APPROVAL SHEET

Title of dissertation: SHARING LEARNED MODELS BETWEEN HETEROGENEOUS ROBOTS: AN IMAGE DRIVEN INTERPRETATION

Name of Candidate: Isha Rahul Potnis
Masters of Science
Computer Science, 2018

Dissertation and Abstract approved: _____

Dr. Cynthia Matuszek
Assistant Professor
Computer Science and Electrical Engineering

Date approved: _____

PERSONAL DETAILS

Name:	Isha Rahul Potnis
Permanent Address:	I-1, State Bank Nagar co-op hsg society, Bibvewadi, Pune - 411037 India
Degree and Date to be conferred:	Masters of Science, July 2018 Computer Science
Date of Birth:	March 30, 1993
Place of birth:	Pune, India
Secondary Education:	Vidya Niketan English Medium School, Bibvewadi, Pune - 411037 India
Collegiate Institutes attended:	University of Maryland, Baltimore County Master of Science, May 2018 University of Pune Bachelor of Engineering, May 2015
Major:	Computer Science
Professional positions held:	Full stack web developer intern Proticy LLC, Baltimore County, MD, USA Research Assistant UMBC, MD, USA
Personal Email ID:	ishapotnis30@gmail.com

ABSTRACT

Title of dissertation: SHARING LEARNED MODELS BETWEEN HETEROGENEOUS ROBOTS: AN IMAGE DRIVEN INTERPRETATION

Isha Rahul Potnis, Master of Science, 2018

Dissertation directed by: Dr. Cynthia Matuszek
Assistant Professor
Computer Science and Electrical Engineering

With the evolution of robotics to produce more affordable and proficient robots, it has become crucial for robots to get acquainted with their environment and tasks quickly. This requires training classifiers to identify objects denoted by natural language, a type of grounded language acquisition and visual perception. The current approaches require extensive training data gathered from humans for robots to learn the contextual models. For robots to work collaboratively, every robot must understand the task requirement and its corresponding environment. Teaching every robot these tasks separately would multiply human interaction with robots. Research in 'transfer learning' is gaining momentum to avoid the repetitive training task and minimize human-robot interaction.

With the advancement of personal assistance in elderly care and teaching domains, where the learned robot models are environment-specific, transferring the learned model to other robots with minimum loss of accuracy is crucial. Homoge-

neous transferred learning is easy as compared to transfer learning in heterogeneous robot environment with different perceptual sensors.

We propose the 'chained learning approach' to transfer data between robots with different perceptual capabilities. These differences in sensory processing and representations may lead to a gradual drop in transfer learning accuracy. We conduct experiments for co-located robots with similar sensory ability, with qualitatively different camera sensors, and for non-co-located robots to test our learning approach. A comparative study of cutting-edge feature extraction algorithms help us build an efficient pipeline for optimal knowledge transfer.

Our preliminary experiments lay a foundation for efficient transfer learning in a heterogeneous robot environment while introducing domain adaptation as a potential research option for grounded language transfer.

SHARING LEARNED MODELS BETWEEN HETEROGENOUS ROBOTS: AN IMAGE-DRIVEN INTERPRETATION

by

Isha Rahul Potnis

Dissertation submitted to the Faculty of the Graduate School of the
University of Maryland, Baltimore County in partial fulfillment
of the requirements for the degree of
Master of Science
2018

Advisory Committee:
Dr. Cynthia Matuszek, Advisor
Dr. Don Engel
Dr. Francis Ferraro

Preface

This dissertation on "Sharing learned models between heterogeneous robots: an image-driven interpretation" tries to answer two main questions: whether we can transfer learned models between heterogeneous robots with minimum loss of accuracy and can this transfer be independent of shared workspace. This dissertation is written to fulfill the graduation requirements of the M.S. Computer Science program at the University of Maryland, Baltimore County in the period between January and June 2018.

The research was carried out under the purview of the Interactive Robotics and Language (IRAL) Lab at UMBC. I was a Research Assistant associated with IRAL from January, 2018 to May, 2018. This topic was chosen to accommodate my areas of interest, which are Robotics, Image Processing and Machine Learning.

Acknowledgments

I owe my gratitude to all the people who have made this thesis possible and because of whom my graduate experience has been one that I will cherish forever.

First and foremost I'd like to thank my advisor, Dr. Cynthia Matuszek for giving me an invaluable opportunity to work on such a challenging and extremely interesting project.

I would like to thank all my committee members for the excellent guidance throughout the entire research process. The brainstorming IRAL sessions were very helpful in gaining insights of the latest trends in industry and how they can affect our research problem. I would like to thank all my IRAL colleagues for giving inputs during moments of confusion. I would especially like to thank my friend, Nisha Pillai, for her guidance and support during the research. I really loved working with Dr.Cynthia Matuszek, my advisor and guide, and IRAL group during the entire duration of my research.

I would, also, like to thank my friends and family, especially Timothy Lewis, Srishty Saha and Tushar Rathi for their valuable suggestions to my research.

Table of Contents

List of Figures

List of Abbreviations

BTOMP	Batch Tree Orthogonal Matching Pursuit
CCG	Combinatory Categorial Grammar
DoG	Difference of Gaussian
FUBL	Factored Unification Based Learning
HMP	Hierarchical Matching pursuit
LDA	Linear Discriminant Analysis
NLP	Natural Language Processing
OMP	Orthogonal Matching Pursuit
PCA	Principle Component Analysis
SVM	Support Vector Machine
SIFT	Scale Invariant Feature Transform
SVD	Singular Value Decomposition

Chapter 1

Introduction

Advancement in technology in the last few years has led to a growth in the availability of robots. Rather than being large, dedicated pieces of equipment operated only by trained technicians, today's robots are smaller, more affordable, and heterogeneous. This development enables the deployment of useful robots in traditionally human-centric environments such as schools, nursing homes, and assisted living facilities. However, to increment the usability of robots in such environments, the robots must be intuitive, and communicating with them should be natural and user-friendly. One approach that is gaining traction uses natural language for interactions with robots, thereby enabling layman users to teach robots about their particular way of expressing instructions to convey information about their needs and desired environment.

Grounded Language Acquisition is a growing research area in robotics and NLP. It involves learning models which map linguistic constructs to perceivable world. The core problem of language dependent robots is mapping the semantics of words and sentences with perceptual world in a noisy environment [49]. Learning about the user specific requirements enable robots to know their user better. It lets the robot learn how a particular user chooses to refer to its surroundings. Such personalized learning is appropriate for assisted technology settings such as elder

care. In such settings, each person has different needs and way to address them vary. Grounded language acquisition plays an important role in such user-oriented domains.

1.1 Motivation

There is a rapid growth in the need for technological assistance in elder care domain. The proportion of the population that is over 65 is rising, a trend that is expected to continue in the coming years. While it is preferable for seniors to maintain the ability to live independently, most people will eventually need assistance with both physical tasks and with managing cognitive decline. Our research has application in assisted living in elder care domain. Assisted robotics may allow an aging population to live independently for longer, with less stress on families and the social network.

In elderly care environment, where users have varying needs, multiple small robots with varying skills are more useful than a single robot with a fixed set of capabilities. Such robots might also be shared, for example, in assisted living facilities where some task is being done collaboratively by more than one robot. Shared robots will have a greater need for informed user customization. To accomplish this while placing the smallest possible burden on the user, heterogeneous robots in this environment should be able to share learned knowledge about language and actions and should be able to distribute tasks.

1.2 Research Objective:

Our primary objective is to develop an approach that lets the heterogeneous robots share the learned models. A collaborative task completion requires multiple heterogeneous robots to work together with common basic understanding of the environment. To make the training task easier for users in a user-customized environment, the robots must know how to transfer the models to other robots. Loss of accuracy when the knowledge models are being transferred is an important factor in health care domain. Our research calculates accuracy loss while sharing the learned models between the robots.

1.3 Key Insights

Current work on grounded language acquisition depends on certain assumptions that limit collaborative learning and performance. Transfer learning between robots with different perceptions is important as the learned models are specific to a robot's perceptual and physical capabilities. For sharing learned models between robots with different dimensional perceptions of the environment, we need a proper understanding of the perceptual powers of other robot. Thus, the goal of the proposed research is to develop a learning mechanism that lets robots build and share personalized models of a person's language, actions, and goals. This will ultimately let heterogeneous robots in an elder care setting to respond to instructions and anticipate needs that are idiosyncratic to individual users, while collaborating on tasks that require a mix of capabilities.

1.4 Thesis Statement

To develop an approach using transfer learning to support collaboration in customized human-robot and robot-robot groups which mainly aim at transferring learned models of idiosyncratic human interaction among robots with different sensor abilities; and calculating the accuracy loss during knowledge transfer from one robot to another.

1.5 Thesis Organization

Rest of the thesis is organized as follows:

Chapter 2 gives an overview of transfer learning and describes the previous work conducted in the shared knowledge transfer domain. It also highlights the shortcomings of these approaches to solve our research problem efficiently.

Chapter 3 explains our approach to solve the research question. It describes our proof of concept and the supporting concepts for the designing the architecture proposed in the thesis.

Chapter 4 describes the experimental setup for the architecture. It gives a detailed analysis of the dataset used and the evaluation of the architecture experiment. A comparative study with the existing methods is included to provide an insight into the accuracy of the system.

Chapter 5 concludes the results in the experimental evaluation section. It states the future scope of the research question and also discusses the approaches that could have been undertaken to reduce the loss of accuracy even further.

Chapter 2

Related Work and Background

The need for untrained users to interact with the robots has resulted in an aroused interest in the study of robots that support natural language learning [48]. Many researchers are exploring the language grounding problem, where the goal is to extract the natural language representations in the physical world.

2.1 Grounded Language Acquisition

Grounded language acquisition involves perceiving words that refer to object attributes present in physical environment. It establishes the meaning of such words by mapping them to a perceptual system. This mapping is further used to identify the physical entities with reference to the contextual representation. Research in language grounding problem has led to success in many domains like robots learning to follow human direction [76], understanding natural language commands to perform tasks [72], understanding and generating spatial references and descriptions [26], and incorporating multi modal context beyond language [5].

Many approaches are put forth to study the grounded language problem. Boteanu et al. [12] provided a framework to identify the assumptions in complex environment situations to complete the task using manually defined words. Arumugam et al. [2] defined a granular structure to provide multiple level of abstraction

for interpreting natural language commands. Williams and Scheutz [81] demonstrated an approach for domain-independent grounded language acquisition when the knowledge is obtained from multiple sources. Misra et al. [56] proposed a method to train the classifiers using reinforcement learning by directly mapping raw visual images with textual input to get the output. For this work, we assume that similar grounded language acquisition approaches provide the labels for the first stage of the learning pipeline.

Ororbia et al. [59] demonstrated that natural language is closely associated with the physical context of the object. They proved that language models trained with visual context outperforms other language models. Mavridis and Roy [51] developed a modular architecture with GSM that provides a smooth integration of language and the sensor-driven information about a particular situation. The modules combine results of language, perception, and action-related modules for grounding words in visual scenes. We use a similar approach of combining language-provided labels with visual concepts, but focus more on transferring those models.

