The SUN Action Database: Collecting and Analyzing Typical Actions for Visual Scene Types

by

Catherine Anne White Olsson

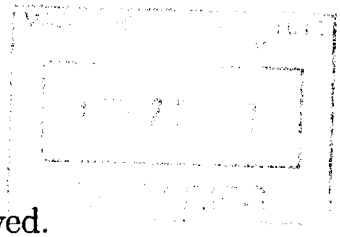
Submitted to the Department of Electrical Engineering and Computer Science
in partial fulfillment of the requirements for the degree of

Master of Engineering in Electrical Engineering and Computer Science

at the

MASSACHUSETTS INSTITUTE OF TECHNOLOGY

June 2013



Author .....
Department of Electrical Engineering and Computer Science
May 24, 2013

Certified by .....
Aude Oliva
Principal Research Scientist
Thesis Supervisor

Certified by .....
Antonio Torralba
Associate Professor
Thesis Supervisor

Accepted by .....
Prof. Dennis M. Freeman
Chairman, Masters of Engineering Thesis Committee

The SUN Action Database: Collecting and Analyzing Typical Actions for Visual Scene Types

by

Catherine Anne White Olsson

Submitted to the Department of Electrical Engineering and Computer Science
on May 24, 2013, in partial fulfillment of the
requirements for the degree of
Master of Engineering in Electrical Engineering and Computer Science

Abstract

Recent work in human and machine vision has increasingly focused on the problem of scene recognition. Scene types are largely defined by the *actions* one might typically do there: an *office* is a place someone would typically "work". I introduce the **SUN Action** database (short for "Scene UNderstanding – Action"): the first effort to collect and analyze free-response data from human subjects about the typical actions associated with different scene types. Responses were gathered on Mechanical Turk for twenty images per catgory, each depicting a characteristic view of one of 397 different scene types. The distribution of phrases is shown to be heavy-tailed and Zipf-like, whereas the distribution of semantic roots is *not* Zipf-like. Categories strongly associated with particular tasks or actions are shown to have lower overall diversity of responses. A hierarchical clustering analysis reveals a heterogeneous clustering structure, with some categories readily grouping together, and other categories remaining apart even at coarse clustering levels. Finally, two simple classifiers are introduced for predicting scene types from associated actions: a nearest centroid classifier, and an empirical maximum likelihood classifier. Both classifiers demonstrate greater than 50% classification performance in a 397-way classification task.

Thesis Supervisor: Aude Oliva
Title: Principal Research Scientist

Thesis Supervisor: Antonio Torralba
Title: Associate Professor

Acknowledgments

This work would not have been possible without the support of a great many people who have played a role in my education up to this point — advisors, mentors, teachers, colleagues, family, and friends. The following people are just a small subset of the vast number of people to whom I am extremely grateful:

- Aude, for being a constant source of enthusiasm and encouragement. I could not have asked for a more understanding, empathetic, and motivating advisor. Aude has been a role model as someone unafraid to "dream big" and to push the frontiers of knowledge at the intersection of academic disciplines. It has been a joy to work alongside one of my academic heroes, sharing in her vision and energy.

- Antonio, for inspiring me to be playful and to stay true to the engineering spirit: when the breadth of the bigger picture gets overwhelming, just try something and see what works!

- My parents, for neverending support despite a distance of thousands of miles between us. Words cannot express my gratitude for their encouragement and pride. "Lots of Love!"

- The many supportive and inspirational teachers, advisors, and mentors I have had over the years: Laura Schulz, Patrick Winston, Josh Tenenbaum, Rebecca Saxe, and Noah Goodman here at MIT, not to mention a great many teachers at Lakeside and in the PRISM program throughout my grade school experience. I am immensely indebted to the teachers throughout my life who have gone beyond simply imparting knowledge; who have taken the time to get to know me personally, and placed enormous trust in me and my abilities.

- The hundreds of workers on Mechanical Turk without whom this work would quite literally not be possible, for their endless patience, and for their delightful sense of humor which kept me afloat during many tedious hours.

- Michelle Greene, for sharing her LabelMe object analyses with me, which helped me immensely in figuring out how to wrap my head around this data and get an initial foothold.

- The Writing and Communication Center, and to anyone who has ever asked me to write anything: every essay, report, or paper I've ever written has gone into preparing me for this.

- Last but *certainly* not least, the friends and communities which imbue my life with meaning, context, stability, fulfillment, and overwhelming joy. You give me a reason to keep smiling, always.

Contents

List of Figures

List of Tables

Chapter 1

Introduction

Recent work in human and machine vision has increasingly focused on the problem of *scene recognition* - that is, answering the question "where are you?" The answer, of course, is not "in a medium-sized room containing a desk, a chair, some carpet, and a computer," but rather, "in an *office*." A scene is defined not just by its objects, materials, and spatial layout. Rather, a scene type is defined by the *actions* one might typically do there: an *office* like the one in Figure 1-1 is more than just a collection of supplies and furniture — it is a place where one would typically "work". In order to take computer vision to the next level, we must move beyond simply identifying scenes, objects, and materials, and begin endowing computers with an understanding of how people interact with the world around them — not just what things and places are, but what they are *for*.

Recent developments in computer vision have enabled computers to identify scene types from a variety of visual cues (Lazebnik *et al.*, 2006) (Torralba *et al.*, 2008) and to use scene context to improve recognition of the objects present and the actions being performed (Torralba *et al.*, 2003) (Torralba *et al.*, 2006) (Hoiem *et al.*, 2006) (Oliva & Torralba, 2007) (Marszalek *et al.*, 2009) (Li & Fei-Fei, 2007). However, while previous work has dealt with correlations between scenes and actions for restricted sets of scene categories and action types, no work to date has characterized the full space of actions one might typically do in different places over a broad range of scene types and actions. To that end, this work presents the **SUN Action** database (short for

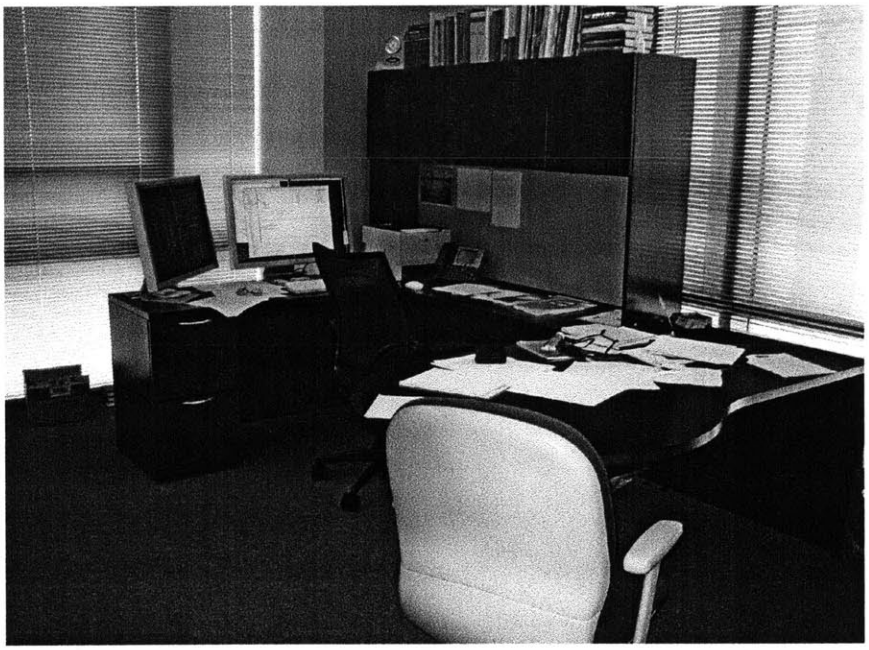


Figure 1-1: In an office, a person might typically work, type, or answer the phone

"Scene UNderstanding – Action"): the first effort to collect and analyze free-response data from human subjects about the typical actions associated with different scene types. We used Mechanical Turk to gather responses for twenty images per catgory, each depicting a characteristic view of one of 397 different scene types.

1.1 Previous work

This project consists of collecting and analyzing a crowdsourced dataset of typical actions, based on prototypical images drawn from large set of fine-grained scene categories. As such, it builds on previous work in a variety of domains across scene understanding, action recognition, and crowdsourced image annotation. In this section I briefly summarize relevant existing work the following areas:

- The general trend in vision research towards studying scene understanding at a holistic level
- Systems that learn and understand relationships between actions and scene contexts, which generally operate over small sets of scene types and action

categories

- The SUN database, a project to identify and catalog a vast breadth of visual scene types

- The scope and nature of existing datasets of actions

- Ongoing projects which, like this project, aim to annotate images in the SUN database with additional semantic information.

I then explain the contributions of this work in the context of previous efforts.

1.1.1 Perceiving and understanding visual scenes

Human and machine vision both have a long history of studying object recognition (Mundy, 2006) (Biederman, 1987), and have shown recent advances in the perception of materials (Adelson, 2001). Lately, the study of *scene perception* has been on the rise (Oliva & Torralba, 2001) (Oliva & Torralba, 2006). In human cognition, scenes and places are of enormous importance: there are brain areas that show selective activation for the processing of places (as opposed to objects, faces, and so on)(Epstein & Kanwisher, 1998), and natural scenes can be recognized and classified extremely quickly from very brief glances (Fabre-Thorpe *et al.*, 2001). On the machine side, image features have been identified that contribute to overall scene recognition (Oliva & Torralba, 2006) (Torralba *et al.*, 2008), and scene context has been shown to improve object recognition and efficiency of search. (Torralba *et al.*, 2006) (Hoiem *et al.*, 2006).

As vision research is broadening its focus and becoming increasingly concerned with the holistic recognition of scenes, and with the role of contextual cues and dependencies, work on the semantic meaning and associated utility of scenes will be increasingly important. One especially important property of a scene is the set of actions that one would typically perform there.

1.1.2 Associating scenes with actions

Existing projects in computer vision have succeeded at learning and modeling the associations between actions and scene types, both in still images (Li & Fei-Fei, 2007) (Farhadi *et al.*, 2010) and in videos (Marszalek *et al.*, 2009). These projects have demonstrated that sensitivity to contextual relations between object labels, scene labels and segmentation improve the performance of each. However, these projects differ from this work in two important ways:

1. Existing work operates over highly restricted sets of scene types and actions — eight to twelve actions, and as few as ten scenes. By contrast, this work aims to shed light on a much more exhaustive space of scene-action associations.

2. Existing work focuses on recognizing actions that people depicted are *currently doing*. By contrast, this work aims to identify *typical* or *possible* actions.

Interestingly, research on typical and possible actions is prevalent in object recognition literature, where it is known as the study of "affordance" (Stark *et al.*, 2008) (Kjellstrm *et al.*, 2011), even though the idea of typical action associations has not been explored in the realm of visual scene recognition.

1.1.3 Defining fine-grained scene categories

Where prior work has focused on a limited set of scenes, this work aims to explore the full range of scene types that people might find themselves in. To that end, we base our work on the Scene UNderstanding Database ("SUN Database") (Xiao *et al.*, 2010). The SUN database is a large-scale effort to represent as fully as possible the set of scene types in human experience (that is, to identify and catalog the set of general place types that could be used to fill in sentences such as "I am in a *place*" or "Let's go to the *place*" and to gather images representing these scene types). Where prior work on scenes (such as (Lazebnik *et al.*, 2006)) had dealt only with limited sets of categories, the SUN database comprises hundreds of thousands of images of visual scenes, drawn from an exhaustive set of 899 scene types. A 397-member subset

of these scene types has been identified as having at least 100 unique photographs in the database and are designated as the "well-sampled" categories.

Since its creation, the SUN Database has been used by researchers in computer science to implement systems for tasks such as automatic description generation (Ordonez *et al.*, 2011), and by cognitive scientists to study cognitive phenomena such as visual memory (Isola *et al.*, 2011).

As with any category structure, category membership is graded (Tversky & Hemenway, 1983) (Rosch & Lloyd, 1978). The images comprising the SUN database vary in how typical they are of their category: some images are very representative of the scene type they depict, whereas some images show unusual or anomalous views. The present work aims to characterize the most typical actions for each category, and so I focus on gathering data for the most typical images in each category. Typicality ratings are based on human judgments previously gathered in (Ehinger *et al.*, 2011).

1.1.4 Datasets of actions: people in videos

In addition to exploring a wide range of scenes, this work aims to gather a broad range of words and phrases people might use to describe typical actions. Although databases of human actions do exist (Liu *et al.*, 2009) (Marszalek *et al.*, 2009) (Schuldt *et al.*, 2004) (Blank *et al.*, 2005), they are of an entirely different nature: these datasets are collections of video data of humans performing a small set of actions against a simple background, rather than the broad range of responses we are looking for.

1.1.5 Crowdsourcing attribute information for scenes with Mechanical Turk

The best way to explore the distribution of responses to the question "What might a person typically do in this place?" is simply to ask many subjects, and the easiest way to gather a vast number of short responses is by crowdsourcing, using a platform such as Amazon Mechanical Turk. This work builds on several previous projects which similarly gather attribute information for scenes in the SUN database by crowdsourc-

ing. The most similar crowdsourced attribute-gathering project is the SUN attribute project(Patterson & Hays, 2012), which sought to label SUN images with dozens of attributes, including a handful of action-related actions such as "sailing" and "camping."

Another crowdsourcing project, which deals with objects rather than actions but can be considered a kindred spirit to the SUN Action project, is the LabelMe object project (Russell *et al.*, 2008). Through the LabelMe project, much of the SUN Database has been annotated with labeled polygons outlining objects. Just as the LabelMe project gathers image-object pairings, the SUN Action project gathers image-*action* pairings. The contextual cues provided by LabelMe have been used to improve algorithms that exploit context for object recognition (Choi *et al.*, 2010) (Oliva & Torralba, 2007); similarly, I hope that the contextual cues derived from the SUN Action database can be used to improve algorithms that exploit context for action recognition (such as Marszalek et al. (Marszalek *et al.*, 2009)). Additionally, the descriptive analyses that have been performed by Greene on the object statistics in the LabelMe dataset (Greene, 2013) have inspired the descriptive analyses described in this work.

However, unlike LabelMe which gathers polygons, and also unlike the SUN attribute database which gathers attribute labels from a discrete set, the SUN Action database contains exclusively free-form text. In that respect, this project has perhaps more in common with the methods of Fei-Fei et al. in gathering free-form text responses for rapid visual understanding (Fei-Fei *et al.*, 2007), although the responses in that dataset were scored by independent coders, whereas no such scoring has been conducted on this data.

1.2 Questions

Once a dataset of human responses about typical actions for visual scenes had been gathered, there were many questions that I could have chosen to investigate based on that a dataset. This work focuses on the following questions:

- **What is the distribution of responses?** When asked to describe the "most

typical" actions for a given scene, do people use only a small collection of phrases, or do they generate a large diversity of responses?

- **How do different categories' response distributions differ?** Do some categories give more sterotyped responses, and others more diverse responses?

- **How do scene types cluster by similar actions?** Which scene types tend to cluster together, and which scene types are distinct from others and tend to remain apart?

- **How well can we predict a category label from a list of typical actions?** Which scene types are unambiguously determined by their associated actions, and which scene types are easily confusable for others? Do simple methods yield good classification performance?

The data gathered so far was generated in response to *typical* images in each category, and therefore lends itself better to between-category comparisons, rather than comparisons between individual images within categories. However, future work should investigate how similarities and differences in the underlying images are reflected in the properties of their responses. For this purpose, more data may need to be gathered showing a broader diversity of responses within a category, including unusual and atypical examples.

1.3 Structure and contributions of this work

In this project I gathered and analyzed information on typical actions for the 397 visual scene types in the SUN database. In contrast with prior approaches to studying scene-action associations, this work focuses not on visible actions depicted in a scene, but rather on the set of *typical* actions for a given place. Furthermore, this data set is not constrained to a narrow set of scene and action classes, but instead explores the breadth of unconstrained natural-language responses for a comprehensive set of scene types.

