%	85.3%	87.1%
Deformable	88.8%	92.2%	91.4%
Combined	60.3%	72.4%	71.9%

mance of most grasp types are improved by object context, except for *Power Sphere* and *Precision Sphere*. I believe the performance deterioration of the two grasp types is due to some false classification of the attribute *Sphere*. For object attributes, classification accuracy for combined attributes is improved by 9.5%. Experiment results strongly support the use of contextual information for improving visual recognition performance.

4.4.4 Action recognition

In this section, experiments are conducted to evaluate the effectiveness of modeling manipulation actions based on semantic information of grasp types and object attributes. The verb part of original action labels in GTEA Gaze Dataset are used as action labels in this work. For example, “Open a jam bottle”” and “Open a peanut bottle” are considered as the same action “Open”. I focus on actions which require two-hand coordination. Seven

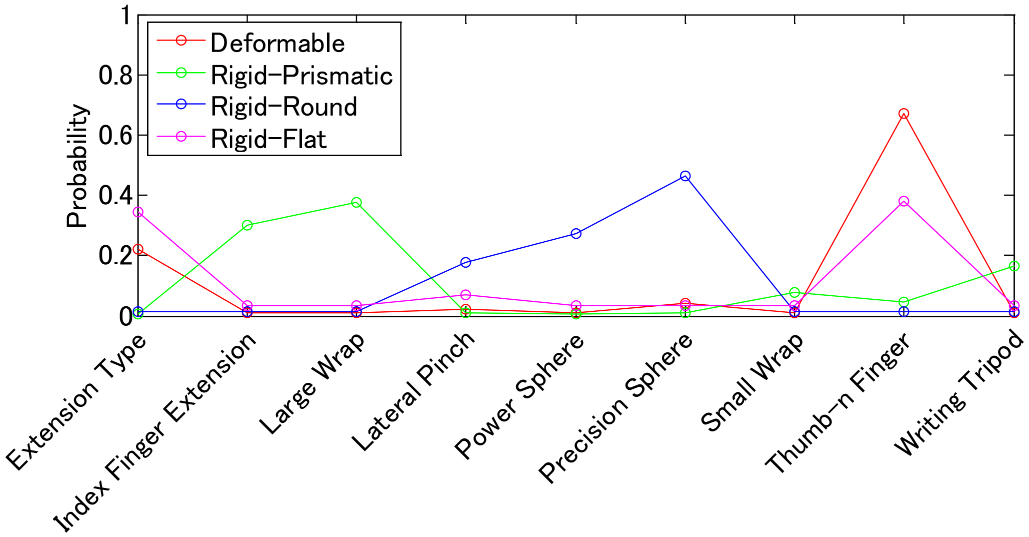


Figure 4.8: Probability of grasp types given object attributes estimated by occurrence frequencies from training data.

action categories are learned in this experiment.

To compare the performance of different components in the proposed action model, linear SVM classifiers are trained based on features from grasp types (GpT), object attributes (OA) and both components (GpT+OA) separately. Note that grasp types were also used in [YLFA15a] for predicting action intention, thus the feature of GpT also serves to evaluate how [YLFA15a] works in modeling manipulation actions. Action recognition performance is also compared with existing methods. Note that no temporal information is used since I focus on recognition from a single image. I choose to compare the method in [FLR12] which utilizes appearance information around gaze location. Since no gaze device is used in this work, an approximate feature representation is composed by concatenating CNN-based features extracted from two hand patches and two object patches (CNN-4). Each CNN-based

Table 4.5: Performance improvement for grasp type recognition by mutual context. F1 measure is evaluated for each grasp type. Accuracy is evaluated for overall performance.

Grasp Category	CNN	CNN+Context
Extension Type	0.166	0.2
Index Finger Extension	0.666	0.949
Large Wrap	0.711	0.818
Lateral Pinch	0.875	0.903
Power Sphere	0.571	0.333
Precision Sphere	0.749	0.666
Small Wrap	0.526	1.0
Thumb-n Finger	0.55	0.59
Writing Tripod	0.733	0.8
Overall	61.2%	74.1%

feature vector is reduced to a 100-dimensional feature vector using Principal Component Analysis (PCA) and the feature dimension for CNN-4 is 400. Performance is evaluated using 5-fold cross validation based on labeled images from GTEA Gaze Dataset.