Steels [71] proposed an operational model for robots to develop a shared grounded communication system. This led to an advancement in the domain of animatronics as well as human-robot, and robot-robot interaction.

in Kollar et al. [37], they first construct visual classifiers that can identify appropriate object properties. Then, they map the meaning of individual words to the built classifiers, then construct a model of compositional semantics to analyze the sentence as a whole. Previous work on grounded language acquisition in computer vision emphasizes on finding the meaning of a hidden word, rather than compositional

semantic analysis. Matuszek et al. [49] demonstrated an unconventional approach for jointly learning visual classifiers and semantic parsers, to produce rich, compositional models that span directly from sensors to meaning. They build probabilistic learning models from the input. The approach is built on the the existing work of probabilistic Combinatorial categorical grammar for semantic parsing [87, 39] and visual attribute classification using depth kernel descriptors [9]. Similar work was put forth in NSF NRI project "Jointly Learning Language and Affordances" using physical accordances associated with humans to learn joint models of the environmental and perceivable grounded language. By comparison, we use a similar paired data set, but the work in this thesis is assumed to occur after word meanings have been discovered.

Current models face the lack of negative examples in the training set. Pillai and Matuszek [62] discussed the solution to overcome this shortcoming by using semantic similarity or cosine similarity to automatically choose negative exemplars. In our preliminary experiments, we have used manual annotation for grounding the natural language. With related study conducted in this domain, we aim to build user-centric models over a generalized language model.

2.2 Visual classification

Visual classification identifies objects in physical environment based on the attributes taken from the visual context. Csurka et al. [17] stated the steps for visual object-based classification as:

1. Feature Extraction

2. Clustering the extracted features to form a dictionary

3. Compare the test image features with the created categorical dictionary.

Felzenszwalb et al. [22] gave an overview of the different methods that can be used for object detection and provided an algorithm that trains on partial input labels. These object recognition systems are typically based on local image descriptors using SIFT over 2D images [46], and spin images over 3D point cloud [31]. Bo et al. [9] developed kernel descriptors that are able to build models with size, shape and depth edges in a single framework. These descriptors experimentally outperformed traditional 3D spin images. Recent work in kernel descriptors leverage the kernel properties. Karmakar et al. [34] proposed an approach to increase the efficiency of the existing kernel descriptors by improving the similarity between the compared patches with respect to any pixel attribute. We used this work to learn about the visual classification pipeline and find the latest development in the classification. With this related work, our next steps would be to adopt the kernel descriptors for visual classification.

Yang et al. [85] proposed a dynamic match kernel on the top of the match kernel [8] which calculated the matching thresholds adaptively based on the pairwise distance among deep CNN features [43] between query and candidate images. In the preliminary experiments, we used linear SVMs as we were using a limited dataset. With a detailed study of the advancement done in field of visual classification, we aim to use one of these more advanced visual classification algorithm in future along

with large dataset.

2.3 Feature Extraction

An important attribute of image classification is selection and extraction of meaningful features from images. Feature extraction derives informative and non-redundant features from the initial dataset provided. But, while considering an algorithm using large data distribution with redundant data, it can be reduced to subset of important features called feature vector and the process is called feature selection. Hira and Gillies [27] provide a comparative survey on the different methods of feature selection and feature extraction algorithms. Lowe [46] provided the steps for feature extraction where Scale-space peak selection efficiently searches for all scales and image locations using Difference of Gaussian (DoG) function to identify scale and orientation invariant potential interest points. After it finds the candidate location based on point stability, a detailed model is fit to determine location, scale, and contrast. Then we assign one or more orientations to each key point location on the basis of image properties. The last step measures local image gradient at region around each key point and transforms into a representation that allows local shape distortion and change in illumination. Harris corners are used to detect interest points.

Below is a summary of the feature extraction steps [46]:

1. Scale-space peak selection

2. Key point localization

3. Orientation assignment

4. Key point descriptor

In our research, we have made our choice of algorithms to be used for feature extraction based on the referenced papers. This comparative study was important to determine if we are making a right choice for the dataset that we considered for the experiment.

2.3.1 Bag of Visual Words

Using a bag of visual words proceeds in three steps. Firstly, extract the feature descriptors from an image dataset of each category to form a visual vocabulary or bag of features. Secondly, group the descriptors iteratively into k mutually exclusive clusters. The resultant clusters are compact and distinguished by similar features where each cluster center represents feature or visual word. The third step involves vector quantization (see 2.1).

Summarizing the steps for bag of visual words,

1. Feature extraction

2. Codebook construction

3. Vector quantization

O'Hara and Draper [58] conducted a detailed survey on the literature of the bag of visual features. It also provided some insights into the open issues to be tackled in the domain of feature extraction. Nowak et al. [57] provided a detailed

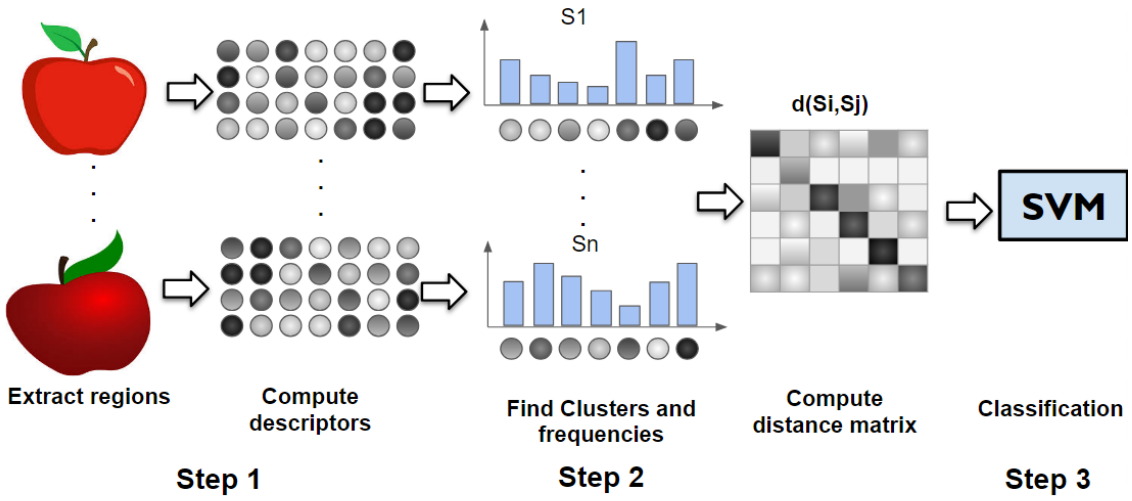

Figure 2.1: Example of bag of visual words feature extraction in image classification. In the first stage, key points are extracted from the image; the descriptors are then clustered using clustering algorithms to find the similar features in the images; we then calculate a vocabulary of visual words. Distance of features extracted from new test images are calculated from the visual vocabulary and then classified.

survey regarding the sampling strategies and the overall effect of these strategies on the performance of the system. Jurie and Triggs [33] provided an experimental survey for creating efficient codebooks. We found these surveys very useful as it contained a consolidation of work done regarding the topic 'bag-of-visual-words.' Csurka et al. [18] introduced the novel method of bag of visual words based on vector quantization [36] of affine invariant descriptors of image patches. This method has demonstrated robustness to background clutter in the exeriments described in the

paper. However, we did not use this method because our dataset consisted of focused images where there was no background clutter.

Recent advancement in the image classification [29] suggests using Local tetra patterns (LTrP) followed by bag of visual words (BoW) pattern for efficient learning of visual words [55]. Sarwar et al. [66] proposes creation of small visual vocabularies of the local intensity order pattern (LIOP) feature and local binary pattern variance (LBPV) feature to integrate to form a large vocabulary. As the primary aim of our research is establishing communication between the robots, we used the simplest bag of visual words approach to classify images. In future, we would like to see how these recent advancements help improve the image classification.

2.4 Feature Extraction Algorithms

Trivedi et al. [74] performed a comparative survey of various feature extraction algorithms, like deep learning approaches, histogram based algorithms [15, 67], color/edge based algorithms [16], textual based features [47], SIFT [46], and SURF [4] algorithms. For our research, we used the survey to compare and choose a feature extraction algorithm appropriate for the dataset considered for the experiment; in our case, we used SIFT for our experiments. SIFT performed better than SURF when we considered various features like blurring, viewpoint, rotation invariance, noise addition. The reason of improved SIFT performance over SURF may be because of more features extracted in SURF than reduced the speed of the process.

2.4.1 Scale Invariant Feature Transform(SIFT)

Lowe [46] proposed a novel method for extracting distinctive features from images which are invariant to image scale, rotation, robust matching to affine distortion to substantial range [3], addition of noise, change in 3D viewpoint, and change in illumination. SIFT provides key point matching needed to find nearest neighbor. Mikolajczyk and Schmid [54] compared the performance of many local descriptors which used recall and precision as the evaluation criterion. SIFT algorithm consists of four phases stated as Scale-space extrema detection, key point localization, orientation assignment and key point descriptor.

2.4.2 PCA-SIFT detector

Juan and Gwun [32] described PCA as a standard technique that enables us to linearly-project high-dimensional samples into a low-dimensional feature space. After applying PCA, we can apply two approaches for the final description of the SIFT key points: majority rule approach and key point histograms approach. Ke and Sukthankar [35] provided a comparative analysis of SIFT and PCA SIFT. PCA is well-suited to represent key-point patches and is more space efficient but observed to be sensitive to the registration error and non-rigid deformations. We elected not to use PCA-SIFT because our dataset contained many objects with non-rigid deformations like garlic, tomato.

2.4.3 Speeded-Up Robust Features (SURF)

Bay et al. [4] showed that we create a stack of features without 2:1 downsampling. This creates images of same resolution. SURF and SIFT show similarity in the performances while PCA-SIFT outperforms for rotation, blur and illumination changes [32]. However, SURF's sensitivity to rotation is inappropriate when a robot may view an object from any angle.

2.5 Hierarchical Matching pursuit

Bo et al. [10] proposed an architecture that constructs a layer-wise feature hierarchy using an efficient matching pursuit encoder. HMP enables linear SVM to match the performance of nonlinear SVM while providing scalability for large data set. Aharon et al. [1] developed a dictionary learning algorithm with K-SVD that uses k-means and updates dictionary sequentially. Matching Pursuit Encoder used in HMP has three modules namely Batch Tree Orthogonal Matching Pursuit, Spatial Pyramid Max Pooling and Contrast Normalization (see 2.2).

Lan et al. [41] proposed multi-channel feature dictionaries based feature learning method using two layers. The features are obtained by performing max pooling on the sparse codes of pixels in a cell. These features are concatenated to form patch features. These features learn the sparse code dictionary in second layer. Last step is to apply spatial pyramid pooling to generate the object features. This method is experimentally proved to be more efficient than other state-of-art methods. We use these HMP features for shape classification.

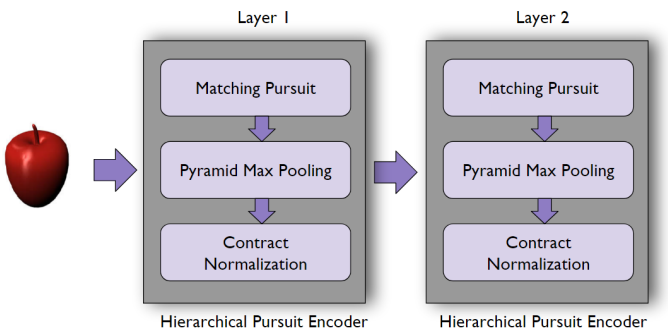

Figure 2.2: Architecture of Hierarchical Matching Pursuit.The sparse codes from small image patches are aggregated to patch-level features in the first level. While in second layer, sparse codes are aggregated to produce the image-level features. Matching pursuit computes sparse codes in each layer. Pyramid max pooling creates a pool of important features while Contrast normalization normalizes the image-patch illumination.