In Chapter 2, I explain my methodology for building the SUN Action dataset. Over 100,000 annotations for 397 visual scene types, gathered using Mechanical Turk, comprise a dataset of typical actions for scene types.

In Chapter 3, I analyze the distribution and diversity of responses, overall and by category. I show the distribution of phrases is to be heavy-tailed and Zipf-like, whereas the distribution of semantic roots is *not* Zipf-like. I identify and explain between-category differences in response distribution and diversity.

In Chapter 4, I examine action similarity relationships using hierarchical clustering. Scene types show a heterogeneous pattern of clustering: some scene types cluster readily together even in fine-grained clusterings, whereas other scene types resist clustering even at the coarsest level of grouping.

In Chapter 5, I describe two classifiers for predicting scene types from associated actions: a nearest centroid classifier, and an empirical maximum likelihood classifier. Both classifiers demonstrate greater than 50% classification performance in a 397-way classification task.

In Chapter 6, I explain directions for future work, including a potential follow-up study on the effect of category name on subjects' responses, and possible next steps for correlating objects, materials, and spatial properties with typical actions.

Chapter 2

Building a Dataset of Typical Actions for Scene Types

The first step in characterizing human judgments of typical actions is to gather a dataset of such judgments. Data was gathered using Amazon Mechanical Turk, an online crowdsourcing platform through which "Requesters" can post small tasks (known as Human Intelligence Tasks, or HITs) for anonymous "Workers" to complete. We chose to use Mechanical Turk because it provides a simple way to gather data quickly and conveniently at low cost. The task of labeling images with descriptive words is familiar to workers and well-suited to the format of multiple small tasks.

Although the use of Mechanical Turk for gathering experimental data has a short history, Mechanical Turk workers have been shown to produce comparable results to laboratory subjects on a variety of measures (Paolacci *et al.*, 2010) (Horton *et al.*, 2011). Many studies similar to this one have used Mechanical Turk with great success (see (Ordonez *et al.*, 2011), (Isola *et al.*, 2011), (Ehinger *et al.*, 2011), (Farhadi *et al.*, 2010), and (Choi *et al.*, 2010)). Working with Mechanical Turk has advantages over working with in-person laboratory subjects, including cost and convenience. For more information on the use of Mechanical Turk for data collection, see Appendix B.

2.1 Stimuli: 397 scene types

Data was gathered for all 397 "well-sampled" scene types in the SUN database. The full list of 397 scene types is enumerated in Appendix A.

In order to gather data most reflective of typical actions for each category, we chose to use images that were prototypical for their category rather than diverse and unusual views. For each scene type, we used the twenty (20) "most typical" images in each category, based on typicality ratings gathered from Mechanical Turk users in a previous study (Ehinger *et al.*, 2011). In that study, workers had repeatedly selected three images out of a set of twenty that *best* or *worst* illustrated the scene category. For each image, a typicality score was calculated from the number of times it was selected as the best exemplar of its category, minus the a fraction of the number of times it was selected as the worst exemplar, normalized by the total number of times that image had appeared. We chose the twenty images in each category with the highest typicality score my this measure.

2.2 Interface: Mechanical Turk

In Experiment 1, users were shown 7 or 8 images per HIT (initial batches accidentally contained only 7 images per HIT due to a bug), and asked to list one to three (1-3) words or short phrases describing typical actions or activities a person might do in that place. An example screenshot of the interface is shown in Figure 2-1.

Subjects were *not* told the name of the scene type they were looking at, because we wanted to examine the actions that were associated with the place itself, unaffected by the language used in the SUN dataset to name its category. However, in future studies such as those described in Chapter 6, we intend to to investigate the effect of telling subjects the name of each scene.

In a given batch, each of the $397 * 20 = 7940$ images was split up pseudorandomly into HIT groups of eight images (with four images appearing twice to round out a full batch of $(7940 + 4)/8 = 993$ HITs per batch). No two images of the same scene

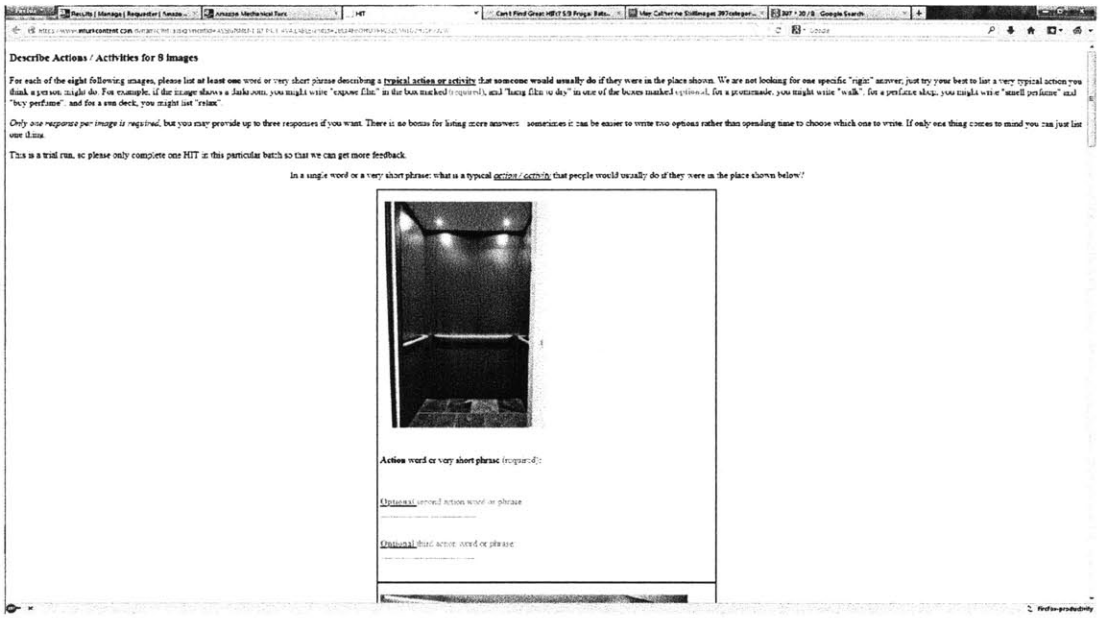


Figure 2-1: The user interface as seen by Mechanical Turk workers. Workers saw a block of instruction text, followed by eight images per page, the majority of which cannot be seen in this screenshot because they are scrolled off the page.

type appeared in the same HIT. Sixteen (16) batches were run. To minimize showing a given worker the same image twice, workers were asked not to complete HITs from multiple batches.

Despite the aforementioned bug that caused 1/8 of the images to be accidentally excluded from early batches, every image appeared at least 8 times over the course of the study. The most frequent image appeared 18 times, with images appearing on average 14.4 times.

The specific instructions shown to workers were as follows:

> **Describe Actions / Activities for 8 images**
>
> For each of the **eight** following images, please list **at least one** word or very short phrase describing a **typical action or activity** that **someone would usually do** if they were at the place shown. We are not looking for one specific "right" answer; just try your best to list a very typical action you think a person might do. For example, if the image shows a darkroom, you might write "expose film" in the box marked (required), and "hang

film to dry" in one of the boxes marked optional; for a promenade, you might write "walk"; for the inside of a perfume shop, you might write "smell perfume" and "buy perfume"; and for a sun deck, you might list "relax".

Only one response per image is required, but you may provide up to three responses if you want. There is no bonus for listing more answers - sometimes it can be easier to write two options rather than spending time to choose which one to write. If only one thing comes to mind you can just list one thing.

This is **Batch 16**. You may only work on HITs from one batch number. If you have done any of our other batches in the past few days, please do not work on this batch. If you do this batch now, please do not do any of our other batches in the next few days. Within this batch you may do as many HITs as you want.

In a single word or a very short phrase: what is a typical action / activity that people would usually do if they were in the place shown below?

2.3 Participants

Participants were US-based Mechanical Turk workers, with at least 100 HITs approved and at least 95% HIT acceptance rate. 621 unique workers participated in the experiment, each completing on average 25.6 HITs per worker (993 HITs per batch * 16 batches / 621 workers). Workers were paid $0.10 per HIT.

No data was excluded from the final dataset, nor was any data cleaning or spell-checking performed. Visual inspection suggests that spelling errors were present but rare, and that virtually all answers provided by workers ere serious attempts to respond to the prompt.

Users were allowed to provide one to three responses per image, and often chose to provide more than one response, resulting in an average of 17.3 responses per image.

Table 2.1: A canonical 33-category subset, as described in Section 2.4, for use in later analyses when viewing data for all 397 categories would be impractical

abbey	bathroom	forest	river
airport_terminal	beach	garage	staircase
alley	bedroom	greenhouse	street
art_studio	bowling_alley	highway	subway_interior
attic	casino	kitchen	waterfall
auditorium	stle	lake	wine_cellar
bakery	church	living_room	
bar	corridor	mountain_snowy	
basement	dining_room	office	

Overall, the study yielded 137,558 responses (words and short phrases describing actions).

2.4 A 33-category subset is easier to visualize

Visualizing results across 397 individual classes can be difficult; many of the analyses in the following sections are better viewed for a smaller selection of scene types. Therefore, a 33-category subset of the data was selected with the goal of capturing a selection of well-recognized and semantically important categories, using the following method.

For each of the 397 categories, the number of images with LabelMe annotations was tabulated. The 80 scene categories with the most annotated images were determined. Four of the author's collaborators were each asked to nominate the 20-40 most important scene types out of these 80. Out of the 80 scene types, 35 of them received at least two nominations and were slated for inclusion. "Cathedral" was removed for not being different enough from "church" and "abbey", and "coast" was removed for containing only aerial shots, leaving a final slate of 33 image, as listed in Table 2.1.

Chapter 3

Describing the Distribution of Responses

The following descriptive statistics attempt to characterize the distribution of subjects' responses, both global trends and category-to-category variation. These analyses aim to shed light on the following questions:

- **What is the distribution of responses?** When asked to describe the "most typical" actions for a given scene, do people use only a small collection of phrases, or do they generate a large diversity of responses?
- **How do different categories' response distributions differ?** Do some categories give more sterotyped responses, and others more diverse responses?

Visualizing data from several hundred categories can be difficult, so many of the following analyses are shown not only for the full 397 categories, but also for the 33-category subset described in Section 2.4.

This analysis examines how responses vary between one category and another, grouping together responses over all subjects and all images within each category. We did not tell subjects which scene type they were looking at, so by grouping at the category level, we can investigate the differences that naturally distinguish different scene types from one another. By showing subjects the most typical images for each

category, we hoped to elicit these category-level effects as clearly as possible. However, future work will likely investigate image-by-image differences within this dataset, and may gather additional data for more diverse and unusual views of the different scene types to better shed light on image-by-image differences as opposed to category-wide commonalities.

3.1 Phrase statistics

All told, 137,558 total responses (words and short phrases) were collected. This number represents 30,056 distinct responses (22% of the total count).

3.1.1 Phrase occurrence distribution is Zipf-like and heavy-tailed

The first question I sought to answer was what the overall distribution of responses was: do subjects provide the same answers over and over again? Is there a heavy tail? Both properties seem to be true. The most frequent responses accounted for a large proportion of the total responses. Seven responses each accounted for greater than 1% of the total responses: namely, "walk", "relax", "work", "swim", "eat", "sleep", and "shop" (listed in 3.1). Furthermore, the 350 most common phrases (1.16% of the total number of distinct phrases) account for 50% of the total responses. At the same time, the majority of responses were quite rare: among the least frequent responses, 71.0% of the responses occurred only once in the dataset, and 91.60% of the responses occurred five or fewer times.

The relationship between rank and frequency (that is, how common the most common words are, and how rare the rarest words are) can be examined by plotting rank and frequency on a log-log graph. If phrase frequency is inversely proportional to phrase rank, then the relationship will appear linear on a log-log graph, and phrase frequency can be said to obey "Zipf's Law" (Zipf, 1935). Indeed, phrase frequency is roughly Zipf-like, as shown in Figure 3-1.

Table 3.1: The seven most common phrases, each of which accounts for more than 1% of the total phrases collected.

Seven most common phrases overall	
Common phrase	Percent of all phrases
walk	1.86%
relax	1.74%
work	1.44%
swim	1.32%
eat	1.25%
sleep	1.23%
shop	1.08%

One corrollary of phrase frequency having a Zipf-like distribution is that it is more *heavy-tailed* than one might expect. That is to say, even though we asked subjects to provide the most common/typical actions, many of the responses subjects provided are rare. For example, consider the response distribution shown in Figure 3-2, in which the "beach" category has been selected as a representative example of the overall diversity. The responses include very common responses lie "swim", "walk", and "tan", but also a long tail of very uncommon responses like "paddle outrigger canoe" and "ride boogie board." The long tail of this distribution implies that the space of typical actions for a scene cannot be adequately characterized by a small collection of phrases — the full distribution will only be accurately reflected if a large number of responses is gathered.

3.1.2 Response diversity varies by category

The distribution of responses varies greatly by category. In terms of response diversity, some categories yield the same responses over and over again, while some categories yield a vast diversity of responses. To visualize this discrepancy in response diversity, the ratio of *distinct responses* to *total responses* is plotted for each category. A ratio is plotted, rather than a total number, to minimize any effect of respondents choosing to provide more total phrases for some categories, since respondents were free to provide up to three phrases per image. The phrase diversity ratio described here is also easier to compare with the stem diversity ratio described in Section 4.1.

Figure 3-3a plots the phrase diversity ratio over all 397 categories, with only a subset of the category labels shown. Figure 3-4 illustrates the phrase diversity ratio for the 33-category subset described in Section 2.4 to illustrate how a few well-known categories behave.

Response diversity varies greatly from category to category. The least diverse category in the entire dataset is *shower*, with 51 distinct phrases out of 337 total phrases (15%). The most diverse category is *veterinarian's office*, with 253 distinct phrases out of 335 total phrases (76%). The full set of phrases for *shower* and *veterinarian's office* is listed in Appendix A. The phrases for *shower* traverse only a limited set of concepts around cleaning onesself (with a few additions like singing, cleaning the shower, and so on), whereas the phrases for "veterinarian's office" explore a vast array of different pet-related and medical procedures, including actions a pet owner could do as well as actions a vet could do.

What drives the differences in diversity between categories? Among the categories with lowest diversity, one key factor seems to be that they generally have just one main associated action – bedrooms are for sleeping, bowling alleys are for bowling, and so on. Indeed, for many of the least diverse scenes (*bedroom*, *bowling alley*, *casino*, and *art studio*), over 45% of the total responses are dominated by the single most common response ("sleep", "bowl", "gamble", and "paint"). And for three of the most diverse scenes (*attic*, *corridor*, and *wine cellar*), less than 10% of the total responses are made up of the most common response. However, dominance of the primary response is not the only driving factor in the overall diversity: scenes like *bathroom* have a very small diversity of responses despite not having one dominant response, and scenes like *alley* have a large overall diversity of responses despite being dominated by the primary response "walk". Table 3.6 shows the single most common phrase for each of the five least phrase-diverse categories, and Table 3.7 shows the most common phrase for the five most phrase-diverse categories.

One property which seems to covary with response diversity is familiarity. In general, low-diversity categories tend to be places that people are familiar with and have a lot of experience with, such as a *shower* or a *bedroom*, and high-diversity

categories tend to be places that most people would be unfamiliar with, like a *server room* or *mansion*. However, this pattern does not hold in every case: some low-diversity categories are places that people might not usually find themselves in, such as a *wave*, that nonetheless have specific strongly-associated actions (such as "surfing"); and some high-diversity categories are places that people would often go to but do not have a single strongly-associated action, such as a *basement* which is defined by its location rather than its intended purpose.

Table 3.2: For the five least phrase-diverse categories, the most common associated phrase is shown. Categories with low phrase diversity tend to have a single highly-associated action word. While low phrase-diversity categories do tend to include categories that people probably have a lot of experience with, such as a *shower*, they also include categories that have a specific known purpose but that people might not usually find themselves in, such as a *wave*.