The classification accuracy for seven actions is shown in Table 4.7. The proposed GT+OA achieves best classification accuracy of 79.3%, which indicate the combination of grasp types and object attributes works better than using grasp types alone. GT+OA also outperforms CNN-4, which verifies the advantage of our action model over appearance-based method. Note that my method only relies on 25 dimensional feature vector from a single image. The confusion matrix for seven manipulation actions is shown in Figure 4.9. The

Table 4.6: Performance improvement for object attribute recognition by mutual context (evaluated by accuracy).

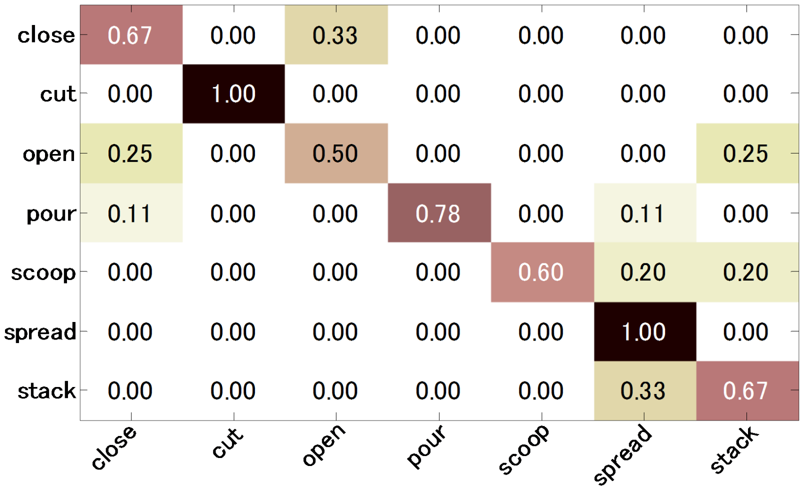
Object Attributes	CNN	CNN+Context
Prismatic	87.9%	88.8%
Round	94.0%	95.7%
Flat	85.3%	88.8%
Deformable	92.2%	92.2%
Combined	72.4%	81.9%

proposed method mainly confuses *Close* with *Open*. I believe that this is because for some objects (such as a bottle) these two actions share similar grasp types and object attributes from a single image.

To demonstrate the correlation between each action and its semantic components of grasp types and object attributes, model parameters from support vectors learned by each linear SVM classifier are computed. Model parameters indicate the correlation between action and its 25 semantic components. Visualization of model parameters is illustrated in Figure 4.10. It can be seen that each action has strong correlation to different grasp types and object attributes.

Table 4.7: Performance comparison for recognizing seven action classes on GTEA Gaze Dataset. CNN-4 is used as a baseline feature approximating the work of [FLR12]. GpT is used as a baseline feature based on grasp types similar to the work of [YLFA15a]. The features of OA and GpT+OA is proposed in this work considering the joint use of object attributes together with grasp types.

	Accuracy
CNN-4 [FLR12]	70.3%
GpT [YLFA15a]	69.0%
OA	70.7%
GpT+OA	79.3%



	close	cut	open	pour	scoop	spread	stack
close	0.67	0.00	0.33	0.00	0.00	0.00	0.00
cut	0.00	1.00	0.00	0.00	0.00	0.00	0.00
open	0.25	0.00	0.50	0.00	0.00	0.00	0.25
pour	0.11	0.00	0.00	0.78	0.00	0.11	0.00
scoop	0.00	0.00	0.00	0.00	0.60	0.20	0.20
spread	0.00	0.00	0.00	0.00	0.00	1.00	0.00
stack	0.00	0.00	0.00	0.00	0.00	0.33	0.67

Figure 4.9: Confusion matrix for manipulation action classification using grasp types and object attributes on GTEA Gaze Dataset.