HMP outperforms both SIFT based single layer sparse coding and other hierarchical feature learning approaches [10]: convolutional deep belief networks [43], convolutional neural networks [38] and deconvolutional networks [83]. The algorithm creates histograms of the image dataset and works well with linear SVM [15, 10]. We are using RGB-D Dataset Collection [40] for our experiment, and Lai et al. [40] states that HMP algorithm for 3D feature extraction works best for the provided dataset.

2.6 Heterogeneous Transfer Learning

Transfer learning involves using the knowledge obtained from one domain to get results in some other domain [90] (see figure 2.3). For example, image classifiers are trained on the motorcycle classifiers. The knowledge learned by the models is then applied to classify cars. This domain has a vast applications in domains where the source domain is labeled but we lack sufficient labels in target domain.

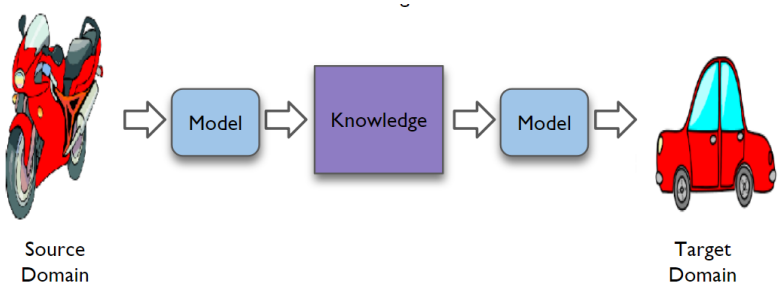

Figure 2.3: Transfer learning: Knowledge learned from one domain is applied to obtain the results in other related domain. An example: knowledge from motorcycle classifiers is used to learn about the car object.

Types of approaches.

Pan et al. [60] provides a survey for different transfer learning techniques with relation to the classification, regression, and clustering problems. They introduce the following transfer learning approaches as:

- Instance-transfer: Re-weight some labeled data in the source domain for use in the target domain [30, 44, 28, 6].

- Feature-representation-transfer: Find a feature representation that minimizes the gap between the source and the target domains and the error of classification [19, 7, 20].

- Parameter-transfer: Discover the common features between the source domain and target domain models, which can be beneficial for transfer learning [42, 11, 68, 24].

- Relational-knowledge-transfer: Build mapping of relational knowledge between the source domain and the target domains where assuptions are relaxed in the domains [53, 52, 21].

Recent approaches in the field of heterogeneous transfer learning has led to various learning approaches like asymmetric heterogeneous transfer learning [23], proactive / complete heterogeneous transfer learning [82]. Our research question focuses on transfer learning in heterogeneous domains. Related work in this domain give us an overview of the approach to be followed for transferring learned models. Our work is most similar to asymmetric heterogeneous transfer learning where we have plenty of labels in source domain but our model uses an unlabeled dataset for the target domain.

2.7 Domain Adaptation

Domain Adaptation is adapting the knowledge from more than one sources and using it to improve the performance of related task in target domain (see 2.4).

In our scenario, we consider the tasks in both the domains to be related but target domain to be unlabeled.

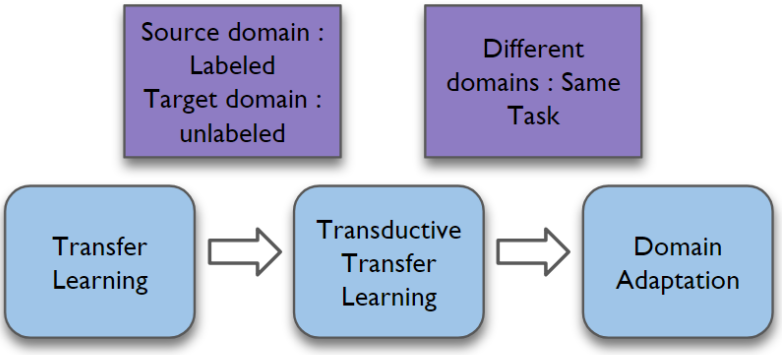

Figure 2.4: Domain adaptation: Transductive transfer learning considers flow of data from labeled source domain where the target domain is unlabeled, while domain adaptation considers the scenario where domains are different but the task to be completed is the same (here, classification).

Our approach adapts learned models to solve the new semi-related tasks using knowledge gained from the initial learned models. In hybrid heterogeneous transfer learning [89], the scenario where the feature space and task concerned with both source and target domain are not related to each other are considered. Recent advancement in heterogeneous domain adaptation suggests some approaches that use manifold alignment [25, 77], correlation subspace [86]. Tuia et al. [75] compared the different approaches and gave an overview of advancement in the field of domain

adaptation. A summary of the domain adaptation approaches is:

1. **Adaptation of the data distributions**, where classifiers are left unmodified while the source and target domain data distributions are made as similar as possible. These approaches extract a common feature space where all domain data distributions may be treated equally.

2. **Adaptation of the classifier**, where a learned classifier model from source domain is adapted to target domain by considering unlabeled samples of the target domain.

3. **Adaptation of the classifier by active learning**, where providing limited amount of smartly chosen label samples from target domain helps to perform adaptation.

Our research considers the scenario where we transfer the learned models from a labeled source domain to an unlabeled target domain, making it most similar to adaptation of the classifiers method of domain adaptation. Such a test case is considered under the study of domain adaptation. Related work introduces us to the types of domain adaptation methods and helps us to select an approach that best matches our scenario.

Chapter 3

Learning Approach

Homogeneous transfer learning is easier as compared to heterogeneous transfer learning where the source and target data distributions are completely different [79]. Supervised learning is a straightforward task. Caruana and Niculescu-Mizil [14] provides a comparative survey of various supervised learning approaches, like SVMs [78], neural nets [69], logistic regression, naive Bayes [64], memory-based learning [50], random forests, decision trees [63]. But for our experiment, we consider unsupervised environment where the target data distribution is completely unlabeled.

The accuracy of the classifiers built by the robot-trained robot is assumed to be much lower than that could have been achieved if a human would have trained it; alternatively, the target domain data distribution would have been provided with partial labels, the reason being we can not train the models with 100% accuracy over complex data models. In an elderly care scenario, where we train our robots on inputs like "Advil is medicine for fever and common cold" and show the particular bottle, the visual classifiers for human-trained robots are supposed to identify the right bottle of Advil when they see it. We, then, transfer the learned models to the robot-trained robot with no prior knowledge about the medicine.

Due to the difference of perceptual sensors, we can not directly transfer the

classifiers to the robot-trained robot. These classifiers have to be modified to adjust to the changed specifications. The robot-trained robot uses the labels from human trained robots as its ground truth. The loss of accuracy due to the transfer of models from one robot to another can affect the environment which are data sensitive like elder care.

At times, training all the robots present on the field is not a feasible option, especially for elder care domain. Each robot, deployed in the elder care facility, has to cater to the needs of individual patients. To avoid this tedious and repetitive task, sharing the models learned by one robot with other is the possible option. Though the robots have the same task at hand, the data distribution present at the source end and target end cannot be assumed to be the same.

Similarly, we can not assume that the robots are always co-located. Sharing of learned models should be possible when the robots are not physically co-located. Sharing these data models among robots with different sensors is difficult as the models cannot be transferred directly from one robot to another. They have to be modified to be used by the robot-trained robots. Bozcuoglu et al. [13] conducted a study in this domain with the assumption that systems use shared pre-knowledge about the existing concepts in their environment.

The motivation for the research question is taken from various research domains [48] like transfer learning, domain adaptation, feature adaptation [45], the evolution of language to include syntax and semantics [70] and sensor difference modeling.

To tackle this challenging problem of sharing learned visual classifiers from

one robot to multiple heterogeneous robots, we came up with the chained learning approach. Chained learning is when a human trains a robot on a specific set of classifiers and we use the shared physical presence of the other robots to transfer the visual classifiers. The trained robot provides ground truth to the new robots and the test set of the human-trained robot are used as a training set for the robot-trained robots. Later, this chain continues to impart its learned visual classifiers to other new robots. With each training session, there would be a loss in the accuracy.

In the next section, we evaluate this methodology experimentally. Our experiments are built on previous work on joint model learning based on visual classifiers and language [49, 62]. To evaluate our theory of chained knowledge transfer, we perform a proof of concept on simple classifiers like the color classifiers.

Chapter 4

Experimental Evaluation

Research in the domain of transfer learning in heterogeneous robots is highly beneficial for some extremely environment-specific domains like elderly care. Such areas require collaborative task execution where we deploy multiple robots for carrying out the subtasks. These tasks involve taking care of the patients, providing them with their medicines timely, catering to their needs. It is essential that the robots are personalized with respect to individual user needs to have a proper understanding of the patients' requirements. In such a case, training all the robots separately would be an arduous task for the patients as well as the faculty.

Our research question is transfer learning using the chained learning approach with minimum loss of accuracy; eventually moving away from the shared workspace. The experimentation involves three crucial phases: data collection, data processing, and data analysis. We perform three experiments to validate the chained learning approach. First, we check the validity of our chained learning approach with a proof of concept experiment which involves co-located robots having similar sensors; second, we conduct our experiment for co-located robots with qualitatively different sensors; third, we move away from the shared workspace where we consider non co-located robots with different source and target data distribution.

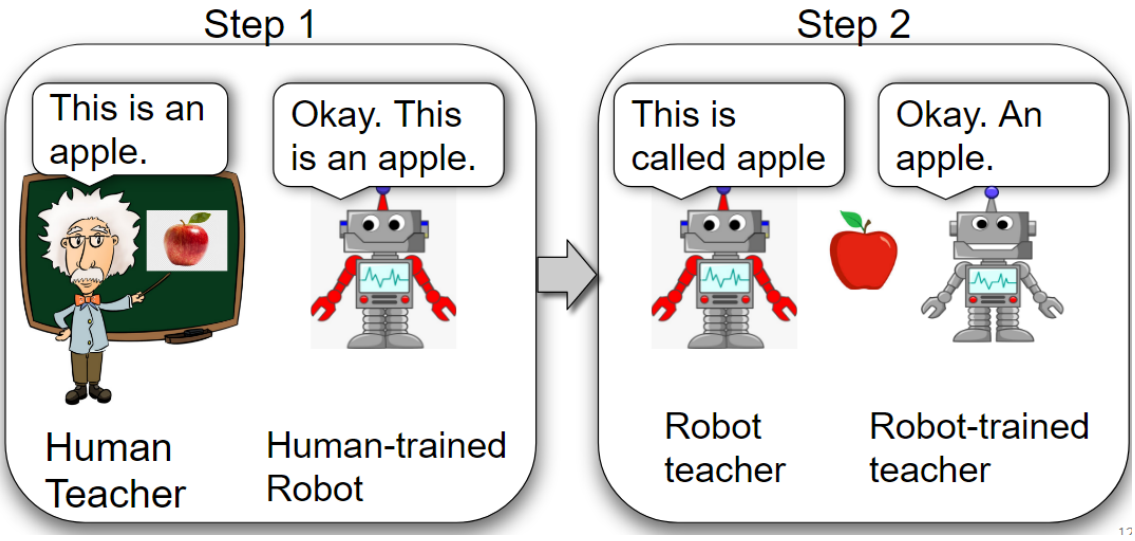

Figure 4.1: Example of chained learning approach

In the chained learning approach, a human teacher provides ground truth or labels to the human-trained robot about its physical surroundings. Human-trained robot builds its classifiers based on the examples taught by the human teacher. For example, human teacher teaches the human-trained robot the classifiers for 'apple' object. Second step of the approach is when the robot teacher provides ground truth or labels to the other robot-trained robots to build their classifiers from the knowledge it has gained from its own experience. Thus the robots can share their learned models to other robots (see 4.1). We consider this entire experiment in three scenarios:

4.0.1 Sharing learned models with co-located, similar sensor robots

In this experiment, the first robot learns the color classifiers and then transfers them further to the other robots according to the data that they perceive in the environment around them. The second robot trains its classifiers from the ground truth obtained by the first robot. Thus the second robot is gradually trained with the fully functioning linguistically meaningful perceptual learned models from the first robot (see 4.2).