Least diverse categories: phrases	
Category	Most common phrase
shower	"shower"
wave	"surf"
swimming_pool/indoor	"swim"
gymnasium/indoor	"exercise"
parking_garage/indoor	"park"

Table 3.3: For the five most phrase-diverse categories, the most common associated phrase is shown. Categories with high phrase diversity tend not to have any one strongly-associated action phrase. High phrase-diversity categories also tend to be places that most people would be unfamiliar with.

Most diverse categories: phrases	
Category	Most common phrase
veterinarians_office	"examine"
booth/indoor	"get information"
lock_chamber	"work"
server_room	"work"
ice_shelf	"take pictures"

3.2 Morphological stem statistics

Although the distribution of *phrases* is Zipf-like, the question remains whether the inverse relationship between rank and frequency also holds true of the *semantic* prop-

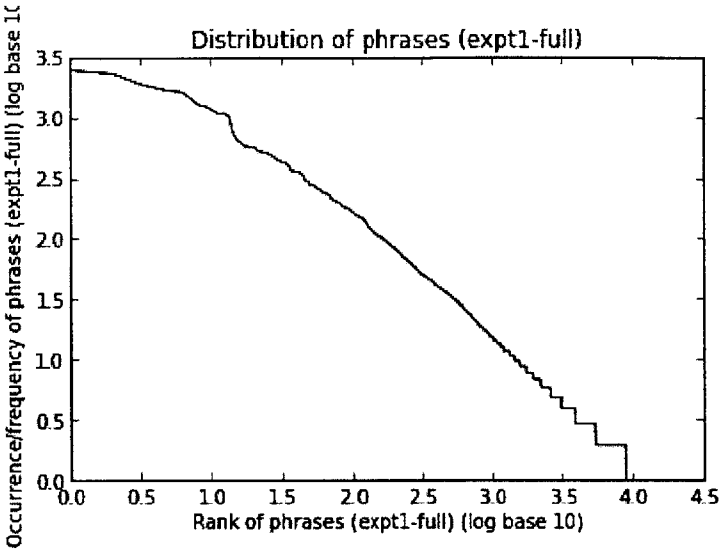


Figure 3-1: Phrase frequency is roughly inversely proportional to phrase rank, following a Zipf-like distribution.

erties of subjects' responses. A Zipf-like distribution of phrases is theoretically compatible with a set of responses that actually refer to only a small set of conceptual actions, but which use a vast spread of phrase structures to describe them.

A cursory glance at the data reveals that responses are in fact somewhat semantically redundant. Even ignoring synonym relationships, the same phrases are expressed in multiple grammatical configurations, and the same words appear in multiple phrases. Consider the class *greenhouse* for an illustration of these phenomena. Table 3.4 shows the top ten responses for *greenhouse*, which use the same root words (such as "garden", "plant", and "grow") over and over again in different configurations and grammatical forms. The following sections explain the use of "stemming" to reduce words to their grammatical roots, as a way to reduce surface variation and better elucidate the underlying concepts in users' responses.

3.2.1 Semantic content can be extracted by reducing constituent words to morphological stems

One approach for unifying information across phrases that differ in structure but not in semantic content is to break each phrase down into tokens (e.g. words), and strip

Table 3.4: The ten most common phrases for 'greenhouse indoor'. The phrases are highly redundant, with the same words appearing in different grammatical constructions (e.g. "garden" and "gardening"), and the same concepts being expressed in multiple different actions (e.g. "plants" in "water plants", "grow plants", and "buy plants".)

Ten most common phrases for 'greenhouse indoor'		
Phrase	Occurrence (out of 356)	Percent of phrases
'water plants'	40	11.2%
'grow plants'	20	5.6%
'smell flowers'	14	3.9%
'grow'	12	3.4%
'gardening'	12	3.4%
'buy plants'	12	3.4%
'plant'	11	3.1%
'water the plants'	10	2.8%
'garden'	10	2.8%
'smell the flowers'	9	2.5%

any grammatical affixes away from each token to bring it down into its morphological stem (e.g. root). These processes, known as "tokenizing" and "stemming", can be performed using standard natural language processing tools. I used the tokenization functionality in the python module `nltk.word_tokenize`, and the Porter Stemmer (Porter, 1980) as implemented in `nltk.stem.porter` (Loper & Bird, 2002). It is worth noting that the stem of a word represents its grammatical base form, stripped of all tense, inflections, and derivational morphology, and is not necessarily a valid English word - for example, the shared stem of "gamble", "gambling", and "gambler" is the non-word "gambl". It is also worth noting that by tokenizing and stemming the entirety of each response (such as "grow plants"), more semantic content is preserved and captured than if we were to stem only the initial verb ("grow"). Additional context is carried in the direct and indirect objects, adjectives, adverbs, and prepositions, that make up action phrases alongside the primary verbs.

As a result of the transformation to stems, responses in each category are condensed down to a bag-of-words representation – or more accurately, a bag-of-stems representation. In this representation, the information about the order and co-occurrence of stems within phrases is discarded. Nonetheless, bag-of-words models have proven useful in domains ranging from natural language processing (Blei *et al.*,

2003) to computer vision (Lazebnik *et al.*, 2006).

3.2.2 Responses were short

The first thing we can verify from the tokenized responses is that subjects followed directions. Specifically, we asked subjects for single words or short phrases, not long responses. Subjects did in fact follow directions: the average response length was 1.98 tokens per response, indicating that participants generally provided short answers.

3.2.3 Morphological stem distribution is *not* Zipf-like

Returning to the original question about response distribution, we can now investigate whether the inverse rank-frequency relationship still holds after the transformation to stems. In this space, we are getting closer to examining the distribution of *what* people talk about, not just *how* they say it. However, in order to examine the underlying semantic meaning further than the stem level, one would need to examine synonym and relatedness relationships, which is beyond the scope of this work.

Responses were tokenized, uninformative words were removed ("stop words", see Appendix A for a list), and the resulting tokens were reduced to their morphological stems. As a result of this process, a total of 223,062 total stems were extracted. This total is composed of only 5,547 distinct stems (2% of the total count). Among the most frequent responses, fifteen responses each accounted for greater than 1% of the total responses: namely, "play", "walk", "take", "go", "watch", "ride", "buy", "eat", "work", "relax", "shop", "swim", "pictur", and "look" (as listed in 3.5). Among the least frequent stems, 40.80% of the stems occurred only once in the dataset, a much smaller percentage than the percent of phrases that occurred only once. Additionally, 67.44% of the responses occurred five or fewer times, which is also much smaller than the corresponding percent of phrases. That said, the 67 most common stems (1.21% of all distinct stems) account for 50% of the responses, which is a similar percentage to the observed proportion of phrases accounting for 50% of occurrences.

As with phrases, the relationship between stem rank and frequency can be visual-

Table 3.5: The fifteen most common stems, each of which accounts for more than 1% of the total stems collected.

Fifteen most common stems overall	
Common stem	Percent of all stems
play	2.84%
walk	2.56%
take	2.50%
go	1.85%
watch	1.77%
ride	1.61%
buy	1.60%
eat	1.59 %
work	1.52%
relax	1.42%
shop	1.13%
swim	1.09%
sit	1.09%
pictur	1.08 %
look	1.02 %

ized by plotting it on a log-log graph. Unlike phrases, stem frequency is *not* inversely proportional to stem rank - that is to say, the distribution of stems does not follow Zipf's law. On the contrary, there are many more "mid-frequency" stems than such a law would predict, as demonstrated here by the fact that the graph in Figure 3-5 is distorted upwards from a straight line. This distortion indicates that the frequency of later-rank stems is elevated, as compared to a Zipf-like distribution.

Although the distribution of morphological stem frequencies is not Zipf-like, the distribution is nonetheless somewhat heavy-tailed, containing a substantial number of rare or unique stems. The response distribution for *beach* is shown in Figure 3-6. The responses which had been very common in the phrase chart (such as "swim" and "relax") are still at the top of the order; words such as "walk" have increased strongly in frequency compared to their position in the phrase chart due to the contribution of forms such as "walk along the shore"; and there are many more mid-frequency stems, including nouns such as "beach" and "sun". That said, the heavy tail remains, with stems like "barbequ" and "explor" remaining relatively rare.

3.2.4 Morphological diversity varies by category

The distribution of stems varies by category, although the variation is not as dramatic as the variation in phrases. Some categories contain references to only a small set of root concepts, whereas other categories' responses draw on a variety of different semantic roots. The category-to-category variation in stem frequency can be seen in Figure A-1 in the Appendix, which shows a histogram of the top 10 stems for every category in the 33-category subset.

As in Section 3.1, the ratio of *distinct stems* to *total stems* is plotted for each category. Figure 3-7a plots the stem diversity ratio over all 397 categories, with only a subset of the category labels shown, and Figure 3-8 illustrates the stem diversity ratio for the 33-category subset described in 2.4.

The least diverse category in the entire dataset is *shoe shop*, with 41 distinct phrases out of 658 total stems (6.3%). The most diverse category is *slum*, with 181 distinct stems out of 524 total stems (34.5%). The full set of stems for *shoe shop* and *slum* is listed in Appendix A.

The distribution of stems for *shoe shop* is dominated almost exclusively by "shoe" and "buy," along with a few synonyms for each (like "loafer" and "purchas"). Additional stems refer to other things the customer might do as part of the purchasing process, like browsing and selecting; as well as actions the staff could do, like selling, measuring, and organizing. By contrast, the stems for "slum" traverse a vast array of concepts. The most common stem is "live", and some of the other stems describe dwellings, but most of the further words describe poverty, filth, disease, and subsistence, along with more positive concepts like helping, fixing, and cleaning. Some of the rarer concepts describe streets and roads, villages and neighborhoods, and children and games.

In general, the least stem-diverse categories are those which are designated for a specific purpose (such as a *greenhouse* for growing plants or an *auditorium* for hearing talks and lectures) whereas the most stem-diverse categories are those which have no designated purpose (such as a *castle* or an *alley*).

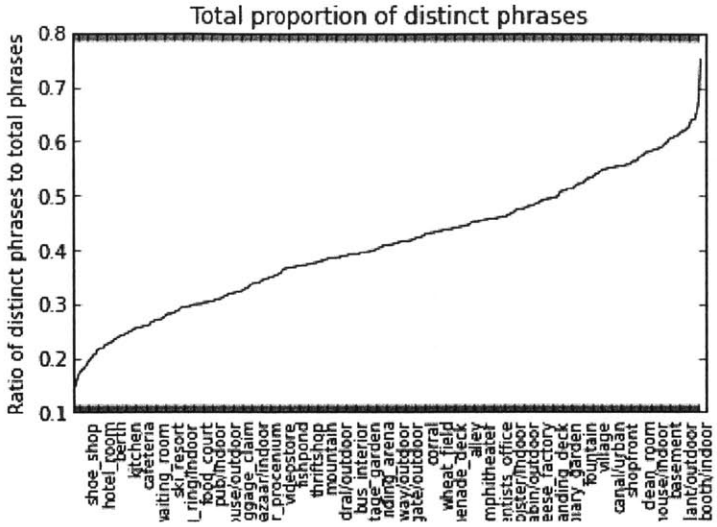
Table 3.6: For the five least stem-diverse categories, the most common associated stem is shown. Categories with low diversity tend to be highly associated with particular actions or particular object types.

Least diverse categories: stems	
Category	Most common stem
shoe_shop	"shoe"
parking_garage/indoor	"park"
shower	"shower"
laundromat	"cloth"
baggage_claim	"luggage"

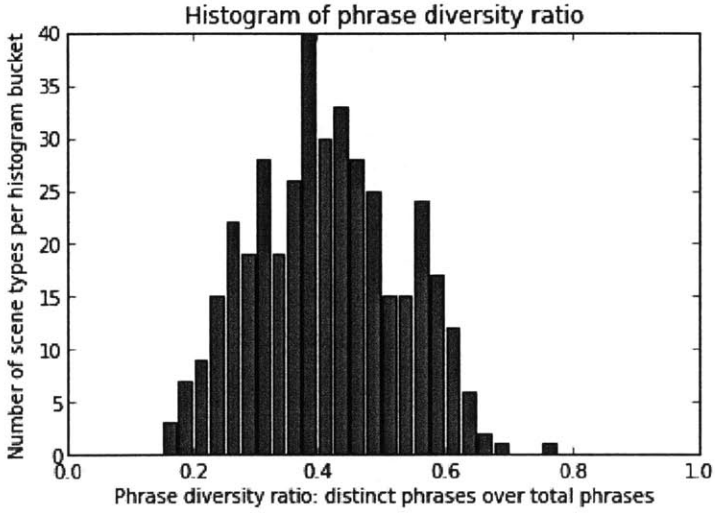
Table 3.7: For the five most stem-diverse categories, the most common associated stem is shown. Categories with high stem diversity tend not to have any one clear purpose, or associated action or object.

Most diverse categories: stems	
Category	Most common stem
slum	"live"
control_room	"work"
power_plant/outdoor	"work"
anechoic_chamber	"record"
oil_refinery/outdoor	"take pictures"

Figure 3-2: The distribution of responses is heavy-tailed, with many of the responses occurring very infrequently. For example, the responses to the *beach* category include very common responses lie "swim", "walk", and "tan", but also very uncommon responses like "paddle outrigger canoe" and "ride boogie board."



(a) Ratio of distinct phrases to total phrases per category



(b) Scene types are sorted into buckets based on their phrase diversity ratio

Figure 3-3: Ratio of number of distinct responses to number of total responses, for all 397 categories. The same information is presented in two different formats: the top plot shows a line plot of phrase diversity ratio for all 397 categories, with a subset of category labels is shown, while the bottom plot presents the same information in a histogram format to better illustrate the distribution in a more familiar "bell-curve" format. Response diversity varies greatly, from the least diverse (*shower*) with the number of distinct responses amounting to only 15% of the total responses, to the most diverse (*veterinarian's office*) with the number of distinct responses amounting to fully 76% of the total responses.

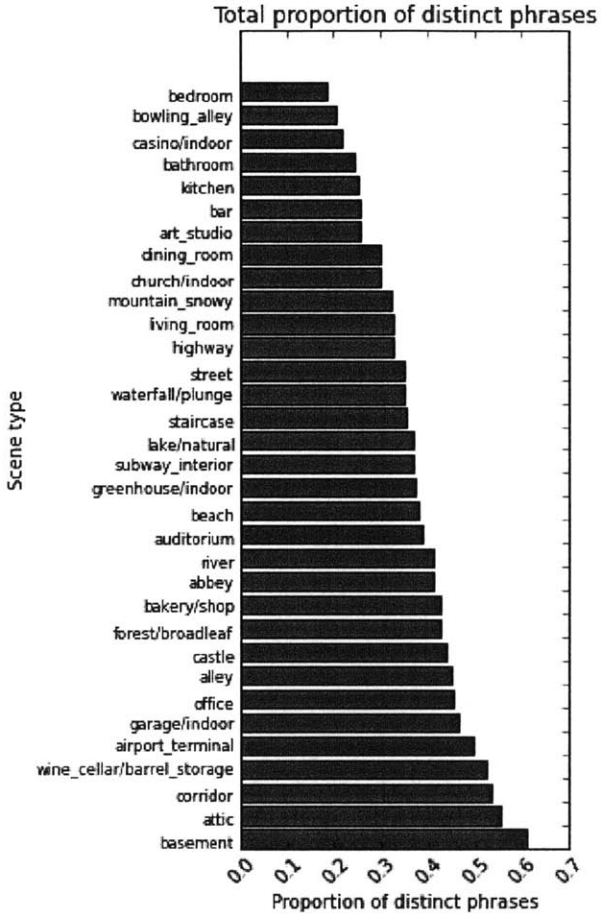


Figure 3-4: Ratio of number of distinct responses to number of total responses, for a well-known subset of categories. In general, categories that are evocative of one specific associated action word (such as "sleep" in a bedroom and "pray" in a church) produce low response diversity, whereas categories that have no single most-associated word (such has *attic* and *basement*) produce high response diversity.