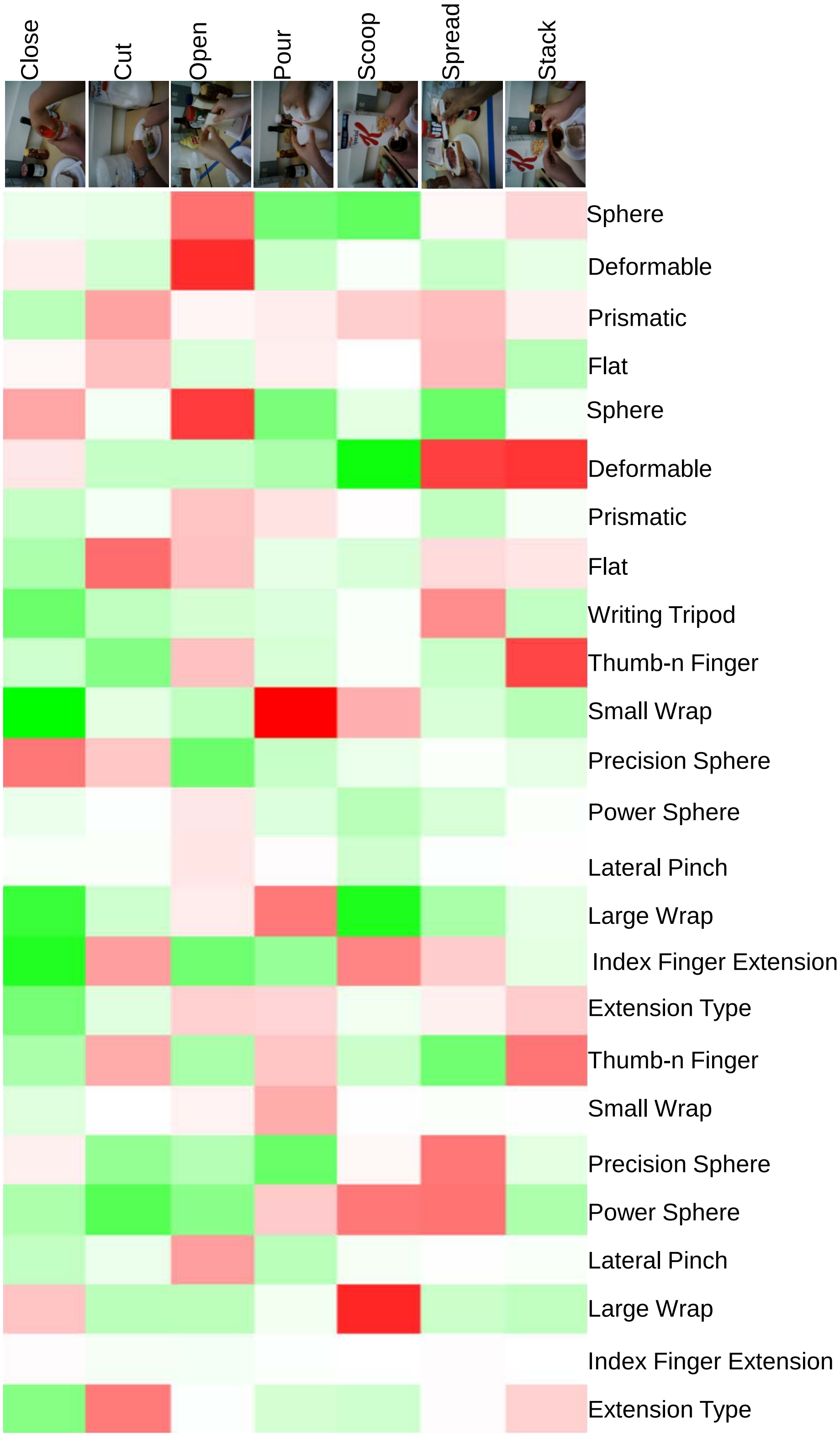


Figure 4.10: Visualization of model parameters for seven action classes. The saturation of red color indicates positive correlation while the saturation of green color indicates negative correlation. White color indicates no correlation.

Table 4.8: Generality evaluation of action models by training on GTEA Gaze Dataset and testing on GTEA Gaze+ Dataset. Appearance-based model is trained based on CNN-4, while the proposed hierarchical model is trained based on GpT+OA.