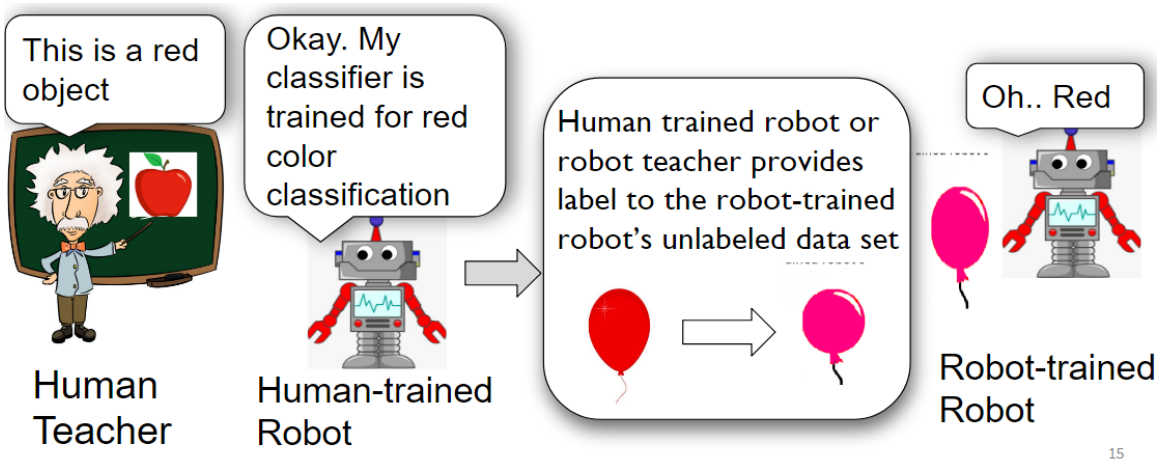

Figure 4.2: Experiment 1 setup: Sharing learned models with co-located, similar sensor robots

4.0.2 Sharing learned models with co-located, with qualitatively different sensor robots

We consider human-trained robot training its classifiers on 2D images while the robot-trained robot perceiving 3D images. Human-trained robots provide ground truth or labels to the target data set of the robot-trained robots (see 4.3). Here, the first, human-trained robot is equipped with only a camera, while the second robot-trained robot has an RGB-D sensor.

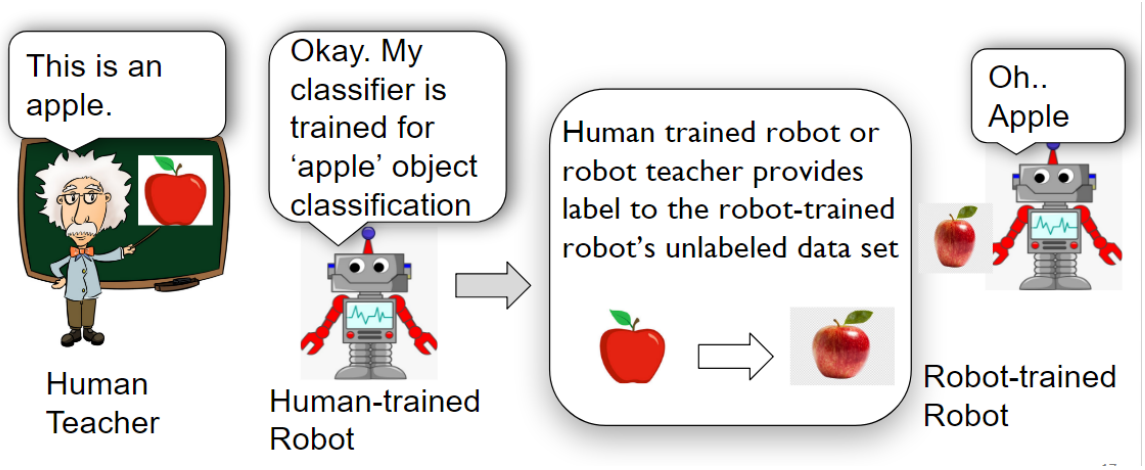

Figure 4.3: Experiment 2 setup: Sharing learned models with co-located, qualitatively different sensor robots

4.0.3 Sharing learned models with non-co-located robots

In this experiment, we consider the scenario where the teacher and the robot being trained are in different locations. Transfer of learned models takes place with

one teacher robot directly providing the robot-trained robot with learned models. Robot-trained robot has an unlabeled data set. So the robot-trained robot directly uses the classifiers provided which brings the research closer to domain adaptation (see 4.4).

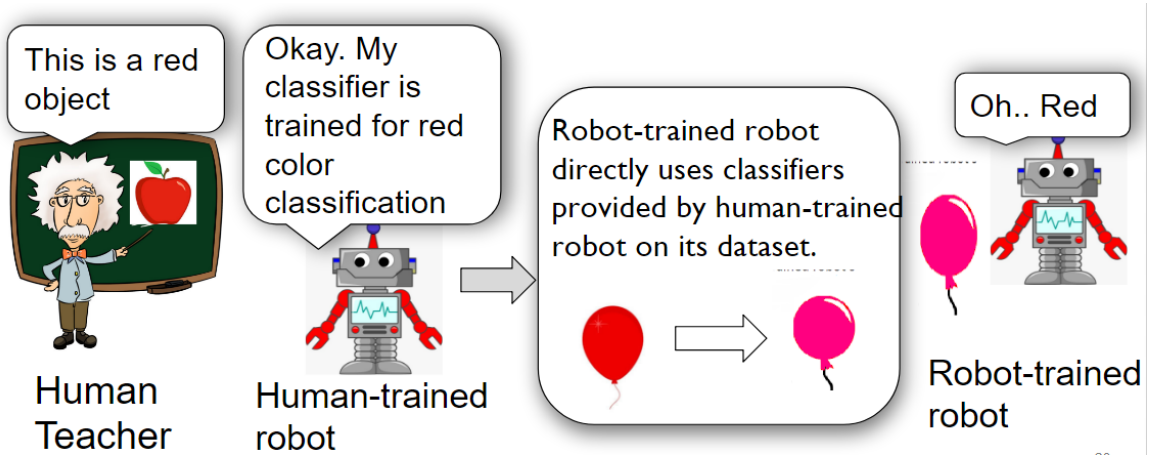

Figure 4.4: Experiment 3 setup: Sharing learned models with non-co-located robots

4.1 Datasets

For the entire experiment, we have used the open source public image repository for RGB-D images from the University of Washington RGB-D dataset [40]. The objective of using such established dataset is to enable rapid progress in building basic classifiers. Another advantage is to avoid the noise due to manual errors introduced while creating a real dataset. The dataset originally consists of 300 instances organized into 51 categories. RGB-D images include color and depth information that substantially improve quality of results.

4.2 Experiment 1: Sharing learned models in co-located similar-sensor Robots

The human-trained robot has no knowledge about the perceptual ability of the robot trained robot. Chained learning is the most straightforward approach for transfer learning in such condition. Applying this learning approach for result-sensitive domain like elder care, we have to calculate accuracy drop over the period. Our experiment checks the loss of accuracy of the transferred models as well as the scalability of the approach.

To test the correctness of our theory with sharing grounded language acquisition models with heterogeneous robots, we first trained language-denoted classifiers using images from an existing dataset [40].

4.2.1 Dataset for Experiment 1: Simple RGB Images

Our first experiment involves sharing learned models in robots with different camera sensors. We consider our human-trained robot to be perceiving the objects as perfect RGB images. For the robot-trained robot, we consider its camera be of low quality. As the research focuses on transfer learning, we assume correct labels assigned by a human teacher; in practice, these labels vary in quality [88, 80], which would reduce downstream accuracy of the pipeline.

Our dataset had a total of 8 categories. To create a dataset for our preliminary experiment, we included 15 images of each category that we are considering for color classification. From amongst these 15 images, we have 8 images of one instance and

7 images of the same object of another instance. The two instances have slight variation in appearance and texture. These 15 images form a positive sample data set for the particular color classifier(see 4.5).

Figure 4.5: Experiment 1: Data set for Co-located Similar-Sensor Robots

This pattern is followed for all the other color classifiers to form the training set for human-trained robot. Instances of other classifiers are then taken as negative samples for a particular color classifier. As we are considering simple color classification problem, we can safely state that all instances of other color classifiers can be taken as negative exemplars. With complexity in the classification problem, where a 'carrot' classifier can not be a negative example for an 'orange' classifier, we can extend on the concept on cosine similarity put forward by Pillai and Matuszek [62]. Our current dataset is highly unbalanced where we have more negative samples than the positive examples.

4.2.2 Training Dataset for proof-of-concept approach for human-trained robot

We build our learned models on linear SVM classifiers [84, 40, 15]. For testing our built classifiers, we consider instances of objects with same color but different category. For example, Lemons were replaced by bananas, tomatoes were replaced by bell peppers. Replacing the object with other objects from the similar color components allowed us to experiment with the actual color classification models being independent of the objects to be classified(see 4.6, 4.7).

Type	Color classifier	Example instances
Tomatoes	Red	15
Lemons	Yellow	15
Oranges	Orange	15
Lime	Green	15
Garlic	Gray	15
Blocks	Blue	15

Table 4.1 Training set for human-trained robot where the number suggests the number of instances of the object present in the dataset.

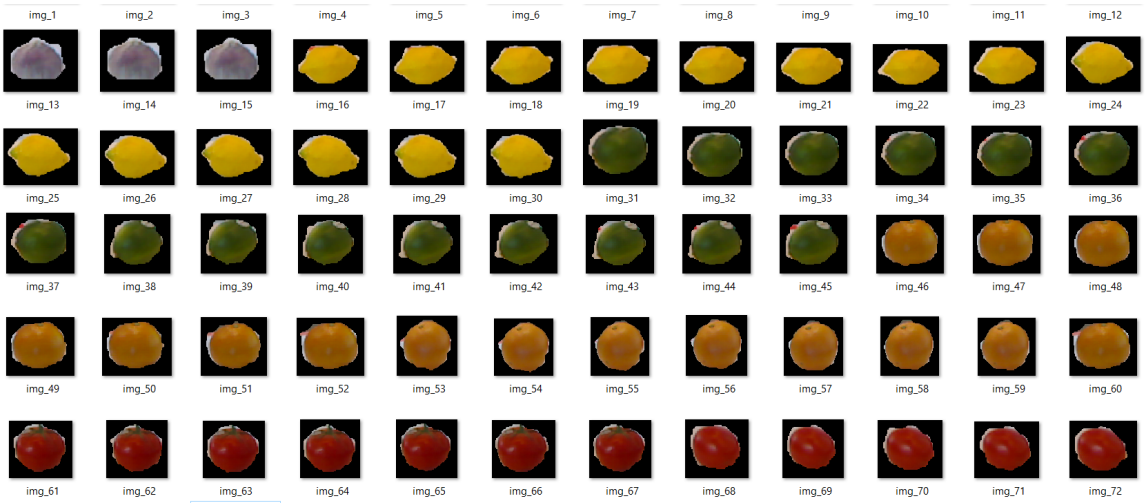

Figure 4.6: Experiment 1: Training set for Co-located Similar-Sensor Robots (visual representation).