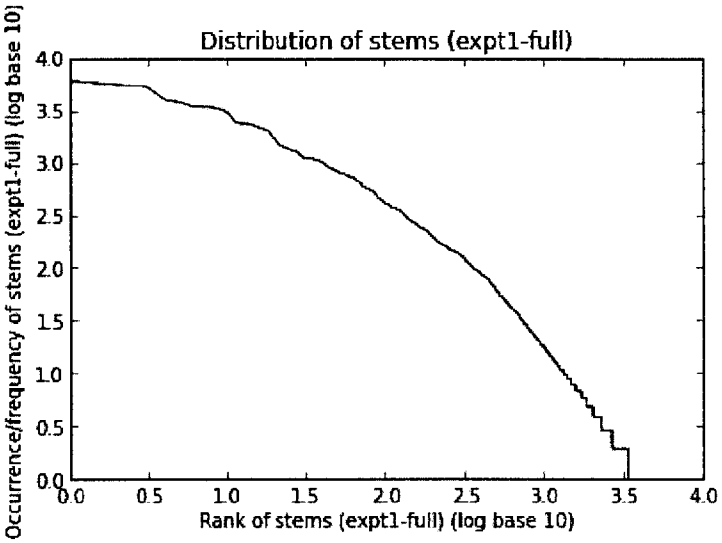


Figure 3-5: Stem frequency is *not* inversely proportional to stem rank. On the contrary, there are many more "mid-frequency" stems than such a law would predict, as demonstrated here by the fact that this graph is distorted upwards from a straight line.

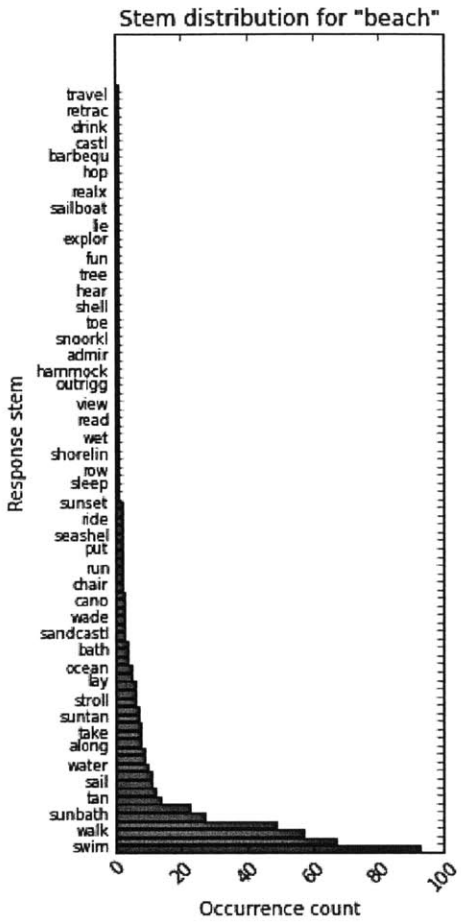
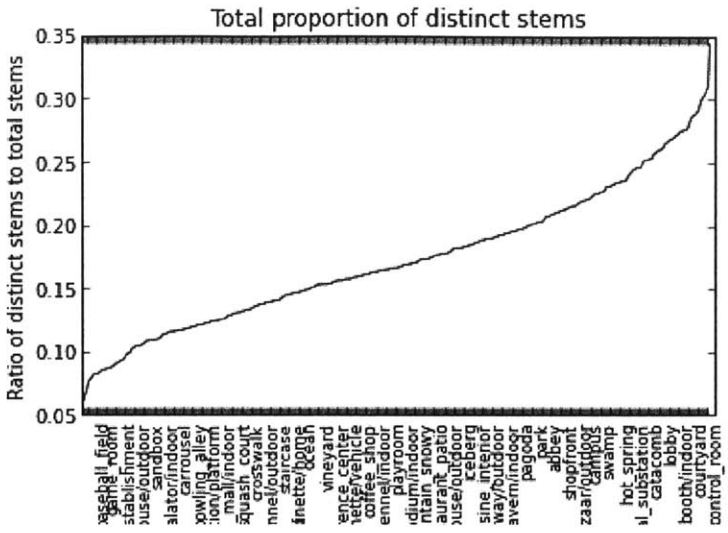
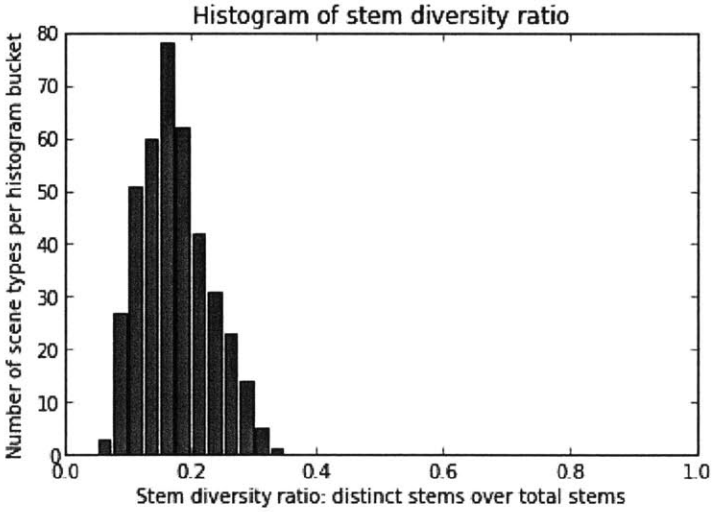


Figure 3-6: Unlike the distribution of phrases, the distribution of stems generally contains more mid-frequency items - that is to say, the drop-off is not sharp enough for the distribution to be properly Zipf-like. The responses that were most common in the phrase plot (such as "swim" and "relax") remain common; some verbs, such as "walk", have increased in frequency because phrases like "walk along the shore" have been incorporated; and words such as "beach" and "sun" appear among the mid-frequency stems.



(a) Ratio of distinct stems to total stems per category.



(b) Scene types are sorted into buckets based on their stem diversity ratio

Figure 3-7: Ratio of number of distinct stems to number of total stems, for all 397 categories. The top plot shows a line plot of phrase diversity ratio for all 397 categories, with a subset of category labels is shown, while the bottom plot presents the same information in a histogram format. Response diversity varies moderately, from the least diverse ("shoe shop") with the number of distinct stems amounting to 7% of the total number of stems, to the most diverse ("slum") with the number of distinct responses amounting to 35% of the total responses.

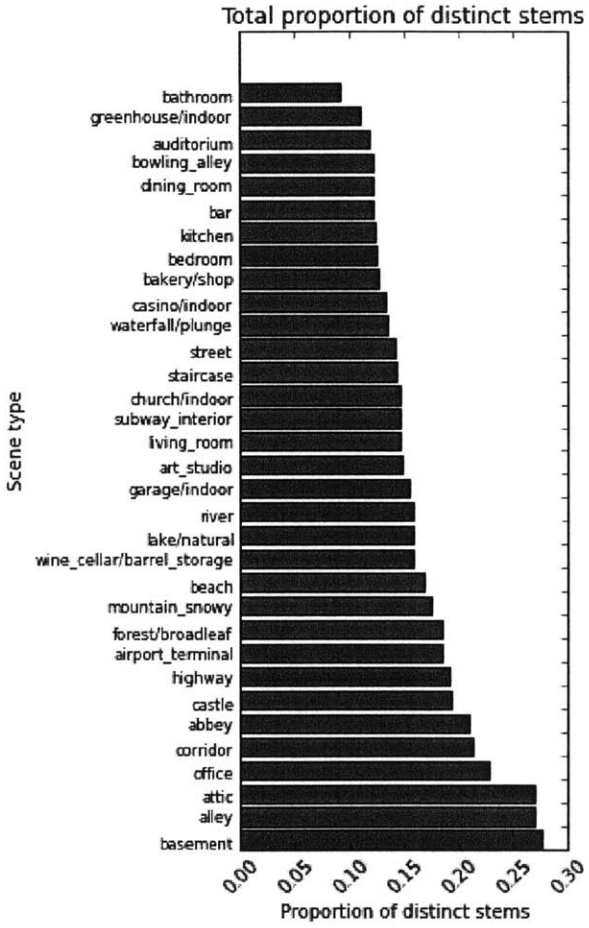


Figure 3-8: Ratio of number of distinct responses to number of total responses, for a well-known subset of categories. In general, categories that are designated for a specific purpose (such as a greenhouse for growing plants, an auditorium for hearing talks and lectures, etc.) produce low morphological diversity, whereas categories that have no specific purpose (such as a "castle" or "alley") produce high morphological diversity.

Chapter 4

Visualizing the Action-Similarity Space of Scenes

One goal of this work is to understand the similarity relationships between different scene types according to their associated actions: **How do scene types cluster by similar actions? Which scene types tend to cluster together, and which scene types are distinct from others and tend to remain apart?**

In order to understand similarity relationships, a similarity measure must be chosen. I use normalized histogram similarity, as explained in Section 4.1. This similarity measure defines a *similarity space* over scene types, which in future work could be compared to similarity spaces defined by other measures. This similarity space is presented in two views: as heatmaps (Figure 4.2), and as a hierarchical clustering tree (Figure 4-2). The heatmap enables visual detection of bright spots corresponding to pairs of scene types of strong similarity, and dark lines corresponding to scene types which are especially distant from other scene types. Hierarchical clustering reveals a heterogeneous clustering structure, with some scene types clustering together readily into large groups even at fine levels of clustering, and other scene types resisting clustering even at the coarsest scale.

4.1 Normalized histogram similarity is a symmetric distance measure for frequency counts

Having transformed the responses per category into a bag-of-stems representation as described in Section , we now need a distance measure that operates on stem frequency counts. One standard distance measure for comparing probability distributions is Kullback-Liebler divergence, also known as relative entropy (Cover & Thomas, 1991):

$$KL(p||q) = \sum_x p(x) \log \frac{p(x)}{q(x)}$$

Note that frequency-count histograms can be transformed into empirical probability distributions simply by dividing each entry by the total number of elements:

$$\frac{a[key]}{sum(a.values())} = \Pr(key|a)$$

However, K-L divergence is not symmetric, and we need a symmetric similarity measure to derive a single similarity value between any two categories. Therefore, we chose the following simple histogram similarity measure, which sums the overlap of the two distributions at each point.

$$\begin{aligned} histogramSimilarity(a,b) &= \sum_{i=1...keys} \min\left(\frac{a[i]}{\sum_{j=1...keys} a[j]}, \frac{b[i]}{\sum_{j=1...keys} b[j]}\right) \\ &= \sum_{keys} \min(\Pr(key|a), \Pr(key|b)) \end{aligned}$$

This normalized histogram similarity measure is 0 when the two frequency counts are completely disjoint, and 1 when they are perfectly identical.

4.2 Similarity heatmaps

A simple way to visualize similarity is as a heatmap: class labels go along the $x-$ and $y-$ axes, and each (x, y) pixel shows the similarity between category x and category y, with bright pixels indicating more similarity, and dark pixels indicating less similarity. Figure 4-1 shows similarity heatmaps, alphabetical by scene type.

A few bright spots are visible, indicating similar pairs, such as *attic* and *basement*; *highway* and *street*; and the trio *lake*, *river*, and *waterfall*. Dark lines can also be seen, indicating scene types that are not similar to any other scene types, such as *casino* and *kitchen*. However, because the $x-$ and $y-$ axes of these plots are organized alphabetically rather than by any particular logical grouping, it is difficult to see any more sophisticated structure in the data. Section 4-2 will describe a clustering method; as a preview, the heatmaps when grouped according to that clustering scheme show a more visible block structure with a wide variety block sizes, as shown in 4-2.

4.3 Hierarchical clustering shows heterogeneous grouping structure

One transformation that can be performed on a similarity space to better understand its structure is *hierarchical clustering*. Hierarchical clustering groups data at multiple scales, so that the absolute most similar scene types cluster together at the finest scales, with more distantly-related scene types clustering together at broader scales of clustering. The results of clustering can be viewed as a *dendrogram* in which branchings closer to the leaves of the tree represent fine-grained clusterings, and branchings closer to the trunk of the tree encompass broader clusterings. The height of each linkage in the dendrogram indicates the distance between the subclusters it connects. To compute hierarchical clusters and dendrograms in the following analyses, I used the `linkage`, `dendrogram`, and `cluster` functions available in the MATLAB Statistics Toolbox (The MathWorks, Inc, 2013).

The aim of this hierarchical clustering analysis is to understand the grouping

structure of the similarity space of actions. Do some scene types group with others more readily, and with which others? Do other scene types remain apart and distinct?

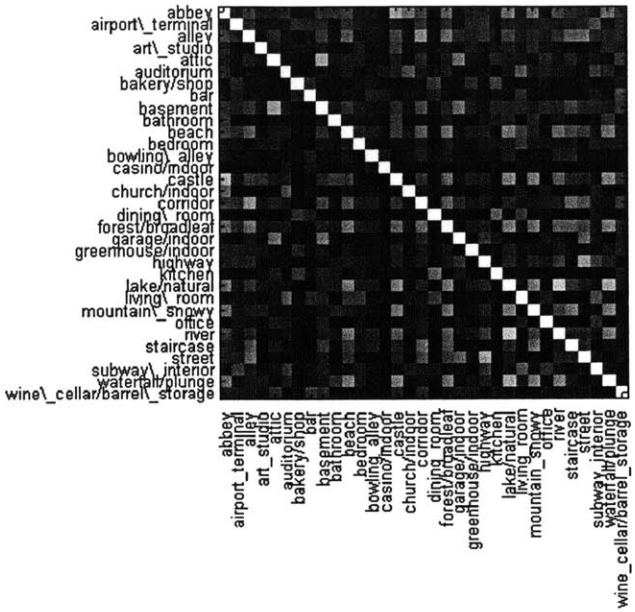
Figure 4-3 shows a dendrogram representation of the hierarchical clustering for the 33-category subset. Overall, this clustering structure is highly heterogeneous, with some scene types (generally shown at the bottom of the chart) clustering together readily into large agglomerations, and some scene types (generally at the top of the chart) being truly distinct from others and resisting clustering.

Starting from the bottom of Figure 4-3, the large cluster at the bottom combines two subclusters: outdoor scenes and scenes specifically for movement and travel. The fact that all the outdoor scenes group together before the indoor scenes begin to converge indicates that outdoor scenes are generally more similar to one another than indoor scenes are to one another. Within the outdoor scenes, The two most action-similar scene types — *river* and *lake/natural* — cluster together most readily. The historical attractions — *castle* and *abbey* — group together next. The group that unifies *street* and *highway*, along with *staircase*, *corridor*, and *alley*, seems to encapsulate places designed for walking, driving, or other forms of locomotion. Among the indoor scenes, more sedentary locations cluster together, where one might sit and wait or do something quietly: *auditorium*, *subway_interior*, *living_room*, and *airport_terminal*.

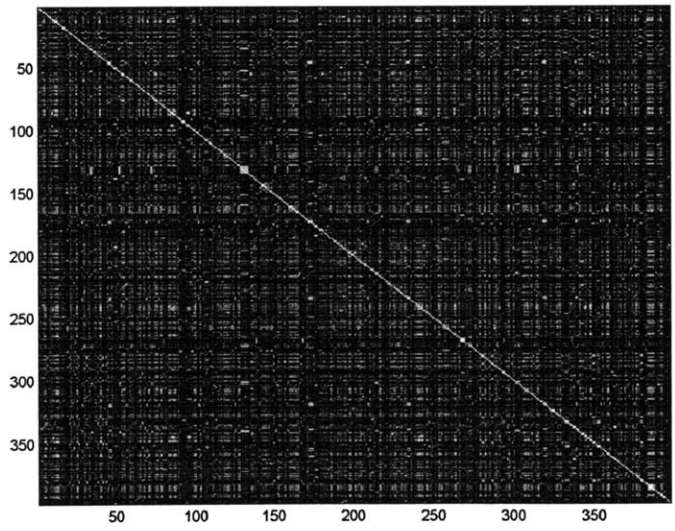
Finally, many of the scenes have very distinct action profiles from other scene types, and resist clustering together at all. Moderately clustering-resistent scene types, like *bedroom*, *office*, and *church/indoor*, are strongly associated with actions that are typically done in a specific place but may occasionally be done elsewhere ("sleep", "work", or "pray"), and are places where one might occasionally do something else. The most highly clustering-resistant scene types are strongly associated with actions that are never done in other places, and may in fact be literally inappropriate to do elsewhere: it is not generally permitted to gamble outside a casino, or to use the toilet outside a bathroom.