	Appearance-based	Proposed
Accuracy	29.2%	50.4%

To compare the generality of the proposed semantic action model with appearance-based model, action recognition is performed by training and testing on different datasets. While all the training procedure is done on GTEA Gaze Dataset, actions are predicted on GTEA Gaze+ Dataset recorded in different environments. 100 images are selected for each action category and a total of 700 images from GTEA Gaze+ Dataset are used for testing. Classification accuracy is shown in Table 4.8. The proposed semantic model outperforms the appearance-based model by over 20%, which indicates that the proposed method is more robust to overfitting.

4.5 Conclusion

In this chapter, I propose an unified model for understanding hand-object manipulation with a wearable camera. From a single image, grasp types are recognized from detected hand patches and object attribute information are extracted from the manipulated objects. Furthermore, mutual context is explored to boost the recognition of both grasp types and object attributes. Finally, actions are recognized based on belief distribution of grasp types and object attributes.

Experiments are conducted to evaluate the proposed approach: (1) Average accuracy of 61.2% is achieved for grasp type recognition and if 72.4% is achieved for object attribute classification. (2) By mutual context, recognition performance is improved by 12.9% for grasp types and by 9.5% for object attributes. (3) Best average accuracy of 79.3% for manipulation action recognition is achieved using the proposed semantic action model. Evaluation results for model generality support my hypothesis that grasp types and object attributes contain consistent information for characterizing different actions.

Chapter 5

Conclusions

5.1 Summary

In this thesis, methods are presented for recognizing and analyzing hand grasp types, and modeling manipulation actions from first-person view video with a wearable monocular camera. Chapter 1 explains the motivation of this work, describing the importance of the topic and the shortcomings of previous approaches. Against these shortcomings, new methods are proposed and introduced in the following chapters. In Chapter 2, a first-person vision system is proposed to recognize hand grasp types and discover visual structures of hand grasp in everyday manipulation tasks. In the system, a wearable camera is used to record hand manipulation tasks. Advances of computer vision techniques are incorporated in the system to do hand detection, and extract appearance-based features for training discriminative grasp classifiers. An iterative clustering method is proposed to learn visual structures between different grasp types. Chapter 3 introduces a new feature presentation based on hand-guided feature tracking which improves the grasp recognition performance and is more robust to hand detection noises than appearance-based features. In Chapter 4, semantic action model is proposed which encodes high-level semantic constrains of actions based on hand grasp types and object attributes. Furthermore, novel methods for extracting attributes of the manipulated object are proposed without any specific object detectors, and the mutual context between grasp types and object attributes is explored to boost the recognition performance. As a whole, the methods presented in this thesis offer a scalable way for studying the use of human hands in daily manipulation tasks at a large scale.

5.2 Contributions

The main contributions of this work are summarized as follow:

- Propose a first-person vision system for hand grasp analysis. The system is capable of recognizing hand grasp types and analyzing grasp structures for everyday manipulation tasks with a single wearable monocular camera. The work shows the potential for using computer vision techniques for analyzing hand grasps with large scale of data in real-life settings.

- Propose a method for recognizing hand grasp types, object attributes and manipulation actions from a single image within a unified model. Attribute information from the manipulated object can be extracted without using specific object detectors. Mutual context of grasp types and object attributes is explored to enhance the recognition of both. Furthermore, the proposed hierarchical semantic action model outperforms appearance-based models and is robust to overfitting.

5.3 Future work

5.3.1 Grasp recognition with wearable RGB-D cameras

In Chapter 3, a feature representation based on trajectory information of hand tracking is proposed to improve grasp recognition performance. However, the classification accuracy is still not good enough for practical use in real world applications.

In recent years, RGB-D cameras capable of recording both appearance and depth information are becoming popular in various estimation techniques such as 3D reconstruction and body pose estimation. With the advancement of hardware and sensing techniques, RGB-D cameras are becoming smaller and smaller from Microsoft Kinect [KIN] to Creative Senz3D [SEN]. Rogez et al. [RIR15] recently proposed method for discrete hand pose recognition with a chest-mounted RGB-D camera although it is not originally designed for wearable usage. I believe that wearable RGB-D cameras will be available in the near future, making it possible for researchers to extract stable 3D features in first-person vision applications. Using 3D features, the spacial configuration between fingers and the geometric information of the grasped object can be explored, and performance of the grasp recognition system will be largely improved.