Type	Color classifier	Example instances
Bell pepper	Red	15
Bananas	Yellow	15
Oranges	Orange	15
Greens	Green	15
Garlic	Gray	15
Blocks	Blue	15

Table 4.2 Test set for human-trained robot where there are fifteen of each object present in the dataset.

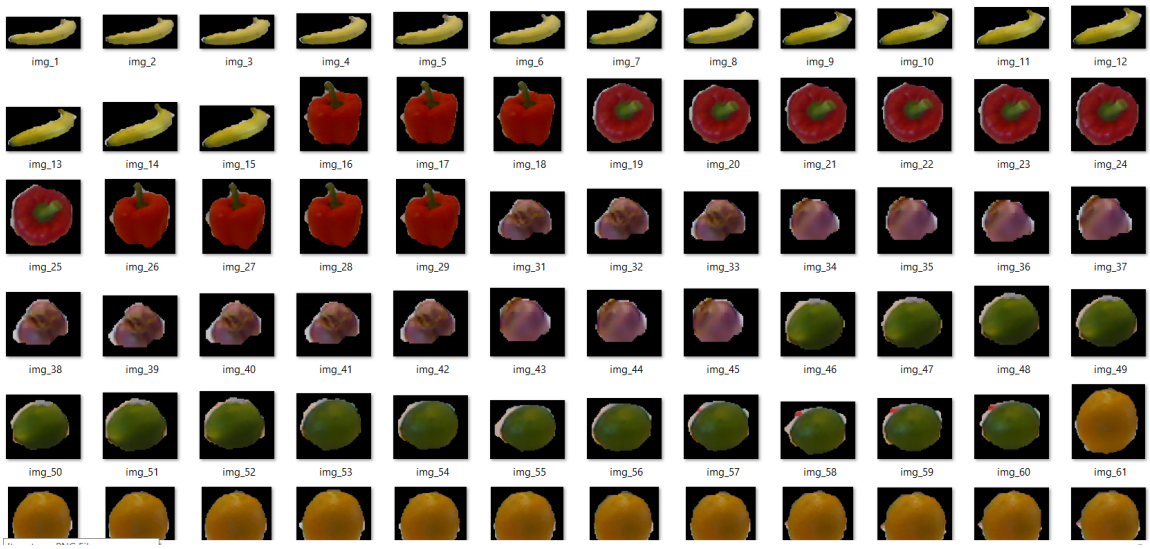

Figure 4.7: Experiment 1: Test set for Co-located Similar-Sensor Robots (visual representation).

4.2.3 Simulation for Experiment 1: Co-located Similar-Sensor Robots

For first experiment, we considered sharing learned models with robots with different sensor quality where our robot-trained robot has low quality camera(see 4.8, 4.9). We considered different instances of objects for training and testing of human-trained robot—for example, different instances of bananas and lemons were considered for the 'yellow' color classifier. Our simulation of chained learning approach involved human-trained robot teaching the robot-trained robot about the instances of colors it sees around itself. The human-trained robot classifies these images and provides the ground truth to the robot-trained robot. For simulating the teaching scenario, we created identical datasets for the two robots. Plain RGB images in the dataset were then labeled by human-trained robot and labels were

provided to the robot-trained robot. The images in the dataset were now converted to HSV image format. We simulated the low quality camera simulation by decreasing the hue and saturation of the images in the dataset. An instance of some of the modification in the database created are shown in figures 4.8 and 4.9:

Figure 4.8: Experiment 1: Instance of banana as seen for human-trained and robot-trained robot.

Figure 4.9: Experiment 1: Instance of bell pepper as seen for human-trained and robot-trained robot.

The robot-trained robot is trained on the low quality images that we created in the database and labels provided by the human-trained robot. With the classifiers built for robot-trained robot using SVM classifiers [88], we tested them against a test dataset created of different instances of object in training, simulating as if the instances of images were taken from a low quality camera.

For conducting the experiment for chained learning approach, we chose MATLAB. It has many built-in functions that are tested over years providing reliability. We built color classifiers for orange, green, gray, blue, yellow, red. Building the models for binary classification involved using linear SVM with 4-fold cross validation. Feature vector obtained from the dataset was an average of the R-G-B components of all pixels values for each image. To overcome the problem of background noise and lighting effect, we decided to apply a binary mask on the images in the dataset before we run any classifiers. Including more instances of negative samples to the training samples also significantly improved the results.

4.3 Experiment 2: Sharing learned models in Qualitatively Different, Co-located Sensors

Transferred learning in heterogeneous robots with qualitatively different sensors is a challenge as compared to the transfer learning in robots with different physical sensors. To test our approach on complex data models, consider a scenario where the two heterogeneous robots have qualitatively different camera. In this experiment, we consider a more complex task domain. Object identification is cor-

rectly identifying the object in the images. For the purpose of experimentation, we assume that the object to be identified covers the entire image size. Human-trained robot is perceiving its environment in the form of 2D image while the robot-trained robot perceives it in the form of 3D images (see 4.10, 4.11). So our input to the first robot are basically flat images.

4.3.1 Dataset for Experiment 2: 2D and 3D Images

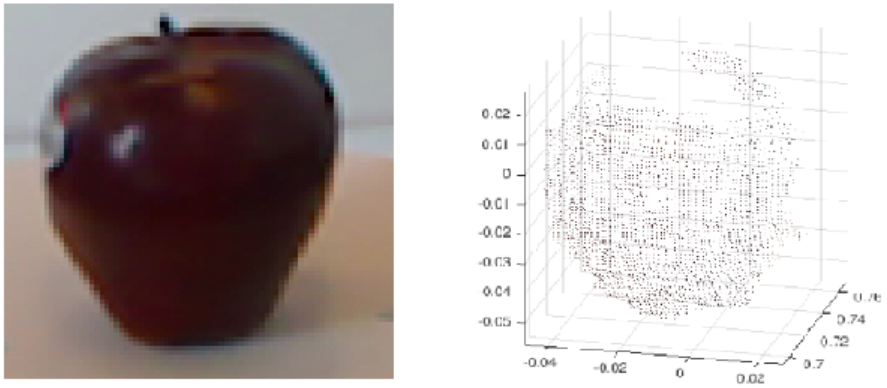

Figure 4.10: Experiment 2: Instance of an apple as seen for human-trained and robot-trained robot.

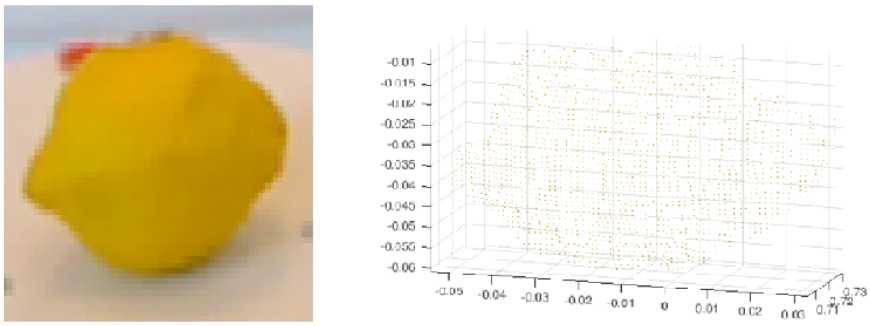

Figure 4.11: Experiment 2: Instance of an lemon as seen for human-trained and robot-trained robot.

Our dataset consists of 10 object categories namely tomatoes, apples, limes, lemons, oranges, staplers, laptops, cereal boxes, napkins, soda cans(see figure 4.12). We used the established image dataset from University of Washington containing RGB-D images [40]. We spliced the total images into training set and test set for the human-trained robot into half of the total number of images. Images considered for the experiment were unmasked and taken with more focus on the object. With a comparative study of the feature extraction tools at hand [32, 61, 35], we decided to use the SIFT feature extraction tool for our transfer learning scenario. The training set containing about 8910 images are applied to SIFT algorithm to extract features from all the images. Total of 1.5 million features were extracted. The algorithm then selected the strongest 80% features from each object category of the feature set and applied k-means clustering on it. In order to improve clustering, we balanced out the number of features in each category by selecting the category with the lowest number of features extracted. We conducted experiments with the number

of clusters to be used in k-means clustering algorithm.

Figure 4.12: Dataset for experiment 2

Number of clusters	Within group distance with centroid (avg)
100	100
200	62
300	51
400	47
500	25
1000	25
2000	24

Table 4.3 Experiment 2: Determining K value for K-means clustering

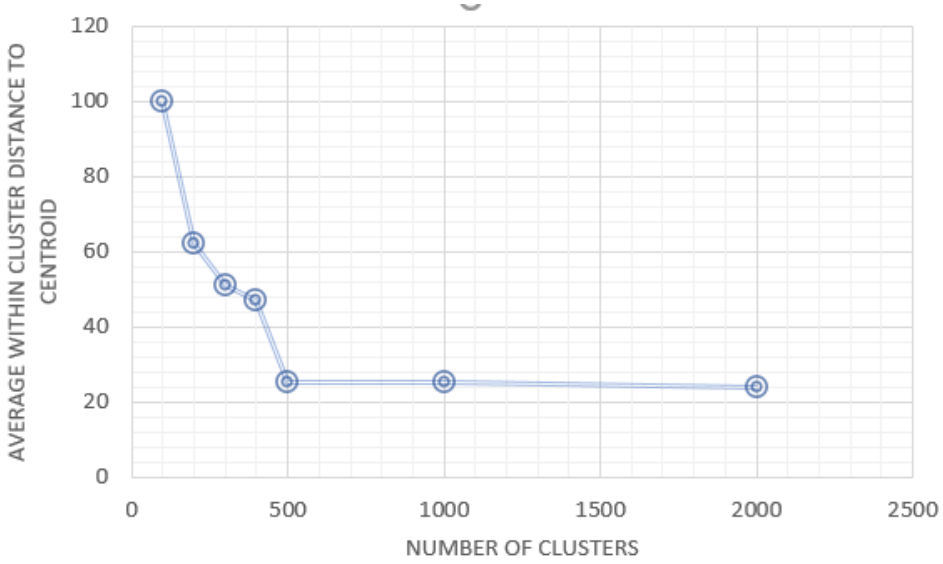

Figure 4.13: Graphical representation for k-means clustering result. We selected k using the elbow method.

We selected 500 because very small number of 'k' will not categorize the objects properly while large number of k may lead to over fitting of the model and will not be able to generalize the category over variety of objects(see 4.13). This entire process reduces the number of features generated to a 'bag of features' consisting of 'k' visual words. We then trained the robots using linear SVM and the generated bag of feature.

4.3.2 Second part of experiment 2: 3D object classification

With the first part of our experimentation working, we considered the robot-trained robot which takes input as 3D images. An important step of object recognition task is to extract expressive features from high-dimensional structured data.

Hierarchical Matching Pursuit (HMP) [40, 10] extracts expressive features and gives good accuracy even for linear SVMs. This makes the feature extraction algorithm perfect fit for our experiment. As the 3D images are large data files which take a lot of computing time, we consider only a subset of the entire dataset for our experiment.