Visualizing the hierarchical clustering representation for all 397 scene types is extremely difficult. Category labels would be illegible at this scale, so Figure 4-4

shows an unlabeled dendrogram of the full dataset. A list of category labels arranged in clusters at an intermediate level of the hierarchy shown is listed in Table A.1. Overall, the pattern of grouping is similar to what was seen in the small dataset. Namely, the some scene types are very resistant to clustering (generally shown as uncolored black lines in the top third of the diagram), while other scene types group readily into large clusters (genreally shown as colored groupings in the bottom two-thirds of the diagram).

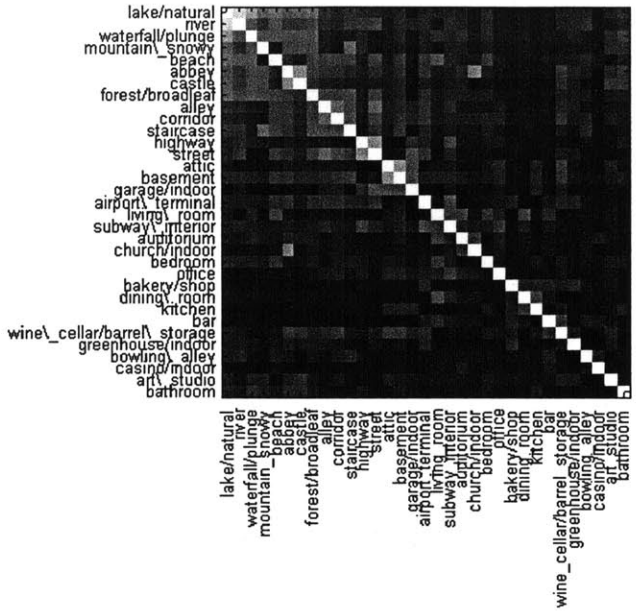


(a) 33-category subset

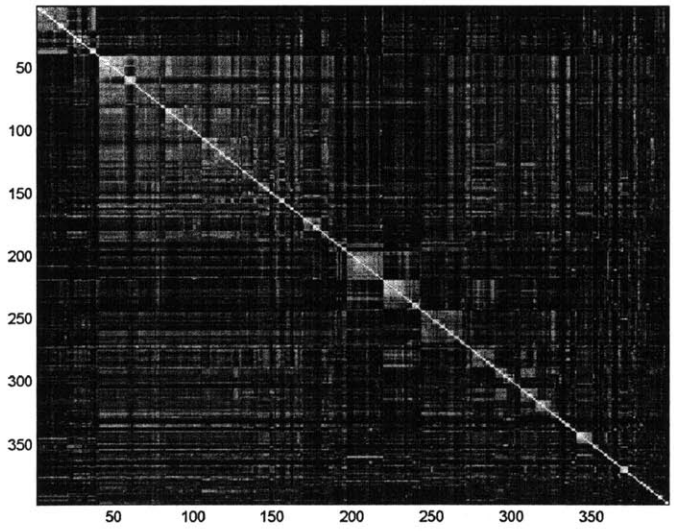


(b) 397 categories

Figure 4-1: Similarity heatmaps, for the 33-category subset and the full 397-category set



(a) 33-category subset



(b) 397 categories

Figure 4-2: Similarity heatmaps, for the 33-category subset and the full 397-category set, clustered by hierarchical clustering as described in

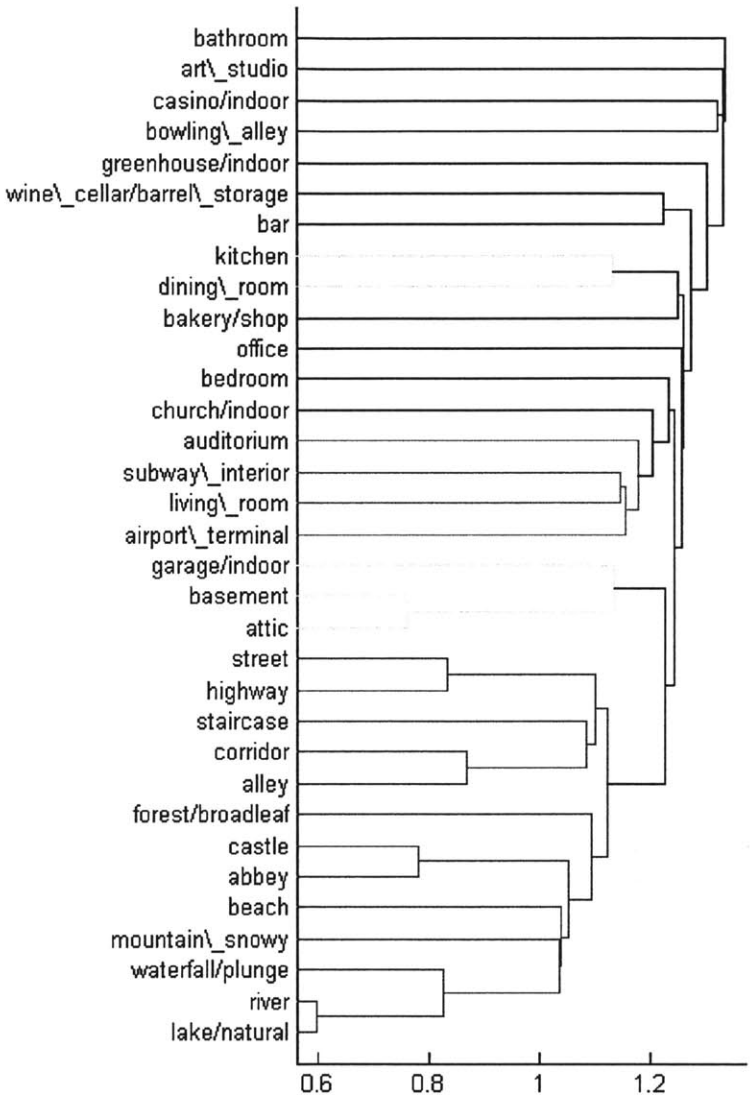


Figure 4-3: Dendrogram showing hierarchical clustering of the 33-category subset. Clusters separated by links of inconsistency coefficent greater than 1.2 are shown in color.

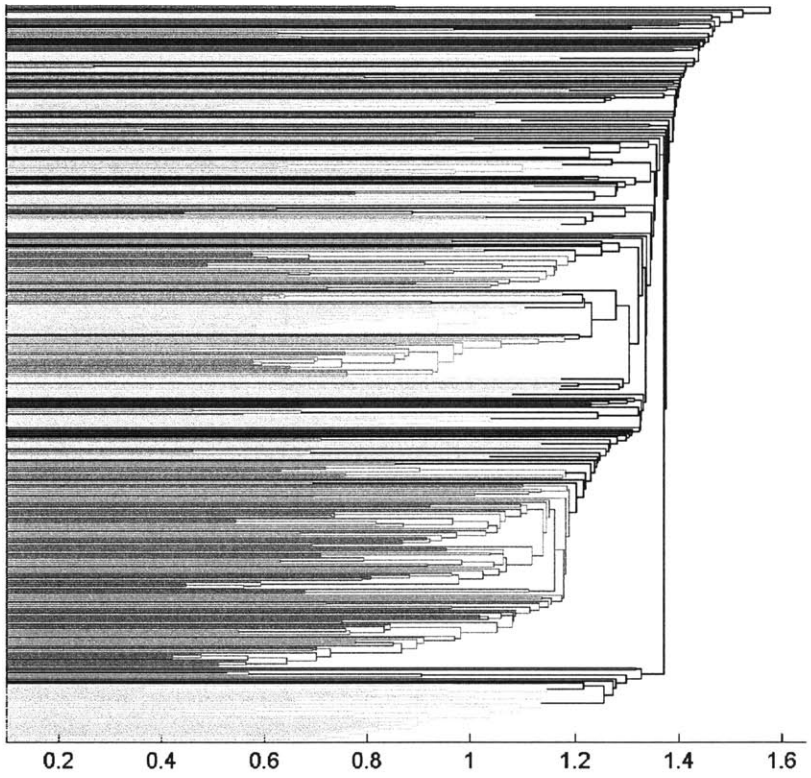


Figure 4-4: Dendrogram showing hierarchical clustering of the 397 categories. A set of subclusters at an intermediate level of clustering (cutoff = 1.2) are shown in color.

Chapter 5

Predicting Scene Type

One possible use of scene-action data is to predict what scene is being described from a set of typical actions. This is a classification problem: given data drawn from a particular category, infer the category label for that data.

Classifying scenes from their associated actions has two purposes. The first purpose is to learn more about the data itself: specifically, **Which scene types are unambiguously determined by their associated actions, and which scene types are easily confusable for others?**. One might expect that some scene types would be readily distinguished from others by a distinctive profile typical actions, whereas other scene types may not show a distinctive pattern of responses, or may be easily confused for other scene types with very similar action profiles.

The second purpose is to establish a baseline for future practical applications. **How well, in general, can we predict a category label from a list of typical actions?** Is the classification problem easy or hard? Possible future applications of predicting a scene type from a set of typical actions might include inferring location context from written material describing actions, such as instruction manuals or movie scripts; or using actions automatically tagged by action recognition on still images or movies to inform guesses of scene context in those images and movies.

In the field of machine learning, there are a variety of approaches to classification. Statistical approaches model the process that generated the data, and perform inference on that model. Other approaches do not perform statistical inference on a

model, but arrive at a predicted label through other algorithms. When the structure of the generative process can be inferred accurately, and its parameters estimated, statistical models can have more predictive power than non-statistical algorithms; the estimated parameters can also provide further insight into the structure of the generating process.

The goal of this work is to establish a baseline for classification accuracy, and to determine which categories are generally more confusable, *not* to richly model the joint distribution between scene type labels and participants' responses. Many hidden factors are likely to govern the process by which people infer the actions one might do in a location and select words to describe that action space. Furthermore, modeling the distribution of responses is made more difficult by the fact that words are discrete, non-numerical datapoints. Chapter 6 outlines some of the possible hidden factors, and presents ideas of directions for future work. Rather than elaborately modeling this possible underlying structure, instead we present two simple, non-parametric prediction algorithms.

The classification problem we tackle in this section is the following: Given n responses each provided as a typical action for a certain category (such as "swim", "relax", "snorkel", "walk on the beach", and "sit", five responses provided about the category "beach"), guess the associated category. We present two non-parametric algorithms to this problem: a *nearest centroid* classifier which is inspired by text classification approaches, and a simple *empirical maximum likelihood* classifier which performs statistical inference on a highly simplified distributional model based on the observed distribution of the data.

5.1 Nearest-centroid classification

This section explains the *nearest centroid*, the first classification algorithm I employ. The use of nearest centroid is inspird by text classification approaches. I first explain why text classification as a general class of approaches is relevant to the task of classifying scenes, and then explain why nearest centroid is particularly well-suited

to the task constraints.

5.1.1 Text classification strategies can inform scene-action classification

One domain in which classification strategies have been explored extensively is that of *text categorization*, in which a document (such as a website or other collection of natural-language text) must be categorized, often as relevant or irrelevant, for the purposes of retrieving documents based on a search query. In the text categorization literature, it is common to transform each document into a vector of word occurrences, where each vector element a_i indicates the weight of word i in the document (Salton *et al.*, 1975). The weight might be a binary variable indicating presence or absence of the word; a frequency count; a term-frequency/inverse-document-frequency (tf-idf) measure to capture the importance of each word to a given document (Manning *et al.*, 2008); or a more complicated weighting (Salton & Buckley, 1988). Additional transformations on this space might be performed, such as dimensionality reduction or feature selection (Aas & Eikvil, 1999). Document vectors are then compared in this high-dimensional space using a distance measure, such as Euclidean distance or cosine distance.

The scene-action classification problem bears a strong similarity to text classification, in that the task is to categorize a collection of natural language text using a corpus of training text. However, there are several subtleties that must be addressed in applying text-classification strategies to scene-action classification, each of which imposes its own constraints.

1. Queries will be small - just a handful of words or phrases - not entire "document-sized" collections of words.

2. It is not clear how to group participants' responses into "documents."

3. Our classifier must choose between 397 options - much more than a simple binary decision.

Queries are small

The queries we would like to classify will be short lists of actions, such as ("swim", "relax", "snorkel", "walk on the beach", "sit"). When represented in a term-frequency (or stem-frequency) vector space, small queries are represented as extremely sparse vectors, because most terms will have zero frequency. Distance measures such as Euclidean distance or cosine distance applied in this space may be problematic, because each "zero" entry will contribute to the similarity. This problem is additionally troublesome given that the total number of dimensions - that is, the number of words or stems in the lexicon - could be very large. Possible solutions include transforming to a new (perhaps lower-dimensional) vector space, or selecting a distance measure which treats the *presence* of terms as informative but ignores the *absence* of terms.

Document groupings are unclear

Responses were generated by subjects in groups of 1-3, not in coherent documents. One possible approach might be to treat each image as a single document even though it was "authored" by many participants. Another approach might be to select a classification mechanism which does not rely on individual-document distinctions, but instead measures query similarity to an entire class.

Classification decision is 397-way

Many classification algorithms work best for binary decisions. For example, some algorithms work by tabulating votes for each category label and selecting the label with the most votes. However, if each vote is for one out of 397 classes, it may take a large number of votes for any one class to emerge as the plurality winner.

5.1.2 Nearest-centroid classification is a simple approach that fulfills our constraints

One simple approach that addresses the subtleties described in the previous section is *nearest centroid classification.* In nearest centroid classification, a test datapoint is

classified by comparing to the centroid of each category in turn (that is, the *average* of all datapoints in each category) and selecting the nearest category centroid. When applied to text classification with an importance-weighting on words, this approach is known as Rocchio's algorithm (Aas & Eikvil, 1999). Our nearest-centroid classifier works as follows:

Data representation

Phrases are tokenized, stopwords are removed, and tokens are stemmed, as described in Section 4.1. A collection of phrases is transformed to a dictionary of stems and associated frequency counts. The frequency counts are then normalized. For example, the collection ("water plants", "buy plants", "garden") would first become {''water'': 1, ''buy'': 1, ''garden'': 1, ''plant'': 2}, and would then be normalized to {''water'': 0.2, ''buy'': 0.2, ''garden'': 0.2, ''plant'': 0.4}

Training

During training, the stem frequency count for each category is tabulated based on all the training data for that category, and is then normalized. This is equivalent to calculating the average over equal-sized collections of stems for each category.

Distance measure

During testing, the query is compared with each category centroid in turn using the following similarity measure (unnormalized histogram distance):

$$\Sigma_{keys} \min(a[key], b[key])$$

The category with the greatest similarity according to this measure is selected as the predicted category label.

Cross-validation

To estimate the classification accuracy over a dataset of phrases, the phrases are first grouped into queries of n responses (for values of n between 2 and 7). For example, ("water plants", "buy plants", "garden") is a 3-response query. These n-response queries are then split into k equal-sized subsamples for k-fold cross-validation.