5.3.2 Temporal dynamics of grasp types in hand manipulation

In Chapter 4, a bottom-up hierarchical model is proposed for recognizing manipulation actions based on hand grasp types and object attributes from a single image. However, the dynamics of hand manipulation is complex, for the patterns of how the hands holds the objects is changing overtime. Thus, the information from a single image is insufficient for identifying fine-grained actions which share similar grasping behavior, such as opening or closing a bottle cap.

Regarding recognition of fine-grained manipulation actions, it is also important to investigate the temporal dynamics of hand grasp types used in certain manipulation actions. In performing a manipulation action, human selection of grasp types for the target object changes according to the variation of task requirements, such as force and dexterity. The temporal dynamics of grasp types can be used as discriminative characterization for different actions. Hence, to completely understand hand manipulation, it is important to consider the temporal dynamics, not only image evidence.

5.3.3 Grasp analysis-based diagnosis system

With practical grasp analysis techniques in first-person vision, many applications can be proposed. One important application is a grasp analysis-based diagnosis system which can provide useful feedback for both clinical diagnosis and task assistance.

In this thesis, computer vision-based techniques are utilized to recognize different hand grasp types of a single user in manipulation tasks. However the detected grasp types can further be utilized to analyze hand grasping behavior of different users in certain manipulation tasks. Profile information can be built for each person by monitoring habitual knowledge, such as the manner in which a manipulation task is performed, and the duration and frequency of grasp types. In clinical diagnosis, deviation from person-tailored profile and typical behaviours can be detected as feedback, helping clinicians in assessing individuals' health status and diagnosing disease-related problems. In task assistance, skill assessment of beginners can be achieved by comparing relevant traits of grasping behaviors with skilled workers.

Bibliography

[BFD13] Ian M Bullock, Thomas Feix, and Aaron M Dollar. Finding small, versatile sets of human grasps to span common objects. In *Robotics and Automation (ICRA), 2013 IEEE International Conference on*, pages 1068–1075. IEEE, 2013.

[BFD14] IM Bullock, T Feix, and AM Dollar. The yale human grasping data set: Grasp, object and task data in household and machine shop environments. 2014.

[BOID05] Keni Bernardin, Koichi Ogawara, Katsushi Ikeuchi, and Ruediger Dillmann. A sensor fusion approach for recognizing continuous human grasping sequences using hidden markov models. *Robotics, IEEE Transactions on*, 21(1):47–57, 2005.

[BPS$^+$14] Lorenzo Baraldi, Federica Paci, Giovanni Serra, Luca Benini, and Rita Cucchiara. Gesture recognition in ego-centric videos using dense trajectories and hand segmentation. In *Computer Vision and Pattern Recognition Workshops (CVPRW), 2014 IEEE Conference on*, pages 702–707. IEEE, 2014.

[BZR$^+$13] Ian M Bullock, Joshua Z Zheng, SDL Rosa, Charlotte Guertler, and Aaron M Dollar. Grasp frequency and usage in daily household and machine shop tasks. *Haptics, IEEE Transactions on*, 6(3):296–308, 2013.

[CKS15] Minjie Cai, Kris M Kitani, and Yoichi Sato. A scalable approach for understanding the visual structures of hand grasps. In *Robotics and Automation (ICRA), 2015 IEEE International Conference on*, pages 1360–1366. IEEE, 2015.

[CL11] Chih-Chung Chang and Chih-Jen Lin. Libsvm: a library for support vector machines. *ACM Transactions on Intelligent Systems and Technology (TIST)*, 2(3):27, 2011.

[CSPA$^+$92] Jane Case-Smith, Charlane Pehoski, American Occupational Therapy Association, et al. *Development of hand skills in children.* American Occupational Therapy Association, 1992.

[Cut89] Mark R Cutkosky. On grasp choice, grasp models, and the design of hands for manufacturing tasks. *Robotics and Automation, IEEE Transactions on*, 5(3):269–279, 1989.

[CYB] http://www.cyberglovesystems.com/.