Number of files processed per category	Features extracted	Accuracy(%)
625	500	100
625	50	99.28
625	5	95.46
100	500	100
100	50	98
100	5	96.5
10	500	100
10	50	100
10	5	100

Table 4.4 Experiment 2: Comparison for various values of number of files, features extracted and accuracy to select the approximate values for HMP in images selection

From the following table, we concluded that around 100 images in each category and 50 extracted features would be a better fit for the feature extraction pipeline. To maintain a uniformity over the entire dataset we randomly extracted 100 images from each category for the object recognition experiment with 2D images

as well. We then applied a 4-fold cross validation on the 3D dataset to give a total accuracy of 94%. To improve the accuracy of 2D object identification, we applied 4 fold cross-validation on the dataset. The resultant accuracy of the model was approximately 98%. To simulate the environment as if a human-trained robot with 2D sensors is transferring models to a robot-trained robot with 3D sensors, we replicated the procedure we carried out for the first experiment. We created a dataset of image instances with depth information where each category had 100 images. We took similar instances of objects from different rotational angle to simulate as if human trained robot is seeing those images. We labeled the 3D image set with the result obtained from 2D images and provided them as an input to the 3D classifier in our pipeline. Accuracy of the classifier dropped to 89.6% when the labels were taken from object identification using sift + SVM classifiers. We provided a pipeline for sharing learned models when robots are co-located but have qualitatively different sensors.

4.4 Experiment 3: Sharing learned models in Non Co-located Robots

We cannot always assume that the human-trained robot and the robot-trained robot would be in the same physical environment. Domain adaptation [75] transfers the learned models from one robot to another without presence of actual shared physical workspace. In this experiment, we considered the same dataset as the first experiment where the robots had similar sensors. The human-trained robot trains on the modified image dataset, that is, simulated images taken from a low resolution

camera. The learned models are then directly transferred to the robot-trained robot where the image dataset consists of original RGB images. The robot-trained robot tries to classify the images in its dataset using the provided classifiers. The result was thus collected. For the subsection of this experiment, we try to determine if the approach is efficient if transfer of classifiers is done from an original RGB camera sensor robot to modified low quality sensor robot. or vice versa.

Chapter 5

Results

In this section, we analyze the outcome of the three experiments conducted. A detailed description of the experiments provided in the previous section gives an insight into the experiments conducted with the three types of scenarios.

5.1 Results for experiment with collocated, similar sensor robots

To avoid the errors due to background noise and lighting effect, we masked the images in the dataset. Masking allowed the classification to be focused on the object rather than on surrounding environment. After applying the binary mask on the images, overall accuracy increased from 73% approximately to 88.27%. With improved results, we included a few more classifiers and images in the database. Scaling the dataset did not affect the accuracy of the trained model for human-trained robot. With our first classifier in place, we simulated the environment for robot-trained robot and calculated the loss of accuracy over the chained transfer. The accuracy dropped further to 79.05% with just two levels of chaining.

		Classifiers					
Ground Truth	KNOWN	RED	YELLOW	ORANGE	GREEN	BLUE	GRAY
	RED	15	0	0	0	0	0
	YELLOW	0	13	0	0	0	0
	ORANGE	4	0	15	0	0	0
	GREEN	0	5	0	9	0	8
	BLUE	0	0	0	0	15	0
	GRAY	0	0	0	0	0	15

Table 5.1 Experiment 1: Results of independent binary classifiers (columns) on a held-out test set of objects (rows). The largest confusion is between red and orange objects, due to color similarities in the test set, and blue and gray objects, because the objects labeled 'gray' were inconsistently colored. Each cell gives the number of objects of that type classified as positive by that classifier.

Table 5.1 shows the confusion matrix for the step 1 of Chained Learning approach for the experiment 1. From this table, we analyze that red and blue color classifiers performed poorly as compared to other classifiers. Though gray did not mis-classify any object but it did have a poor performance on recognizing gray objects.

Some reasons for misclassification:

As some the tomatoes were ripen, some tinge of orange color was observed on the skin of ripen tomatoes and hence they were misclassified as oranges(see 5.1).

Figure 5.1: Reason for confusion in red classifier

Gray image instances had variation in color where the shades of purple varied in the garlic example. An example of this variation can be seen in the figure 5.2.

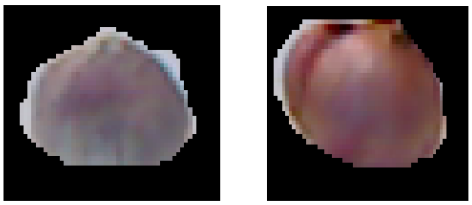

Figure 5.2: Reason for confusion in gray classifier

Ground Truth	Classifiers					
KNOWN	RED	YELLOW	ORANGE	GREEN	BLUE	GRAY
RED	15	0	3	0	0	0
YELLOW	0	10	5	0	0	0
ORANGE	6	0	15	0	0	0
GREEN	0	4	0	12	0	8
BLUE	0	0	0	0	15	0
GRAY	0	0	0	0	0	15

Table 5.2 Experiment 1: Results of independent binary classifiers (columns) on a held-out test set of objects (rows). The largest confusion is between red, yellow and orange objects, due to color similarities in the test set, and blue and gray objects, because the objects labeled 'gray' were inconsistently colored. Each cell gives the number of objects of that type classified as positive by that classifier.

Table 5.2 provides a confusion matrix for the step 2 of chained learning algorithm. The misclassification rate increased when robot-trained robot is provided labels as ground truth from human-trained robot. Here, blue classifier behaves abnormally due to the similarity in the feature vector(R-G-B) of the two color images(see 5.3).

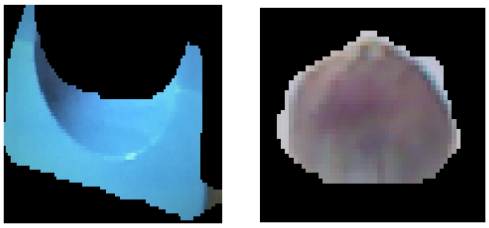

Figure 5.3: Reason for confusion in blue and gray classifier

Some common reasons of misclassification are:

Orange, red and yellow images generate much similar shades when they are modified as per the color sensor of robot trained robot. An instance can be seen in figure 5.4.

Figure 5.4: Reason for confusion in orange, red and yellow classifiers

Also, initial training dataset had background which highly confused the system. Masking the images helped to improve the accuracy of the classifiers greatly. The problem that still persists in the system is effect of lighting on the classifiers. In the final experiment, an accuracy of almost 88.27% was attained which further reduced on applying the 'chained learning approach to 79.05%.

5.2 Results for experiment with collocated, qualitatively different sensor robots

For the experiment involving chained learning where models were transferred between robots with different dimensionality, main reason for the fluctuating accuracy was consideration of huge dataset. . Also the number of features extracted were large. This led to over fitting of the resultant dataset. With decrease in the number of images in each category, we conducted the experiment with realistic results.

We got an accuracy of about 98% when we used the training set while 80% accuracy on evaluation with the test set data. We generated a confusion matrix for the test data.

		Classifiers									
	KNOWN	APPLE	CEREAL_BOX	HAND-TOWEL	KEYBOARD	LEMON	LIME	ORANGE	SODA CAN	STAPLER	TOMATO
Ground Truth	APPLE	0.41	0	0	0	0.35	0.02	0	0	0	0.23
	CEREAL_BOX	0	0.96	0.02	0.01	0.01	0	0	0	0	0
	HAND-TOWEL	0	0	0.98	0.01	0	0	0	0	0	0
	KEYBOARD	0	0	0	1	0	0	0	0	0	0
	LEMON	0	0	0	0	0.98	0.01	0	0	0	0
	LIME	0	0	0	0	0.03	0.57	0	0	0.01	0.39
	ORANGE	0	0	0	0	0	0	1	0	0	0
	SODA CAN	0.01	0	0	0	0.02	0.01	0	0.96	0	0
	STAPLER	0	0	0	0	0.03	0.2	0.01	0	0.61	0.14
	TOMATO	0.27	0	0	0	0	0.16	0	0	0.01	0.56

Table 5.3 Experiment 2: Results of independent binary classifiers (columns) on a held-out test set of objects (rows). The largest confusion is between apple and tomato objects, due to appearance similarities in the test set, and stapler objects, because the objects labeled 'stapler' were with inconsistent shape. Each cell gives the percentage of objects of that type classified as positive by that classifier.

The above experiment for object classification on 2D images gave a precision of 0.8033, recall of 0.6453 while the sensitivity of the classifier was about 0.8193. The overall model accuracy is 0.8033 for SIFT feature extraction while the f-score measure is about 0.7157.

Some of the reasons for misclassification: Apple and tomato classifiers showed the highest misclassification rate due to similarity in shapes, feature vectors and variation in the color of instances of these objects(see 5.5).

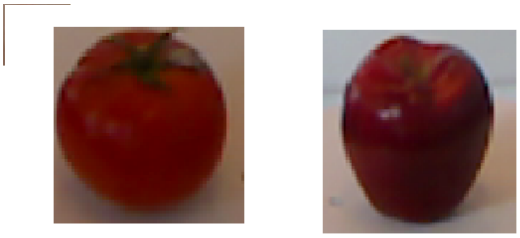

Figure 5.5: Misclassification in tomato (left) and apple (right) instances

Mis-classification in 2D images due to illumination is avoided by applying SIFT algorithm which is invariant to illumination. Stapler was not able to identify stapler objects because of the variety in the shapes of stapler instances used in the dataset(see 5.6).

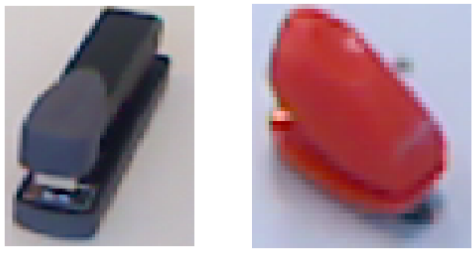

Figure 5.6: Misclassification in stapler instances

This experimentation was further expanded to the different dimensionality, where we trained our model on the 2D images with RGB values and expanded the result to include RGB-D features of the resultant 3D images. Accuracy of the pipeline with perfect labeling was calculated to be 98% which dropped to 89.6% on providing ground truth from the human-trained robot.

Thus, we could transfer learned models between two robots with varying camera quality but the process is lossy.

5.3 Results for experiment with non co-located robots

		Classifiers					
	KNOWN	RED	YELLOW	ORANGE	GREEN	BLUE	GRAY
Ground Truth	RED	15	0	15	15	0	0
	YELLOW	0	15	15	15	0	0
	ORANGE	15	0	15	15	0	0
	GREEN	15	15	15	15	0	0
	BLUE	0	3	0	0	15	0
	GRAY	15	0	0	0	0	0

Table 5.4 Experiment 3: Results of independent binary classifiers (columns) on a held-out test set of objects (rows). We see the worst performance for orange and green classifiers, while gray was not able to classify any object labeled as 'gray' Each cell gives the number of objects of that type classified as positive by that classifier.

In this experiment, we consider two non co-located robots to share learned models where the human-trained robot and robot-trained robots have different sensors. As the classifiers were trained on modified image set, which had hue and saturation reduced to half, and then were transferred to the robot-trained robot, the rate of misclassification was high. For the purpose of simplicity of the experiment, we have made some assumptions like learned models are built on classifiers that are not tuned. Also, considered the hue/saturation/resolution reduced to half as an arbitrary case. We did not experiment with the different values of hue/saturation/resolution.

	RED	YELLOW	ORANGE	GREEN	BLUE	GRAY
Precision	0.25	0.454	0.25	0.25	1	0
Recall	1	1	1	1	1	0
F measure	0.4	0.624	0.4	0.4	1	0

Table 5.5 Experiment 3: Precision, recall and f measure values for each classifier. We see the worst performance for almost all the classifiers except 'blue'.