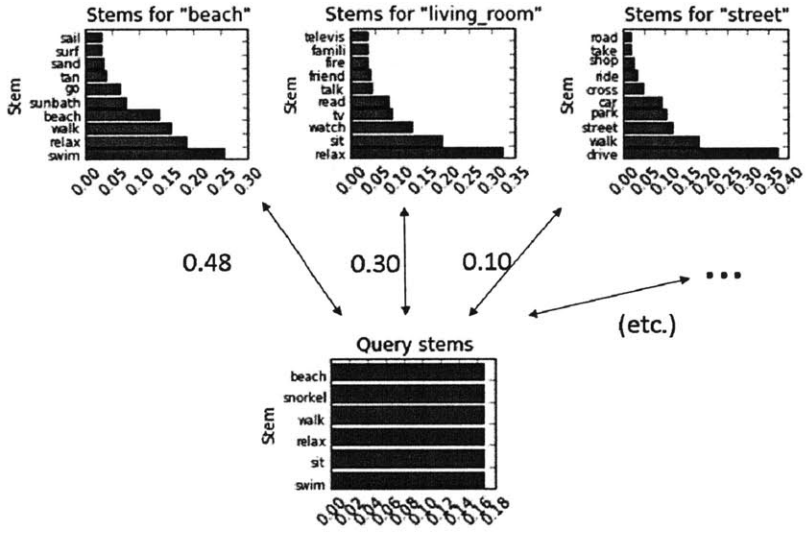


Figure 5-1: Illustration of the nearest centroid algorithm for an example five-response query: "swim", "relax", "snorkel", "walk on the beach", and "sit". The query is reduced to stems, and the normalized histogram of stems is compared to the normalized category histograms for each of the 397 categories (partial histograms for 3 categories shown here) using unnormalized histogram similarity. The category of greatest similarity (here *beach*) is selected as the predicted label.

Suitability for our task

Nearest-centroid is a very simple algorithm which is well-suited to the application of classifying scenes by their typical actions:

1. Histogram similarity does not incorporate stems with zero frequency, so the sparseness of queries does not cause any problems, nor is there any need to know the "true" dimension of the space of all stems.

2. Responses do not need to be grouped into documents - a class-wide average is

sufficient. Comparing to 397 class averages is more efficient than comparing to potentially thousands of individual documents.

3. A single nearest class is chosen, which is appropriate to a 397-way decision - no voting takes place.

4. This method has a direct parallel in the prototype model of categorization in cognitive science: that is, the idea that an item's category membership is determined by calculating its similarity to a prototype (Gardenfors & Williams, 2001).

Finally, it is worth noting that the method described here of normalizing frequency histograms and then performing an unnormalized similarity comparison is completely equivalent to performing a normalized similarity comparison (as described in Equation ??) on un-normalized frequency histograms. That is to say: when using normalized histogram similarity, comparing against the *average* of each category is equivalent to comparing against each category's *sum*.

5.1.3 Results: Nearest centroid shows 51% classification accuracy over 397 classes

Nearest-centroid classification was tested on both the 33-category subset and the full 397-category dataset, with 10-fold cross-validation, with test queries composed of n responses each for values of n between 2 and 7 (for example, ("swim", "relax") is a 2-response query, and ("swim", "relax", "snorkel", "walk on the beach", "sit", "lay in sand", "sunbathe") is a 7-response query). Most of the following plots show results from $n = 5$. At $n = 5$, classification accuracy over the 33-category subset was 87%, compared to a chance level of 3%; accuracy over all 397 categories was similarly high, at 50.75% compared to a chance level of 0.25%.

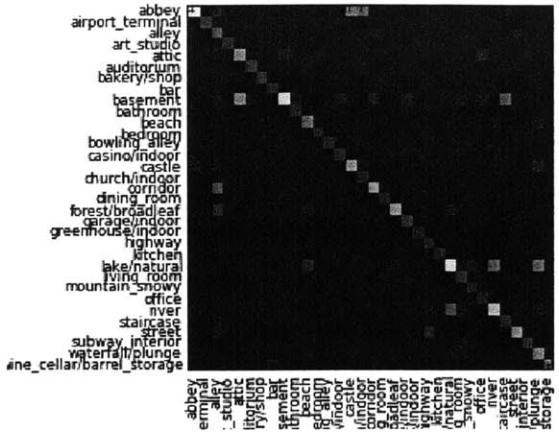
Figure 5-2 depicts heatmaps of the classification accuracy by class, for both the 33-category subset and all 397 categories. The 33-category dataset shows much better performance because most of the classes selected for inclusion in the subset are quite

distinct from one another in terms of actions that can be done there; Figure 5-3 illustrates that all but three classes in the 33-category subset achieve greater than 70% classification accuracy, whereas accuracies in the 397-category set are evenly distributed around the average, with 48% of the categories falling below the mean and 52% above.

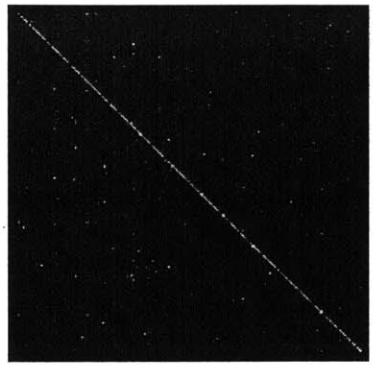
In the 397-category set, ten (10) categories were *always* classified correctly, and five (5) categories were *never* classified correctly, over ten folds of cross-validation, as shown in 5.1. The classes that achieved perfect classification accuracy are those that have very strongly-associated actions (like bowling in a bowling alley) or objects that are acted upon (like shoes in a shoe shop). The classes that achieved minimum classification accuracy afford very little affordance: places one might visit to look at (like a monastery outdoor) or would hope to never visit at all (like a burial chamber).

For example, *aqueduct* was one of the frequently confused classes. The most common stems for *aqueduct* are "take", "pictur", "walk", "drive", "bridge", "sightse", "explor", "visit", "tour", and "cross". The most easily confused categories for *aqueduct* are "ruin", "viaduct", "rope bridge", "rock arch", and "covered bridge exterior". This confusion set makes sense, as aqueducts appear bridge-like or arch-like, and do not afford particularly different actions from other attractions which are primarily of interest as historical or architectural artifacts to be viewed and photographed.

As shown in Figure 5-4, classification accuracy for the 33-category subset ranged from 77.16% with 2 responses per query, to 94.27% with 7 responses per query. Classification accuracy for the full 397-category set ranged from 38.07% with 2 responses per query, to 62.77% with 7 responses per query. This increasing pattern of accuracy is exactly what would be expected: The more responses are included in each query, the more information is available to the classifier for predicting that query's category.



(a) Accuracy over a 33-category subset was 87%, compared to a chance level of 3%.

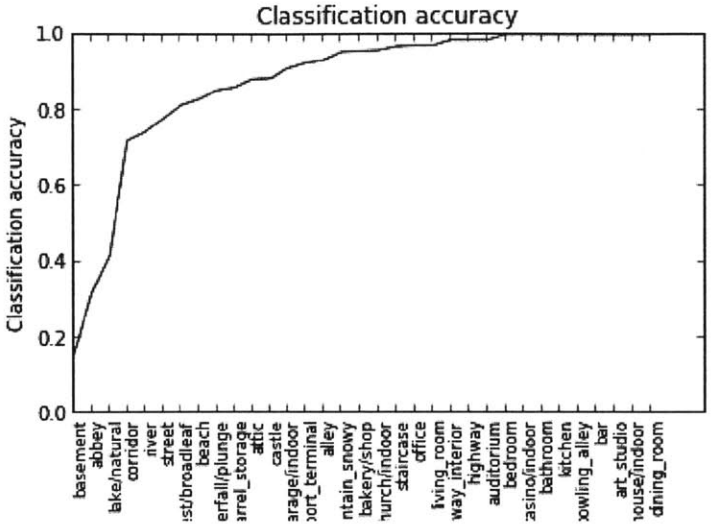


(b) Accuracy over all 397 categories was 50.75%, compared to a chance level of 0.25%.

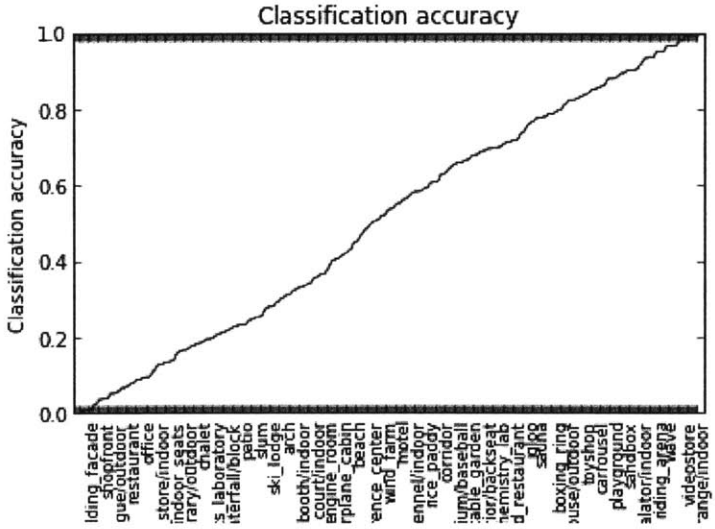
Figure 5-2: Heatmaps of classification accuracy by class for nearest centroid classification, over a 33-category subset and over the full 397-category set, using nearest-centroid with 5 responses per query. Although individual category labels are not shown for the full set, the overall color gives a sense of the pattern of classification.

Table 5.1: Out of the 397 categories, over ten folds of cross-validation, ten (10) categories were always classified correctly by nearest-centroid classification, and five (5) categories were never classified correctly.

Nearest centroid classification	
Always correct:	Never correct:
waiting room	monastery outdoor
shoe shop	inn outdoor
track outdoor	courtyard
phone booth	burial chamber
gas station	aqueduct
firing range indoor	
discotheque	
bowling alley	
baggage claim	
amusement arcade	

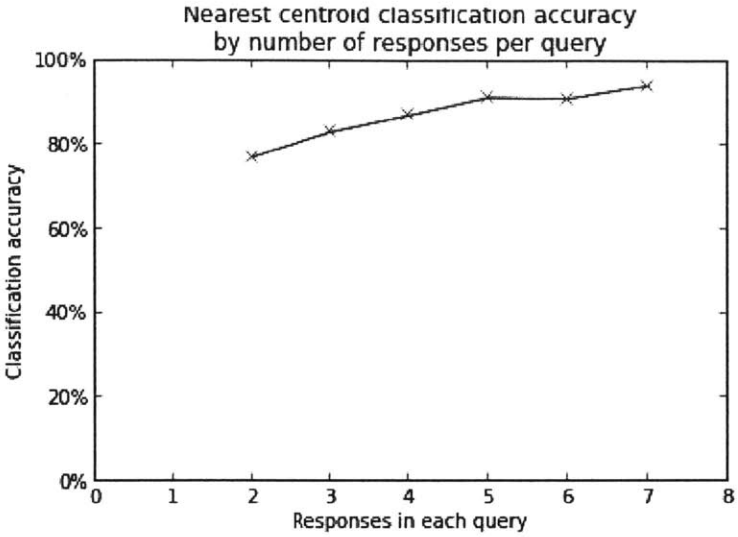


(a) 33-category subset

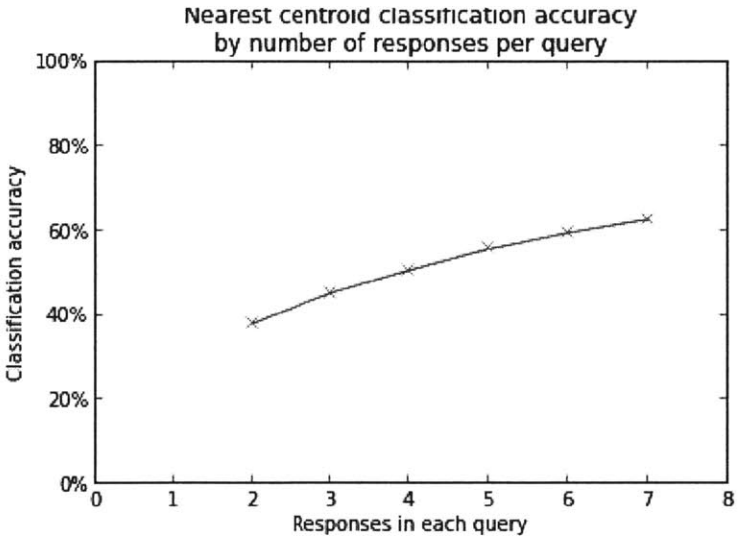


(b) Full 397-category dataset

Figure 5-3: Classification accuracy by class for nearest-centroid classification, over the 33-category subset and for all 397 categories, using 5 responses per query. A subset of category labels is shown for the 397-category plot. Classification accuracy varied widely over categories.



(a) 33-category subset



(b) Full 397-category dataset

Figure 5-4: Overall classification accuracy using nearest-centroid classification, as a function the number of responses per query. Each query, comprised of between 2 and 7 responses, was tested against the average query of each category. The more responses are included in each query, the more information is available to the classifier for predicting that query's category, and so the classifier performs better.

5.2 Empirical maximum likelihood classification

The second classification approach outlined here is a basic classifier based on the observed distribution of responses for each class. The central idea is to use simplifying assumptions to estimate the likelihood, Pr(*actions*|*scene*); if the prior on different scenes is non-uniform, then the likelihood can be combined with the prior Pr(*scene*) using Bayes' rule to estimate the posterior, Pr(*scene*|*actions*).

5.2.1 Bayes' Rule enables reasoning about hidden factors

Bayes' rule is a simple but powerful law of probability which relates conditional probabilities to one another, and is commonly used to reason about hidden factors based on observed data. In the general case:

$$\Pr(Y|X) = \frac{\Pr(X|Y) * \Pr(Y)}{\Pr(X)} \qquad = \frac{\Pr(X|Y) * \Pr(Y)}{\Sigma_Y \Pr(X|Y) * \Pr(Y)} \Pr(X)$$

The causal reasoning story behind this equation is that x a collection of observed data, and y is an unobserved cause or state. Based on some *prior* estimate of how probable a given unobserved state is ($\Pr(Y = y)$ and the *likelihood* of the observed data x given state y ($\Pr(X = x|Y = y)$), Bayes rule enables us to infer a *posterior* estimate of the probability of state y ($\Pr(Y = y|X = x)$ which incorporates the data we have observed.

Once the posterior distribution of unobserved states has been calculated, one way to draw a single guess from that distribution is by choosing the *maximum a posteriori* (MAP) estimate. If the prior distribution on unobserved states is assumed to be uniform, then the posterior is directly proportional to the likelihood, and the *maximum likelihood* (ML) estimate is equivalent to the MAP estimate.

5.2.2 Simplifying assumptions enable us to estimate the likelihood

By way of notation, we will choose $X = x_1, x_2, \ldots, x_k$ to be a list of k observed stems, drawn from an underlying distribution of stems associated with a single scene type $Y = y$.

We make two simplifying assumptions to enable us to easily estimate the likelihood $\Pr(x_1, x_2, \ldots, x_k|y)$: first, we assume that x_1 and x_2 are conditionally independent given y; and second, we estimate each $\Pr(x_i|y)$ using the empirical distribution.

Conditional independence of observed data

The first assumption we make is that, for a given scene category, the occurrence of any observed stem is *independent* of the other observed stems. That is to say:

$$\Pr(x_1, x_2, \ldots, x_k|y) = \Pr(x_1|y) * \Pr(x_2|y) * \ldots * \Pr(x_k|y)$$

This is known as the *Naive Bayes assumption.* This assumption is not a faithful approximation to the distribution of the data: even if the responses we are attempting to classify are in fact drawn randomly from the set of all possible responses (which would not be true if we asked one individual for a set of responses: they would be unlikely to give the same response twice), their constituent stems would not be independently distributed. For example, "buy shoes" is a very common response for "shoe shop", which means that the stem "buy" and the stem "shoe" covary strongly when conditioned on the scene type "shoe shop".

Although this assumption an unrealistic oversimplification, Naive Bayes classifiers based on the conditional independence assumption have been shown to perform well in many real-world applications (Rish, 2001).