[DB14] Raphael Deimel and Oliver Brock. A novel type of compliant, underactuated robotic hand for dexterous grasping. *Robotics: Science and Systems, Berkeley, CA*, pages 1687–1692, 2014.

[DPZ91] Hristo N Djidjev, Grammati E Pantziou, and Christos D Zaroliagis. Computing shortest paths and distances in planar graphs. In *Automata, Languages and Programming*, pages 327–338. Springer, 1991.

[DT05] Navneet Dalal and Bill Triggs. Histograms of oriented gradients for human detection. In *Computer Vision and Pattern Recognition, 2005. CVPR 2005. IEEE Computer Society Conference on*, volume 1, pages 886–893. IEEE, 2005.

[EK05] Staffan Ekvall and Danica Kragic. Grasp recognition for programming by demonstration. In *Robotics and Automation, 2005. ICRA 2005. Proceedings of the 2005 IEEE International Conference on*, pages 748–753. IEEE, 2005.

[Far03] Gunnar Farnebäck. Two-frame motion estimation based on polynomial expansion. In *Image Analysis*, pages 363–370. Springer, 2003.

[FBD14] Thomas Feix, I Bullock, and A Dollar. Analysis of human grasping behavior: Object characteristics and grasp type. *Haptics, IEEE Transactions on*, 7(3):311–323, 2014.

[FEHF09] Alireza Farhadi, Ian Endres, Derek Hoiem, and David Forsyth. Describing objects by their attributes. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1778–1785. IEEE, 2009.

[FFR11] Alireza Fathi, Ali Farhadi, and James M Rehg. Understanding egocentric activities. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 407–414. IEEE, 2011.

[FGE$^+$99] Holger Friedrich, Volker Grossmann, Markus Ehrenmann, Oliver Rogalla, R Zöllner, and Rudiger Dillmann. Towards cognitive elementary operators: grasp classification using neural network classifiers. In *Proceedings of the IASTED Interna-*

tional Conference on Intelligent Systems and Control (ISC), volume 1, pages 88–93, 1999.

[FLR12] Alireza Fathi, Yin Li, and James M Rehg. Learning to recognize daily actions using gaze. In *Computer Vision–ECCV 2012*, pages 314–327. Springer, 2012.

[FPS$^+$09] Thomas Feix, Roland Pawlik, Heinz-Bodo Schmiedmayer, Javier Romero, and Danica Kragic. A comprehensive grasp taxonomy. In *Robotics, Science and Systems: Workshop on Understanding the Human Hand for Advancing Robotic Manipulation*, pages 2–3, 2009.

[GHD12] René Gilster, Constanze Hesse, and Heiner Deubel. Contact points during multidigit grasping of geometric objects. *Experimental brain research*, 217(1):137–151, 2012.

[GOP] https://gopro.com/update/hero3.

[HMMK15] De-An Huang, Minghuang Ma, Wei-Chiu Ma, and Kris M Kitani. How do we use our hands? discovering a diverse set of common grasps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 666–675, 2015.

[HSKMVG09] Henning Hamer, Konrad Schindler, Esther Koller-Meier, and Luc Van Gool. Tracking a hand manipulating an object. In *Computer Vision, 2009 IEEE 12th International Conference On*, pages 1475–1482. IEEE, 2009.

[HX-] http://www.panasonic.com/uk/consumer/cameras-camcorders/camcorders/active-hd-camcorders/hx-a1me.html.

[IBA86] Thea Iberall, Geoffrey Bingham, and MA Arbib. Opposition space as a structuring concept for the analysis of skilled hand movements. *Experimental brain research series*, 15:158–173, 1986.

[IKM+15] Tatsuya Ishihara, Kris M Kitani, Wei-Chiu Ma, Hironobu Takagi, and Chieko Asakawa. Recognizing hand-object interactions in wearable camera videos. In *IEEE International Conference on Image Processing (ICIP)*, 2015.

[ISO] http://www.isotechnology.net/content/isotouch.

[JLSZ14] Jungseock Joo, Weixin Li, Francis F Steen, and Song-Chun Zhu. Visual persuasion: Inferring communicative intents of images. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 216–223. IEEE, 2014.