We think that the reason for poor performance of remaining classifiers is due to lack of labels for the robot-trained robot. Some reasons for misclassification:

Modified images of red and orange had a similarity in the color vector. The presence of a tinge of red in garlic instance is the major reason for mis-classification of garlic by red classifier. Modified images of gray were as good as feature vector of white classifier. Thus, it was not able to identify any of the color correctly. Main reason for not getting a good accuracy is because we used the images from low quality camera to train the classifiers and used them over original images without any modification. So the color classifier were not the original representation of the colors. This was the first step of our experiment to transfer learned models in non co-located robots. Future steps of the experiment include assigning some labels to the robot trained robot and to see how the classifiers behave for such a scenario. In the experiment, if we were to consider all the objects to be true examples, the accuracy of the transferred learning would have been 12.5% but as we are using the base prior knowledge of previous classifier to train the robot-trained robot, accuracy

of the experiment is about 67.22%. We calculate the precision, recall and f-score measure as 0.329, 0.833 and 0.471 respectively. From these values we conclude that the classifiers returned almost all relevant results. While we got a low precision means it was able to predict many irrelevant objects as well. F1 score of the entire experiment is 0.471.

For the second part of this scenario, we considered a transfer of learned models from ordinary RGB camera to a low resolution camera.

		Classifiers					
	KNOWN	RED	YELLOW	ORANGE	GREEN	BLUE	GRAY
Ground Truth	RED	0	8	0	0	0	15
	YELLOW	0	15	0	0	0	15
	ORANGE	0	0	0	0	0	0
	GREEN	0	15	15	0	0	15
	BLUE	0	4	0	0	15	13
	GRAY	0	15	0	0	0	15

Table 5.6 Experiment 3: Results of independent binary classifiers (columns) on a held-out test set of objects (rows). We see the worst performance for almost all the classifiers except 'blue'. Each cell gives the number of objects of that type classified as positive by that classifier.

	RED	YELLOW	ORANGE	GREEN	BLUE	GRAY
Precision	0	0.263	0	0	1	0.263
Recall	0	1	0	0	1	0
F measure	0	0.416	0	0	1	0.416

Table 5.7 Experiment 3: Precision, recall and f measure values for each classifier. We see the worst performance for almost all the classifiers except 'blue'.

In this experiment, we considered the scenario where we are transferring the classifiers trained on RGB images directly to be tested on modified images. We calculate the precision, recall and f-score measure as 0.310, 0.5 and 0.861 respectively. Though both the tests perform equally bad in both the scenarios, from the f-score measure we conclude that transfer learning performs better when we transfer the classifiers from original good quality camera to a low quality camera rather than conducting the transfer vice versa. Though this was overall a failed attempt at transfer learning, we learned that providing labels may be a better point of improvement in the experiment.

Chapter 6

Conclusion and Future Work

Transfer learning in heterogeneous robot environment is an unexplored topic [48]. We conducted three experiments concerning our research question to validate the possibility of sharing the learned models in any conditions between heterogeneous robots. Our first experiment considered the scenario where two communicating robots were collocated and had similar sensors. The robot-trained robot had a low resolution camera that made the source and target domain have different data distributions. Physical presence of the human-trained robot to teach the robot-trained robot proved to be a major advantage for labeling the target data. In the second experiment, we repeated the same experiment of sharing learned models between robots with qualitatively different sensors. While in the third experiment, the two robots communicating their learned models were considered to be at different locations. By conducting the experiments, we conclude that it is possible to transfer the learned models between robots with different sensors irrespective of if they are non collated or they share a common workspace. Sharing of learned models has shown to give a better accuracy as compared to self learning systems [65]. 'Chained learning approach' introduced in the dissertation lays a foundation for research in the shared learning domain. Scalability of the dataset does not affect the loss of accuracy over the pipeline.

As this area of research lies unexplored, there is plenty of scope for future work. For the purpose of experiment, we stuck to the default value for threshold of the classifiers. We did not experiment enough with the values. We can conduct the experiment with many different values to see how that affects the loss of accuracy of the model. Also, we did not experiment with the hue, saturation and resolution values. So we can see the function of how the change in quality of images have an effect on the experiment performance. We can also, combine the work on joint model learning [49, 87] and negative exemplars [62] to build better learned models to be transferred. Domain adaptation [25] is emerging as an interesting field of research. For this research, we conducted the experiments by simulating the environment in which the two robots communicate with each other. But in future, we plan to use the real dataset collected from actual robot sensors and process the entire 'chained learning approach' in real robot environment. Thiel et al. [73] discussed the novel idea of collaborative work in hospital environment where multiple robots efficiently work together to take care of patients by distributing and scheduling tasks. With sharing of the learned models between robots, this task can be accomplished with minimum human interaction.

Bibliography

[1] Michal Aharon, Michael Elad, Alfred Bruckstein, et al. K-svd: An algorithm for designing overcomplete dictionaries for sparse representation. *IEEE Transactions on signal processing*, 54(11):4311, 2006.

[2] Dilip Arumugam, Siddharth Karamcheti, Nakul Gopalan, Lawson LS Wong, and Stefanie Tellex. Accurately and efficiently interpreting human-robot instructions of varying granularities. *arXiv preprint arXiv:1704.06616*, 2017.

[3] Adam Baumberg. Reliable feature matching across widely separated views. In *Computer Vision and Pattern Recognition, 2000. Proceedings. IEEE Conference on*, volume 1, pages 774–781. IEEE, 2000.

[4] Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. Surf: Speeded up robust features. In *European conference on computer vision*, pages 404–417. Springer, 2006.

[5] Tony Belpaeme, Paul Baxter, Robin Read, Rachel Wood, Heriberto Cuayáhuitl, Bernd Kiefer, Stefania Racioppa, Ivana Kruijff-Korbayová, Georgios Athanasopoulos, Valentin Enescu, et al. Multimodal child-robot interaction: Building social bonds. *Journal of Human-Robot Interaction*, 1(2):33–53, 2013.

[6] Steffen Bickel, Michael Brückner, and Tobias Scheffer. Discriminative learning for differing training and test distributions. In *Proceedings of the 24th international conference on Machine learning*, pages 81–88. ACM, 2007.

[7] John Blitzer, Ryan McDonald, and Fernando Pereira. Domain adaptation with structural correspondence learning. In *Proceedings of the 2006 conference on empirical methods in natural language processing*, pages 120–128. Association for Computational Linguistics, 2006.

[8] Liefeng Bo and Cristian Sminchisescu. Efficient match kernel between sets of features for visual recognition. In *Advances in neural information processing systems*, pages 135–143, 2009.

[9] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Depth kernel descriptors for object recognition. In *Intelligent Robots and Systems (IROS), 2011 IEEE/RSJ International Conference on*, pages 821–826. IEEE, 2011.

[10] Liefeng Bo, Xiaofeng Ren, and Dieter Fox. Hierarchical matching pursuit for image classification: Architecture and fast algorithms. In *Advances in neural information processing systems*, pages 2115–2123, 2011.

[11] Edwin V Bonilla, Kian M Chai, and Christopher Williams. Multi-task gaussian process prediction. In *Advances in neural information processing systems*, pages 153–160, 2008.

[12] Adrian Boteanu, Jacob Arkin, Siddharth Patki, Thomas Howard, and Hadas Kress-Gazit. Robot-initiated specification repair through grounded language interaction. *arXiv preprint arXiv:1710.01417*, 2017.

[13] Asil Kaan Bozcuoglu, Gayane Kazhoyan, Yuki Furuta, Simon Stelter, Michael Beetz, Kei Okada, and Masayuki Inaba. The exchange of knowledge using cloud robotics. *Robotics and Automation Letters*, 3(2):1072–1079, April 2018. doi: 10.1109/LRA.2018.2794626.

[14] Rich Caruana and Alexandru Niculescu-Mizil. An empirical comparison of supervised learning algorithms. In *Proceedings of the 23rd international conference on Machine learning*, pages 161–168. ACM, 2006.

[15] Olivier Chapelle, Patrick Haffner, and Vladimir N Vapnik. Support vector machines for histogram-based image classification. *IEEE transactions on Neural Networks*, 10(5):1055–1064, 1999.

[16] Yongsung Cheon and Chulhee Lee. Color edge detection based on bhattacharyya distance. In *14th International Conference on Informatics in Control, Automation and Robotics, ICINCO 2017*. SciTePress, 2017.

[17] Gabriella Csurka, Christopher Dance, Lixin Fan, Jutta Willamowski, and Cédric Bray. Visual categorization with bags of keypoints. In *Workshop on statistical learning in computer vision, ECCV*, volume 1, pages 1–2. Prague, 2004.

[18] Gabriella Csurka, Christopher R. Dance, Lixin Fan, Jutta Willamowski, and Cdric Bray. Visual categorization with bags of keypoints. In *In Workshop on Statistical Learning in Computer Vision, ECCV*, pages 1–22, 2004.

[19] Wenyuan Dai, Gui-Rong Xue, Qiang Yang, and Yong Yu. Co-clustering based classification for out-of-domain documents. In *Proceedings of the 13th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 210–219. ACM, 2007.

[20] Hal Daumé III. Frustratingly easy domain adaptation. *arXiv preprint arXiv:0907.1815*, 2009.

[21] Jesse Davis and Pedro Domingos. Deep transfer via second-order markov logic. In *Proceedings of the 26th annual international conference on machine learning*, pages 217–224. ACM, 2009.

[22] Pedro F Felzenszwalb, Ross B Girshick, David McAllester, and Deva Ramanan. Object detection with discriminatively trained part-based models. *IEEE transactions on pattern analysis and machine intelligence*, 32(9):1627–1645, 2010.

[23] Magda Friedjungová and Marcel Jiřina. An overview of transfer learning focused on asymmetric heterogeneous approaches. In *International Conference on Data Management Technologies and Applications*, pages 3–26. Springer, 2017.

[24] Jing Gao, Wei Fan, Jing Jiang, and Jiawei Han. Knowledge transfer via multiple model local structure mapping. In *Proceedings of the 14th ACM SIGKDD international conference on Knowledge discovery and data mining*, pages 283–291. ACM, 2008.

[25] Raghuraman Gopalan, Ruonan Li, and Rama Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 999–1006. IEEE, 2011.

[26] Sergio Guadarrama, Lorenzo Riano, Dave Golland, Daniel Go, Yangqing Jia, Dan Klein, Pieter Abbeel, Trevor Darrell, et al. Grounding spatial relations for human-robot interaction. In *Intelligent Robots and Systems (IROS), 2013 IEEE/RSJ International Conference on*, pages 1640–1647. IEEE, 2013.

[27] Zena M Hira and Duncan F Gillies. A review of feature selection and feature extraction methods applied on microarray data. *Advances in bioinformatics*, 2015, 2015.

[28] Jiayuan Huang, Arthur Gretton, Karsten M Borgwardt, Bernhard Schölkopf, and Alex J Smola. Correcting sample selection bias by unlabeled data. In *Advances in neural information processing systems*, pages 601–608, 2007.

[29] Hervé Jégou, Matthijs Douze, and Cordelia Schmid. Recent advance in image search.

[30] Jing Jiang and ChengXiang Zhai. Instance weighting for domain adaptation in nlp. In *Proceedings of the 45th annual meeting of the association of computational linguistics*, pages 264–271, 2007.

[31] Andrew E Johnson and Martial Hebert. Using spin images for efficient object recognition in cluttered 3d scenes. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (5):433–449, 1999.

[32] Luo Juan and Oubong Gwun. A comparison of sift, pca-sift and surf. *International Journal of Image Processing (IJIP)*, 3(4):143–152, 2009.

[33] Frederic Jurie and Bill Triggs. Creating efficient codebooks for visual recognition. In *Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, volume 1, pages 604–610. IEEE, 2005.

[34] Priyabrata Karmakar, Shy Wei Teng, Dengsheng Zhang, Ying Liu, and Guojun Lu. Improved kernel descriptors for effective and efficient image classification. In *Digital Image Computing: Techniques and Applications (DICTA), 2017 International Conference on*, pages 1–8. IEEE, 2017.