Empirical distribution estimate of the posterior

The second assumption we make is that the class-conditional probability of a given stem can be accurately estimated directly from the observed frequency of that stem (with the exception of stems with no occurrences in a given class, as explained below).

We estimate the probability of stem x_k given class y to be the number of occurrences of x_k in class y divided by the total number of stems observed in class y:

$$\Pr(x_k|y) = \frac{(\#x_k|y)}{\Sigma_i(\#x_i|y)}$$

The problem with this approach is that it does not handle stems that have not been seen with a given class. It is not reasonable to assume that never having seen a certain stem associated with a certain class implies that such a pairing is literally impossible (zero probability). A more reasonable assumption is that such a pairing is extremely rare.

There are many possible approaches for dealing with unseen stem-class pairings. One possible approach is to add a "pseudocount" to all frequency counts. Another possibility is to adopt a more sophisticated frequency estimation strategy, such as Good-Turing frequency estimation (Good, 1953). Good-Turing is an algorithm is an algorithm to generate smoothed estimates of the species probability of an unknown number of species, based on past observations of objects drawn from those species. This algorithm assumes that the total probability for all never-before-seen species is $\frac{N_1}{N}$ where N_1 is the number of single-occurrence species, and adjusts the rest of the estimates by smoothing.

However, rather than adopting a sophisticated estimation strategy, we use some simple rules to fudge our estimations for never-before-seen stems. If we see a stem that has never been seen in the entire training set, we simply ignore it, as it would provide no information in support of any of the classes under the scheme described here. If we see a stem that has been seen elsewhere in the training set, but not for

this particular class y, then we assign its probability to be the smallest $\Pr(x|y)$ value otherwise calculated (that is, the probability assigned to a 1-ocurrence stem in the class with most stems). This strategy has the advantage of making the calculation of $\Pr(x_1, x_2, \ldots, x_n|y)$ very simple, but it has the downside that it does *not* actually lead to a legitimate probability distribution $\Pr(X|Y)$. Small "probability" has been assigned to many never-before-seen stem-scene pairings, and the sum over all possible values of X for a given scene is much greater than 1. Nonetheless, this shortcut works adequately for our purposes.

Log likelihoods avoid numeric precision difficulties

The computation of the likelihood $\Pr(x_1, x_2, \ldots, x_n|y)$ requires n probabilities, which might be small.

$$\Pr(x_1, x_2, \ldots, x_n|y) = \Pi_{k=1\ldots n} \Pr(x_k|y)$$

Rather than computing the likelihood directly, we compute the logarithm of the likelihood, which enables us to *add* log likelihoods, avoiding numeric precision errors which arise when working with small numbers.

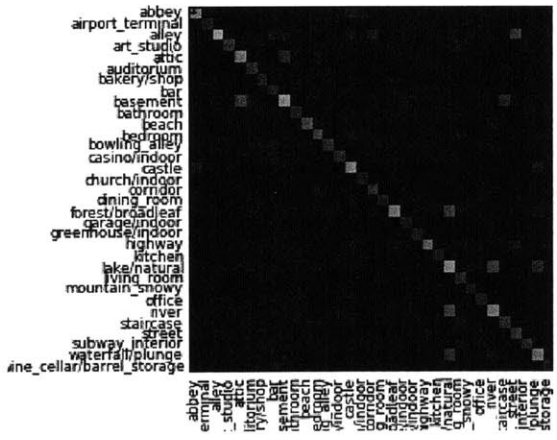
$$\log \Pr(x_1, x_2, \ldots, x_n|y) = \Sigma_{k=1\ldots n} \log \Pr(x_k|y)$$

5.2.3 Results: Maximum likelihood shows 67% classification accuracy over 397 classes

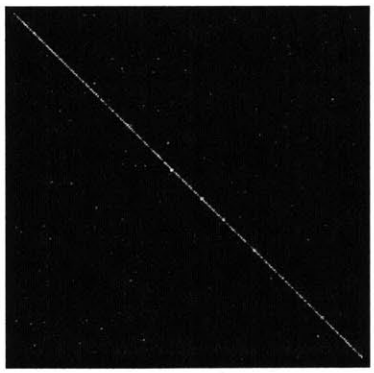
Maximum likelihood classification was tested on both the 33-category subset and the full 397-category dataset, with 10-fold cross-validation, with test queries composed of n responses each for values of n between 2 and 7. At $n = 5$, classification accuracy over the 33-category subset was 94%, compared to a chance level of 3%; accuracy over all 397 categories was similarly high, at 67% compared to a chance level of 0.25%.

Figure 5-5 depicts heatmaps of the classification accuracy by class, for both the 33-category subset and all 397 categories. The 33 categories are very difficult to confuse with one another: an attic occasionally has the action profile of a basement; lakes, rivers, and waterfalls have very similar action profile; an airport terminal sometimes seems like a subway interior; and a highway sometimes seems like a street; but otherwise, most entries in the 33-category set are quite distinct. Figure 5-6 shows a line plot of classification accuracy by class for each dataset, illustrating a rounded curve for both categories: a few categories in each dataset are fairly easily confused with other scene types, but most of them are readily distinguished with one another. Unlike with nearest centroid classification, none of the categories were never classified correctly.

As shown in Figure 5-7, classification accuracy for the 33-category subset ranged from 88% with 2 responses per query, to 97% with 7 responses per query. Classification accuracy for the full 397-category set ranged from 55% with 2 responses per query, to 74% with 7 responses per query.



(a) Accuracy over a 33-category subset was 94%, compared to a chance level of 3%.



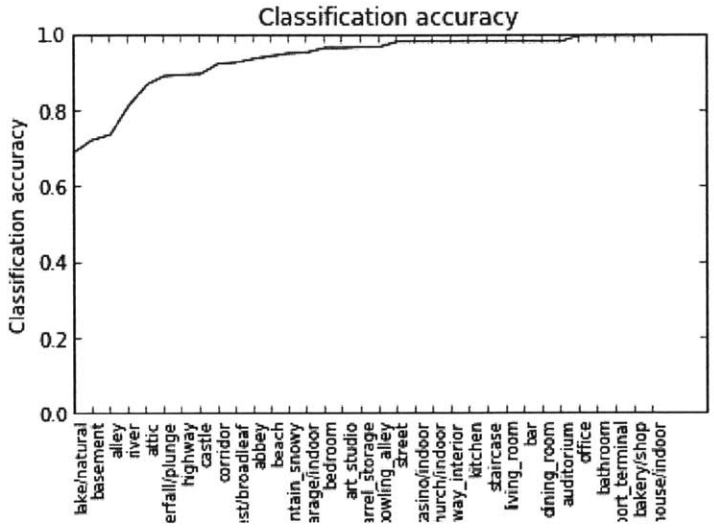
(b) Accuracy over all 397 categories was 66.60%, compared to a chance level of 0.25%.

Figure 5-5: Heatmaps of maximum likelihood classification accuracy by class, over a 33-category subset and over the full 397-category set, using nearest-centroid with 5 responses per query. Although individual category labels are not shown for the full set, the overall color gives a sense of the pattern of classification.

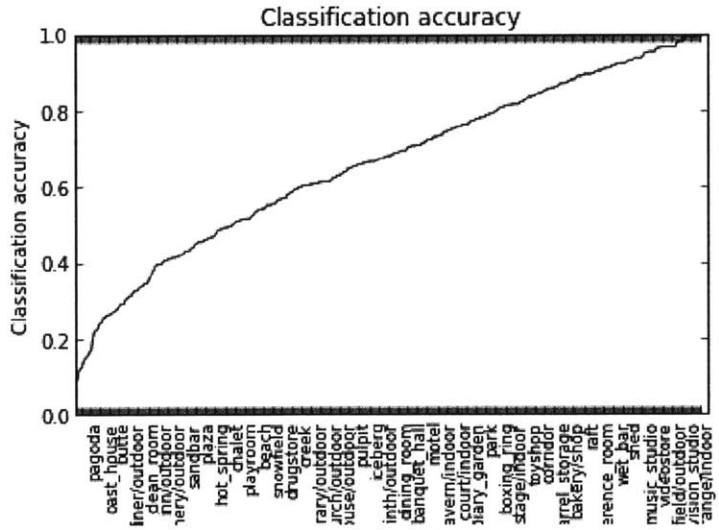
5.3 Discussion

Why is classification performance so good with these two simple classifiers? The answer is perhaps because any given set of responses is likely to contain at least one very diagnostic stem. The long tail of subjects' responses is to our advantage in this respect. Subjects describe actions in specific terms, rather than restricting their answers to common verbs like "walk" and "eat." In this detail they provide a large amount of context-specific information. For example, "highway" is distinguished from a "street" or a "car_interior/frontseat" by relatively rare stems such as "speed", "race", and "hitchik".

The empirical maximum likelihood classifier performed better than the nearest centroid clasifier, at 67% with $n = 5$-response-queries as compared to 51%. That said, the comparison is not a fair one because the nearest-centroid weighting does not currently incorporate any importance-weighting. It would be interesting to incorporate a weighting such as TF-IDF (Manning *et al.*, 2008) into the nearest-neighbor classifier.

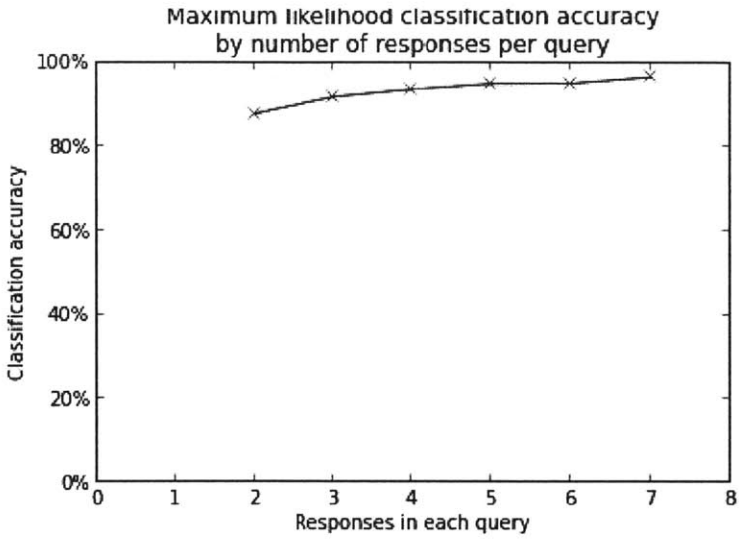


(a) 33-category subset

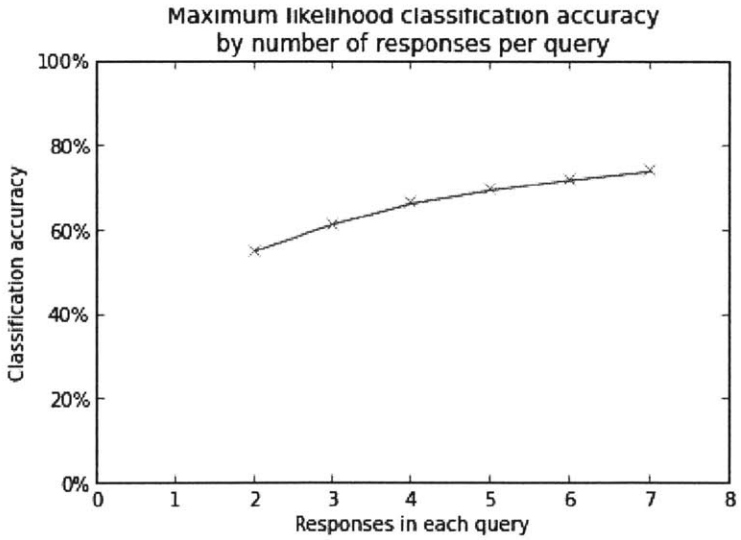


(b) Full 397-category dataset

Figure 5-6: Classification accuracy by class for maximum likelihood classification, over the 33-category subset and for all 397 categories, using 5 responses per query. A subset of category labels is shown for the 397-category plot.



(a) 33-category subset



(b) Full 397-category dataset

Figure 5-7: Overall classification accuracy using maximum likelihood classification, as a function the number of responses per query. The empirical maximum likelihood class for each query was computed. Each query was comprised of between 2 and 7 responses; with more responses per query, classifier performance improved.

Chapter 6

Future Directions

There are a variety of potential future directions for this work, in both computer science and cognitive science. Some of the most promising directions are elaborated in detail below, along with briefer mentions of more nebulous directions.

6.1 Examine the effects of category name on actions

In the present study, subjects were not told the name of the scene type they were viewing a picture of. How would subjects' responses change if they were told the name of the place as listed in the collection of 397 SUN categories? The SUN category name does not necessarily correspond to the "basic-level" category name that subjects themselves would typically use to describe the scene type (Rosch & Lloyd, 1978) — for example, subjects are unlikely to use the word "apse" to refer to the semicircular recess at the front of a church, or "medina" to refer to a certain type of historic district in a North African city. In this study, the experiment would be run again, but subjects would be told the fine-grained category name and definition for the image they were viewing. The hypothesis is that subjects' descriptions in response to named scenes would more precisely describe scene types that are visually hard to distinguish. This may lead to better classification accuracy and more diversity of responses.

In particular, I expect that increased classification accuracy would be seen primarily for categories whose SUN category name does *not* match the basic level of categorization — that is, categories which SUN separates, but which subjects themselves do not distinguish between. By contrast, categories whose SUN category name matches subjects' own judgments of scene type would not show much difference in subjects' responses. The relationship between changes in subjects' responses and mismatch between the SUN categorization structure and subjects' basic-level categorization structure could be verified by further asking subjects to provide basic-level category labels for the images shown, and looking for a correlation between the set of images that subjects most often group together using their own categorization, and the set of images that show the msot improved classification accuracy in the class-labeled dataset as compared to the no-category-name-shown dataset.

6.2 Correlate objects, materials, and spatial properties with actions

Visually speaking, scenes are made up of objects and materials, arranged in a physical space of some shape and size. How much are these objects, materials, and properties diagnostic of typical actions? Intuitively, it seems that many of the actions people report are driven by the objects, materials, and spatial properties of the scene (such as "bake" if an oven is present, "swim" if a body of water is present, or "walk" if the space is long and narrow), as contrasted with phrases that are driven primarily by the known or inferred purpose of the scene (such as "heal animals" for a vet's office even if no animals or medical equipment are immediately visible).

A potential future question is whether some actions are more strongly associated with specific physical properties, and other actions are more strongly associated with more abstract scene types. Some actions might best be explained by the presence or absence of certain objects, materials, and spatial properties independent of scene type, whereas some actions may be best explained as properties relating to the overall

inferred purpose of the whole scene as a property above and beyond the objects, materials, and spatial properties it is composed of.

One path forward in investigating this question is to use ground truth annotation of objects, materials, and spatial properties of each scene, and measure the correlation of these attributes with the actions provided by subjects. For example, ground truth annotation of objects is available in the LabelMe dataset (Russell *et al.*, 2008); the SUN attribute dataset (Patterson & Hays, 2012) contains a variety of material properties ("concrete", "vegetation", or "rusty"), spatial properties (such as "no horizon", "open area", or "symmetrical"), semantic properties (such as "soothing" or "man-made"); and Oliva et al. have generated a database ranked on subjective visual complexity (Oliva *et al.*, 2004). These attributes could be correlated with the occurrence of particular stems or phrases to look for the strongest correspondences, or more complex generative models could be constructed.

Another possible path forward is to use computer vision to detect or approximate these attributes. How well can the typical actions for a scene be predicted from the features extracted by standard computer vision algorithms? Are certain "visual words" as extracted by SIFT (Lowe, 1999) associated with certain actions? The gist descriptor can give an approximation of properties like openness, expansion, naturalness, roughness, and so on (Oliva & Torralba, 2001); do these inferred properties correlate with different actions?