[JSD+14] Yangqing Jia, Evan Shelhamer, Jeff Donahue, Sergey Karayev, Jonathan Long, Ross Girshick, Sergio Guadarrama, and Trevor Darrell. Caffe: Convolutional architecture for fast feature embedding. In *Proceedings of the ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[KBBN09] Neeraj Kumar, Alexander C Berg, Peter N Belhumeur, and Shree K Nayar. Attribute and simile classifiers for face verification. In *Computer Vision, 2009 IEEE 12th International Conference on*, pages 365–372. IEEE, 2009.

[Kel47] Adrian D Keller. *Studies to determine the functional requirements for hand and arm prosthesis*. Department of Engineering University of California, 1947.

[KI93] Sing Bing Kang and Katsushi Ikeuchi. Toward automatic robot instruction from perception-recognizing a grasp from observation. *Robotics and Automation, IEEE Transactions on*, 9(4):432–443, 1993.

[KIN] https://en.wikipedia.org/wiki/Kinect.

[KMD$^+$87] Roberta L Klatzky, Brian McCloskey, Sally Doherty, James Pellegrino, and Terence Smith. Knowledge about hand shaping and knowledge about objects. *Journal of motor behavior*, 19(2):187–213, 1987.

[KRK08] H Kjellstrom, Javier Romero, and Danica Kragic. Visual recognition of grasps for human-to-robot mapping. In *Intelligent Robots and Systems, 2008. IROS 2008. IEEE/RSJ International Conference on*, pages 3192–3199. IEEE, 2008.

[KSH12] 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.

[LFR13] Yin Li, Alahoum Fathi, and James M Rehg. Learning to predict gaze in egocentric video. In *Computer Vision (ICCV),* *2013 IEEE International Conference on*, pages 3216–3223. IEEE, 2013.

[LK13] Cheng Li and Kris M Kitani. Pixel-level hand detection in egocentric videos. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 3570–3577. IEEE, 2013.

[LNH09] Christoph H Lampert, Hannes Nickisch, and Stefan Harmeling. Learning to detect unseen object classes by between-class attribute transfer. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 951–958. IEEE, 2009.

[Low04] David G Lowe. Distinctive image features from scale-invariant keypoints. *International journal of computer vision*, 60(2):91–110, 2004.

[Nap56] John R Napier. The prehensile movements of the human hand. *Journal of bone and Joint surgery*, 38(4):902–913, 1956.

[OKA11] Iasonas Oikonomidis, Nikolaos Kyriazis, and Antonis A Argyros. Full dof tracking of a hand interacting with an object by modeling occlusions and physical constraints. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2088–2095. IEEE, 2011.

[PG11] Devi Parikh and Kristen Grauman. Relative attributes. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 503–510. IEEE, 2011.

[Pla99] John C Platt. Probabilistic outputs for support vector machines and comparisons to regularized likelihood methods. In *Advances in large margin classifiers*. Citeseer, 1999.

[PR12] Hamed Pirsiavash and Deva Ramanan. Detecting activities of daily living in first-person camera views. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 2847–2854. IEEE, 2012.

[PSM10] Florent Perronnin, Jorge Sánchez, and Thomas Mensink. Improving the fisher kernel for large-scale image classification. In *Computer Vision–ECCV 2010*, pages 143–156. Springer, 2010.

[RFKK10] Javier Romero, Thomas Feix, Hedvig Kjellstrom, and Danica Kragic. Spatio-temporal modeling of grasping actions. In *Intelligent Robots and Systems (IROS), 2010 IEEE/RSJ International Conference on*, pages 2103–2108. IEEE, 2010.

[RIR15] Grégory Rogez, James S Supancic III, and Deva Ramanan. First-person pose recognition using egocentric workspaces. In *Computer Vision and Pattern Recognition, 2015 IEEE Conference on*, 2015.

[RKEK13] Javier Romero, Hedvig Kjellström, Carl Henrik Ek, and Danica Kragic. Non-parametric hand pose estimation with object context. *Image and Vision Computing*, 31(8):555–564, 2013.