[35] Yan Ke and Rahul Sukthankar. Pca-sift: A more distinctive representation for local image descriptors. In *Computer Vision and Pattern Recognition, 2004. CVPR 2004. Proceedings of the 2004 IEEE Computer Society Conference on*, volume 2, pages II–II. IEEE, 2004.

[36] Teuvo Kohonen. Improved versions of learning vector quantization. In *Neural Networks, 1990., 1990 IJCNN International Joint Conference on*, pages 545–550. IEEE, 1990.

[37] Thomas Kollar, Jayant Krishnamurthy, and Grant P Strimel. Toward interactive grounded language acqusition. In *Robotics: Science and systems*, volume 1, pages 721–732, 2013.

[38] 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.

[39] Tom Kwiatkowski, Luke Zettlemoyer, Sharon Goldwater, and Mark Steedman. Lexical generalization in ccg grammar induction for semantic parsing. In *Proceedings of the conference on empirical methods in natural language processing*, pages 1512–1523. Association for Computational Linguistics, 2011.

[40] Kevin Lai, Liefeng Bo, Xiaofeng Ren, and Dieter Fox. A large-scale hierarchical multi-view rgb-d object dataset. In *Robotics and Automation (ICRA), 2011 IEEE International Conference on*, pages 1817–1824. IEEE, 2011.

[41] Xiaodong Lan, Qiming Li, Mina Chong, Jian Song, and Jun Li. Multi-channel feature dictionaries for rgb-d object recognition. In *Ninth International Conference on Graphic and Image Processing (ICGIP 2017)*, volume 10615, page 1061515. International Society for Optics and Photonics, 2018.

[42] Neil D Lawrence and John C Platt. Learning to learn with the informative vector machine. In *Proceedings of the twenty-first international conference on Machine learning*, page 65. ACM, 2004.

[43] Honglak Lee, Roger Grosse, Rajesh Ranganath, and Andrew Y Ng. Convolutional deep belief networks for scalable unsupervised learning of hierarchical representations. In *Proceedings of the 26th annual international conference on machine learning*, pages 609–616. ACM, 2009.

[44] Xuejun Liao, Ya Xue, and Lawrence Carin. Logistic regression with an auxiliary data source. In *Proceedings of the 22nd international conference on Machine learning*, pages 505–512. ACM, 2005.

[45] Mingsheng Long, Yue Cao, Jianmin Wang, and Michael I Jordan. Learning transferable features with deep adaptation networks. *arXiv preprint arXiv:1502.02791*, 2015.

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

[47] Bangalore S Manjunath and Wei-Ying Ma. Texture features for browsing and retrieval of image data. *IEEE Transactions on pattern analysis and machine intelligence*, 18(8):837–842, 1996.

[48] Cynthia Matuszek. Grounded language learning: Where robotics and nlp meet. *Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence (IJCAI-18)*, 2018.

[49] Cynthia Matuszek, Nicholas FitzGerald, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. A Joint Model of Language and Perception for Grounded Attribute Learning. In *Proc. of the 2012 International Conference on Machine Learning*, Edinburgh, Scotland, June 2012.

[50] Cynthia Matuszek, Evan Herbst, Luke Zettlemoyer, and Dieter Fox. Learning to parse natural language commands to a robot control system. In *Experimental Robotics*, pages 403–415. Springer, 2013.

[51] Nikolaos Mavridis and Deb Roy. Grounded situation models for robots: Where words and percepts meet. In *Intelligent Robots and Systems, 2006 IEEE/RSJ International Conference on*, pages 4690–4697. IEEE, 2006.

[52] Lilyana Mihalkova and Raymond J Mooney. Transfer learning by mapping with minimal target data. In *Proceedings of the AAAI-08 workshop on transfer learning for complex tasks*, 2008.

[53] Lilyana Mihalkova, Tuyen Huynh, and Raymond J Mooney. Mapping and revising markov logic networks for transfer learning. In *AAAI*, volume 7, pages 608–614, 2007.

[54] Krystian Mikolajczyk and Cordelia Schmid. Scale & affine invariant interest point detectors. *International journal of computer vision*, 60(1):63–86, 2004.

[55] Shantanu Misale and Abid Mulla. Learning visual words for content based image retrieval. In *2018 2nd International Conference on Inventive Systems and Control (ICISC)*. IEEE, 2018.

[56] Dipendra Misra, John Langford, and Yoav Artzi. Mapping instructions and visual observations to actions with reinforcement learning. *arXiv preprint arXiv:1704.08795*, 2017.

[57] Eric Nowak, Frédéric Jurie, and Bill Triggs. Sampling strategies for bag-of-features image classification. In *European conference on computer vision*, pages 490–503. Springer, 2006.

[58] Stephen O'Hara and Bruce A Draper. Introduction to the bag of features paradigm for image classification and retrieval. *arXiv preprint arXiv:1101.3354*, 2011.

[59] Alexander G Ororbia, Ankur Mali, Matthew A Kelly, and David Reitter. Visually grounded, situated learning in neural models. *arXiv preprint arXiv:1805.11546*, 2018.

[60] Sinno Jialin Pan, Qiang Yang, et al. A survey on transfer learning. *IEEE Transactions on knowledge and data engineering*, 22(10):1345–1359, 2010.

[61] PM Panchal, SR Panchal, and SK Shah. A comparison of sift and surf. *International Journal of Innovative Research in Computer and Communication Engineering*, 1(2):323–327, 2013.

[62] Nisha Pillai and Cynthia Matuszek. Unsupervised selection of negative examples for grounded language learning. In *Proc. of the Thirty-second AAAI Conference on Artificial Intelligence (AAAI)*, New Orleans, Louisiana, USA, February 2018.

[63] J. Ross Quinlan. Simplifying decision trees. *International journal of man-machine studies*, 27(3):221–234, 1987.

[64] Irina Rish et al. An empirical study of the naive bayes classifier. In *IJCAI 2001 workshop on empirical methods in artificial intelligence*, volume 3, pages 41–46. IBM New York, 2001.

[65] Ahmed Rubaai, Daniel Ricketts, and M David Kankam. Development and implementation of an adaptive fuzzy-neural-network controller for brushless drives. In *Industry Applications Conference, 2000. Conference Record of the 2000 IEEE*, volume 2, pages 947–953. IEEE, 2000.

[66] Amna Sarwar, Zahid Mehmood, Tanzila Saba, Khurram Ashfaq Qazi, Ahmed Adnan, and Habibullah Jamal. A novel method for content-based image retrieval to improve the effectiveness of the bag-of-words model using a support vector machine. *Journal of Information Science*, page 0165551518782825, 2018.

[67] Bernt Schiele and James L Crowley. Recognition without correspondence using multidimensional receptive field histograms. *International Journal of Computer Vision*, 36(1):31–50, 2000.

[68] Anton Schwaighofer, Volker Tresp, and Kai Yu. Learning gaussian process kernels via hierarchical bayes. In *Advances in neural information processing systems*, pages 1209–1216, 2005.

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

[70] Michael Spranger and Luc Steels. Co-acquisition of syntax and semantics-an investigation in spatial language. 2015.

[71] Luc Steels. Evolving grounded communication for robots. *Trends in cognitive sciences*, 7(7):308–312, 2003.

[72] Stefanie Tellex, Thomas Kollar, Steven Dickerson, Matthew R Walter, Ashis Gopal Banerjee, Seth J Teller, and Nicholas Roy. Understanding natural language commands for robotic navigation and mobile manipulation. In *AAAI*, volume 1, page 2, 2011.

[73] Simon Thiel, Dagmar Häbe, and Micha Block. Co-operative robot teams in a hospital environment. In *Intelligent Computing and Intelligent Systems, 2009. ICIS 2009. IEEE International Conference on*, volume 2, pages 843–847. IEEE, 2009.

[74] Suyog Trivedi, Rajesh Kumar, Gopichand Agnihotram, and Pandurang Naik. Unsupervised feature learning using deep learning approaches and applying on the image matching context. In *2017 3rd International Conference on Applied and Theoretical Computing and Communication Technology (iCATccT)*, pages 220–225. IEEE, 2017.

[75] Devis Tuia, Claudio Persello, and Lorenzo Bruzzone. Domain adaptation for the classification of remote sensing data: An overview of recent advances. *IEEE geoscience and remote sensing magazine*, 4(2):41–57, 2016.

[76] Michael Van den Bergh, Daniel Carton, Roderick De Nijs, Nikos Mitsou, Christian Landsiedel, Kolja Kuehnlenz, Dirk Wollherr, Luc Van Gool, and Martin Buss. Real-time 3d hand gesture interaction with a robot for understanding directions from humans. In *RO-MAN, 2011 IEEE*, pages 357–362. IEEE, 2011.

[77] Chang Wang and Sridhar Mahadevan. Heterogeneous domain adaptation using manifold alignment. In *IJCAI proceedings-international joint conference on artificial intelligence*, volume 22, page 1541, 2011.

[78] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. A survey of transfer learning. *Journal of Big Data*, 3(1):9, 2016.

[79] Karl Weiss, Taghi M Khoshgoftaar, and DingDing Wang. Transfer learning techniques. In *Big Data Technologies and Applications*, pages 53–99. Springer, 2016.

[80] David Williams, Xuejun Liao, Ya Xue, Lawrence Carin, and Balaji Krishnapuram. On classification with incomplete data. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, (3):427–436, 2007.

[81] Tom Williams and Matthias Scheutz. A framework for resolving open-world referential expressions in distributed heterogeneous knowledge bases. In *AAAI*, pages 3958–3965, 2016.

[82] Lu-Qi Xiao, Jun-Yun Zhang, Rui Wang, Stanley A Klein, Dennis M Levi, and Cong Yu. Complete transfer of perceptual learning across retinal locations enabled by double training. *Current Biology*, 18(24):1922–1926, 2008.

[83] Li Xu, Jimmy SJ Ren, Ce Liu, and Jiaya Jia. Deep convolutional neural network for image deconvolution. In *Advances in Neural Information Processing Systems*, pages 1790–1798, 2014.

[84] Jianchao Yang, Kai Yu, Yihong Gong, and Thomas Huang. Linear spatial pyramid matching using sparse coding for image classification. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 1794–1801. IEEE, 2009.

[85] Jufeng Yang, Jie Liang, Hui Shen, Kai Wang, Paul L Rosin, and Ming-Hsuan Yang. Dynamic match kernel with deep convolutional features for image retrieval. *IEEE Transactions on Image Processing*, 2018.

[86] Yi-Ren Yeh, Chun-Hao Huang, and Yu-Chiang Frank Wang. Heterogeneous domain adaptation and classification by exploiting the correlation subspace. *IEEE Transactions on Image Processing*, 23(5):2009–2018, 2014.

[87] Luke S Zettlemoyer and Michael Collins. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. *arXiv preprint arXiv:1207.1420*, 2012.

[88] Timothy J Zeyl and Tom Chau. A case study of linear classifiers adapted using imperfect labels derived from human event-related potentials. *Pattern Recognition Letters*, 37:54–62, 2014.

[89] Joey Tianyi Zhou, Sinno Jialin Pan, Ivor W Tsang, and Yan Yan. Hybrid heterogeneous transfer learning through deep learning. In *AAAI*, pages 2213–2220, 2014.

[90] Yin Zhu, Yuqiang Chen, Zhongqi Lu, Sinno Jialin Pan, Gui-Rong Xue, Yong Yu, and Qiang Yang. Heterogeneous transfer learning for image classification. In *AAAI*, 2011.