[Sch19] G. Schlesinger. Der mechanische aufbau der kunstlichen glieder. *Ersatzglieder und Arbeitshilfen fur Kriegsbeschadigte und Unfallverletzte*, pages 321–661, 1919.

[SEN] http://us.creative.com/p/web-cameras/creative-senz3d#.

[SFD11] Behjat Siddiquie, Rogerio S Feris, and Larry S Davis. Image ranking and retrieval based on multi-attribute queries. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 801–808. IEEE, 2011.

[SFS98] Marco Santello, Martha Flanders, and John F Soechting. Postural hand synergies for tool use. *The Journal of Neuroscience*, 18(23):10105–10115, 1998.

[STK15] Akanksha Saran, Damien Teney, and Kris M Kitani. Hand parsing for fine-grained recognition of human grasps in monocular images. In *Intelligent Robots and Systems (IROS), 2015 IEEE/RSJ International Conference on*. IEEE, 2015.

[VDC12] Eleonora Vig, Michael Dorr, and David Cox. Space-variant descriptor sampling for action recognition based on saliency and eye movements. In *Computer Vision–ECCV 2012*, pages 84–97. Springer, 2012.

[VMT$^+$14] Andrea Vedaldi, Siddarth Mahendran, Stavros Tsogkas, Subhrajyoti Maji, Ross Girshick, Juho Kannala, Esa Rahtu, Iasonas Kokkinos, Matthew B Blaschko, Daniel Weiss, et al. Understanding objects in detail with fine-grained attributes. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 3622–3629. IEEE, 2014.

[WCE$^+$01] Steven L Wolf, Pamela A Catlin, Michael Ellis, Audrey Link Archer, Bryn Morgan, and Aimee Piacentino. Assessing wolf motor function test as outcome measure for research in patients after stroke. *Stroke*, 32(7):1635–1639, 2001.

[WKSL11] Heng Wang, Alexander Kläser, Cordelia Schmid, and Cheng-Lin Liu. Action recognition by dense trajectories. In *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on*, pages 3169–3176. IEEE, 2011.

[WS13] Heng Wang and Cordelia Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558. IEEE, 2013.

[YFA13] Yezhou Yang, Cornelia Fermuller, and Yiannis Aloimonos. Detection of manipulation action consequences (mac). In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 2563–2570. IEEE, 2013.

[YLFA15a] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. Grasp type revisited: A modern perspective of a classical feature for vision. In *Computer Vision and Pattern Recognition, 2015 IEEE Conference on*, 2015.

[YLFA15b] Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. Robot learning manipulation action plans by" watching" unconstrained videos from the world wide web. In *Twenty-Ninth AAAI Conference on Artificial Intelligence*, 2015.

[ZDLRD11] Joshua Z Zheng, Sara De La Rosa, and Aaron M Dollar. An investigation of grasp type and frequency in daily household and machine shop tasks. In *Robotics and Automation (ICRA),* *2011 IEEE International Conference on*, pages 4169–4175. IEEE, 2011.

[ZPR+14] Ning Zhang, Manohar Paluri, Marc'Aurelio Ranzato, Trevor Darrell, and Lubomir Bourdev. Panda: Pose aligned networks for deep attribute modeling. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 1637–1644. IEEE, 2014.

Publications

[1] Minjie Cai, Kris M. Kitani, and Yoichi Sato. “A Scalable Approach for Understanding the Visual Structures of Hand Grasps”. In proc. *IEEE International Conference on Robotics and Automation* (ICRA2015), May 2015.

[2] Minjie Cai, Kris M. Kitani, and Yoichi Sato. “Hand Grasp Recognition from Egocentric Videos”. In proc. *IEEE Computer Society Workshop on Observing and understanding hands in action* (HANDS2015), June 2015.

[3] Minjie Cai, Kris M. Kitani, and Yoichi Sato. “Discovering Appearance-based Grasp Structures with Wearable Cameras”. 電子情報通信学会クラウドネットワークロボット研究会 (IEICE-CNR2014), December 2014.

[4] Minjie Cai, Kris M. Kitani, and Yoichi Sato. “Hand Skeleton Pruning based on Contour Partition with Fingertip Detection”. 画像の認識・理解シンポジウム (MIRU2014), July 2014.