6.3 Compare the similarity space of actions with other similarity spaces

One contribution of this work as described in Section 4 was to extract a similarity space for scenes based on their associated actions. Does this similarity space resemble other similarity spaces extracted from other scene data by other measures? For example, does the hierarchical clustering correlate at all with the SUN 7-layer hierarchy (Xiao *et al.*, 2010)? Or with scene types that are found physically close to one another

in space? Or with which scene types show more similar neural pattern activation in the human brain (Mur *et al.*, 2009), and in which brain areas?

In performing these comparisons, similarity heatmaps can be compared directly using correlation or Euclidean distances, or the extracted clusterings can be compared using methos designed specifically for hierarchical clustering (Fowlkes & Mallows, 1983) or with more general clustering-comparison methods (Meil, 2005) (Hubert & Arabie, 1985).

6.4 Analyze image-to-image variability

As mentioned in Chapter 2, this dataset was gathered for typical images in order to better elucidate category-to-category variation. However, investigating image-to-image variation is a natural next step. For example, images vary in typicality rating; one might hypothesize that an image's distance from the category average in action-similarity-space would be anti-correlatered with its typicality rating, such that exemplars rated as more typical do in fact elicit more typical responses.

Examining image-to-image variation might explain some of the category-level variation. For example, images in the category "kasbah" can often look quite ruined, so when the classifiers described in Chapter 5 classify a kasbah as a ruin, that may not actually be a particularly noteworthy misclassification, unlike misclassifications of a "highway" as a "car interior", which are distinct locations despite having similar action profiles.

Finally, any image property could be correlated with the properties described in this work, such as response diversity, average response length, classification accuracy, and so on. For example, do cluttered scenes generate a greater diversity of responses? Are man-made scenes easier to classify? Candidate image properties include the attributes for memorability as labeled in in (Isola *et al.*, 2011); the number and type of LabelMe objects (Russell *et al.*, 2008); the SUN attributes described in (Patterson & Hays, 2012); or visual complexity as described in (Oliva *et al.*, 2004).

6.5 Richly model the joint distribution

The probability of a certain action phrase being produced in response to a certain image is a product of many unobserved factors. While it would be impossible to model literally all the factors underlying subjects' response patterns, a richer model than the simple empirical model described in Chapter 5 would not only shed more light on these unobserved factors, but may also show improved classification performance, and could generalize to unobserved phrases and scenes. If scenes and words are each modeled as being composed of a set of features, or generated by latent factors, then graphical models could be employed that relate the representations to the other according to a generative probabilistic structure.

Existing work in the space of matching words to actions, such as the work of Blei (Blei *et al.*, 2003) and Barnard (Barnard & Forsyth, 2001) (Barnard *et al.*, 2003) in the space of image captions, could be readily be applied to the domain of typical actions. As an alternative to the graphical models approach, more sophisticated latent topic models could be borrowed from text classification, including non-probabilistic models such as latent semantic analysis (Manning *et al.*, 2008). As demonstrated in Chapter 5, generic term-document models can readily be translated to the problem of associating actions with scene types. If a latent topic model were employed, it would be interesting to visually inspect the topics it generates, to see if any of them correspond to identifiable scene features.

One final possibility would be to model the dependency structure of actions; for example, a tree structured dependency model would be relatively simple to build using the Chow-Liu algorithm (Chow & Liu, 1968), as demonstrated in (Choi *et al.*, 2010).

6.6 Incorporate semantic properties of natural language

Aside from tokenizing and stemming, this work largely ignores the rich semantic content of natural language. Projects such as Wordnet (Princeton University, 2010) model the synonym structure of word senses in the English language. Furthermore, a variety of word similarity measures have been devised, some based on structures like Wordnet, and some based on statistical properties such as corpus co-occurrence (Pedersen *et al.*, 2004).

A few obstacles remain before the semantic hierarchy of Wordnet can be leveraged towards understanding subjects' responses. For one, subject's responses must be *disambiguated*: words have multiple meanings ("senses"), and while wordsense disambiguation algorithms do exist (Warin, 2004) (Resnik, 1995), they tend to function best with more surrounding context than the responses gathered here tend to provide. Furthermore, even once wordsense has been disambiguated, the WordNet hierarchy does not capture all the relatedness between terms: for example, "bathe" and "swim" are unrelated according to the WordNet hierarchy, because bathing is a type of cleansing, and swimming is a type of travel.

6.7 Relate action words directly to constituent objects

Many of subjects' responses describe objects directly visible in the scene. A webtool could be deployed to allow users to directly connect object polygons provided by LabelMe with the words they use to describe an image. A prototype for such a tool has been designed, although the future steps for potential word-object correspondence data are still unclear.

6.8 Gather responses for scene type names without images

In this study we showed people images without names for two reasons: one, because we intend to eventually link typical actions with visually-detectable properties, and two, because we wanted to gathe subjects' responses to the scenes themselves rather than associations with the words. As described in Section 6.1 we hope to compare reponses to images alone with responses to images with names and definitions. What if we don't show any picture at all? How do subjects' responses differ if they have only the name of a scene type and a definition, and no visual information at all?

Chapter 7

Contributions

In summary, the SUN Action project comprises the first effort to gather and analyze unconstrained natural-language annotations of typical actions over a comprehensive set of scene types. This work offers the following contributions to the field of scene understanding:

- I gathered a large dataset of typical actions by scene type, consisting of over 100,000 free-form responses over all 397 "well-sampled" SUN scene categories.
- I analyzed the response distribution and diversity, overall and between categories.
- I defined a similarity space over actions, and demonstrated heterogeneous hierarchical clustering over this space.
- I implemented two classifiers that successfuly classify scenes by their associated actions at considerably higher than chance rates.
- I outlined a plan for future work in both computer science and human vision.

Appendix A

Additional tables

This appendix includes tables of verbatim responses and figures that are too long to include in the main text.

All 397 Scene Types

'abbey'
'airplane_cabin'
'airport_terminal'
'alley'
'amphitheater'
'amusement_arcade'
'amusement_park'
'anechoic_chamber'
'apartment_building/outdoor'
'apse/indoor'
'aquarium'
'aqueduct'
'arch'
'archive'
'arrival_gate/outdoor'
'art_gallery'
'art_school'
'art_studio'
'assembly_line'
'athletic_field/outdoor'
'atrium/public'
'attic'
'auditorium'
'auto_factory'
'badlands'
'badminton_court/indoor'
'baggage_claim'
'bakery/shop'
'balcony/exterior'
'balcony/interior'
'ball_pit'
'ballroom'
'bamboo_forest'
'banquet_hall'
'bar'
'barn'
'barndoor'
'baseball_field'
'basement'
'basilica'
'basketball_court/outdoor'
'bathroom'
'batters_box'
'bayou'
'bazaar/indoor'
'bazaar/outdoor'
'beach'
'beauty_salon'
'bedroom'
'berth'
'biology_laboratory'
'bistro/indoor'
'boardwalk'
'boat_deck'

'boathouse'
'bookstore'
'booth/indoor'
'botanical_garden'
'bow_window/indoor'
'bow_window/outdoor'
'bowling_alley'
'boxing_ring'
'brewery/indoor'
'bridge'
'building_facade'
'bullring'
'burial_chamber'
'bus_interior'
'butchers_shop'
'butte'
'cabin/outdoor'
'cafeteria'
'campsite'
'campus'
'canal/natural'
'canal/urban'
'candy_store'
'canyon'
'car_interior/backseat'
'car_interior/frontseat'
'carrousel'
'casino/indoor'
'castle'
'catacomb'
'cathedral/indoor'
'cathedral/outdoor'
'cavern/indoor'
'cemetery'
'chalet'
'cheese_factory'
'chemistry_lab'
'chicken_coop/indoor'
'chicken_coop/outdoor'
'childs_room'
'church/indoor'
'church/outdoor'
'classroom'
'clean_room'
'cliff'
'cloister/indoor'
'closet'
'clothing_store'
'coast'
'cockpit'
'coffee_shop'
'computer_room'
'conference_center'
'conference_room'
'construction_site'
'control_room'
'control_tower/outdoor'
'corn_field'
'corral'
'corridor'
'cottage_garden'
'courthouse'
'courtroom'
'courtyard'
'covered_bridge/exterior'
'creek'
'crevasse'
'crosswalk'
'cubicle/office'
'dam'
'delicatessen'
'dentists_office'
'desert/sand'
'desert/vegetation'
'diner/indoor'
'diner/outdoor'
'dinette/home'
'dinette/vehicle'
'dining_car'
'dining_room'
'discotheque'
'dock'
'doorway/outdoor'
'dorm_room'
'driveway'
'driving_range/outdoor'
'drugstore'
'electrical_substation'
'elevator/door'
'elevator/interior'
'elevator_shaft'
'engine_room'
'escalator/indoor'
'excavation'

'factory/indoor'
'fairway'
'fastfood_restaurant'
'field/cultivated'
'field/wild'
'fire_escape'
'fire_station'
'firing_range/indoor'
'fishpond'
'florist_shop/indoor'
'food_court'
'forest/broadleaf'
'forest/needleleaf'
'forest_path'
'forest_road'
'formal_garden'
'fountain'
'galley'
'game_room'
'garage/indoor'
'garbage_dump'
'gas_station'
'gazebo/exterior'
'general_store/indoor'
'general_store/outdoor'
'gift_shop'
'golf_course'
'greenhouse/indoor'
'greenhouse/outdoor'
'gymnasium/indoor'
'hangar/indoor'
'hangar/outdoor'
'harbor'
'hayfield'
'heliport'
'herb_garden'
'highway'
'hill'
'home_office'
'hospital'
'hospital_room'
'hot_spring'
'hot_tub/outdoor'
'hotel/outdoor'
'hotel_room'
'house'
'hunting_lodge/outdoor'
'ice_cream_parlor'
'ice_floe'
'ice_shelf'
'ice_skating_rink/indoor'
'ice_skating_rink/outdoor'
'iceberg'
'igloo'
'industrial_area'
'inn/outdoor'
'islet'
'jacuzzi/indoor'
'jail/indoor'
'jail_cell'
'jewelry_shop'
'kasbah'
'kennel/indoor'
'kennel/outdoor'
'kindergarden_classroom'
'kitchen'
'kitchenette'
'labyrinth/outdoor'
'lake/natural'
'landfill'
'landing_deck'
'laundromat'
'lecture_room'
'library/indoor'
'library/outdoor'
'lido_deck/outdoor'
'lift_bridge'
'lighthouse'
'limousine_interior'
'living_room'
'lobby'
'lock_chamber'
'locker_room'
'mansion'
'manufactured_home'
'market/indoor'
'market/outdoor'
'marsh'
'martial_arts_gym'
'mausoleum'
'medina'
'moat/water'
'monastery/outdoor'
'mosque/indoor'

'mosque/outdoor'
'motel'
'mountain'
'mountain_snowy'
'movie_theater/indoor'
'museum/indoor'
'music_store'
'music_studio'
'nuclear_power_plant/outdoor'
'nursery'
'oast_house'
'observatory/outdoor'
'ocean'
'office'
'office_building'
'oil_refinery/outdoor'
'oilrig'
'operating_room'
'orchard'
'outhouse/outdoor'
'pagoda'
'palace'
'pantry'
'park'
'parking_garage/indoor'
'parking_garage/outdoor'
'parking_lot'
'parlor'
'pasture'
'patio'
'pavilion'
'pharmacy'
'phone_booth'
'physics_laboratory'
'picnic_area'
'pilothouse/indoor'
'planetarium/outdoor'
'playground'
'playroom'
'plaza'
'podium/indoor'
'podium/outdoor'
'pond'
'poolroom/establishment'
'poolroom/home'
'power_plant/outdoor'
'promenade_deck'
'pub/indoor'
'pulpit'
'putting_green'
'racecourse'
'raceway'
'raft'
'railroad_track'
'rainforest'
'reception'
'recreation_room'
'residential_neighborhood'
'restaurant'
'restaurant_kitchen'
'restaurant_patio'
'rice_paddy'
'riding_arena'
'river'
'rock_arch'
'rope_bridge'
'ruin'
'runway'
'sandbar'
'sandbox'
'sauna'
'schoolhouse'
'sea_cliff'
'server_room'
'shed'
'shoe_shop'
'shopfront'
'shopping_mall/indoor'
'shower'
'skatepark'
'ski_lodge'
'ski_resort'
'ski_slope'
'sky'
'skyscraper'
'slum'
'snowfield'
'squash_court'
'stable'
'stadium/baseball'
'stadium/football'
'stage/indoor'
'staircase'
'street'

'subway_interior'
'subway_station/platform'
'supermarket'
'sushi_bar'
'swamp'
'swimming_pool/indoor'
'swimming_pool/outdoor'
'synagogue/indoor'
'synagogue/outdoor'
'television_studio'
'temple/east_asia'
'temple/south_asia'
'tennis_court/indoor'
'tennis_court/outdoor'
'tent/outdoor'
'theater/indoor_procenium'
'theater/indoor_seats'
'thriftshop'
'throne_room'
'ticket_booth'
'toll_plaza'
'topiary_garden'
'tower'
'toyshop'
'track/outdoor'
'train_railway'
'train_station/platform'
'tree_farm'
'tree_house'
'trench'
'underwater/coral_reef'
'utility_room'
'valley'
'van_interior'
'vegetable_garden'
'veranda'
'veterinarians_office'
'viaduct'
'videostore'
'village'
'vineyard'
'volcano'
'volleyball_court/indoor'
'volleyball_court/outdoor'
'waiting_room'
'warehouse/indoor'
'water_tower'
'waterfall/block'
'waterfall/fan'
'waterfall/plunge'
'watering_hole'
'wave'
'wet_bar'
'wheat_field'
'wind_farm'
'windmill'
'wine_cellar/barrel_storage'
'wine_cellar/bottle_storage'
'wrestling_ring/indoor'
'yard'
'youth_hostel'

Stop words

The following words "stopwords" carrying little semantic information, as defined in the `nltk` python module in `nltk.corpus.stopwords.words('english')`, were removed before further morphological processing:

'i',	'your',	'her',	'theirs',
'me',	'yours',	'hers',	'themselves',
'my',	'yourself',	'herself',	'what',
'myself',	'yourselves',	'it',	'which',
'we',	'he',	'its',	'who',
'our',	'him',	'itself',	'whom',
'ours',	'his',	'they',	'this',
'ourselves',	'himself',	'them',	'that',
'you',	'she',	'their',	'these',

```
'those',          'or',             'down',           'most',
'am',             'because',        'in',             'other',
'is',             'as',             'out',            'some',
'are',            'until',          'on',             'such',
'was',            'while',          'off',            'no',
'were',           'of',             'over',           'nor',
'be',             'at',             'under',          'not',
'been',           'by',             'again',          'only',
'being',          'for',            'further',        'own',
'have',           'with',           'then',           'same',
'has',            'about',          'once',           'so',
'had',            'against',        'here',           'than',
'having',         'between',        'there',          'too',
'do',             'into',           'when',           'very',
'does',           'through',        'where',          's',
'did',            'during',         'why',            't',
'doing',          'before',         'how',            'can',
'a',              'after',          'all',            'will',
'an',             'above',          'any',            'just',
'the',            'below',          'both',           'don',
'and',            'to',             'each',           'should',
'but',            'from',           'few',            'now'
'if',             'up',             'more',
```

Phrase diversity

Least diverse: 'shower' with 51 phrases:

```
(112, 'shower')                    (1, 'washing')
(78, 'take a shower')              (1, 'wash your body')
(19, 'take shower')                (1, 'wash up')
(17, 'showering')                  (1, 'wash oneself')
(12, 'clean')                      (1, 'wash hands')
(12, 'bathe')                      (1, 'wash body')
(11, 'get clean')                  (1, 'turn on the shower')
(7, 'wash')                        (1, 'taking shower')
(7, 'taking a shower')             (1, 'take showers')
(5, 'relax')                       (1, 'take a luxurious shower')
(4, 'wash hair')                   (1, 'take a bath')
(4, 'turn on water')               (1, 'stand')
(3, 'use shower')                  (1, 'sit')
(3, 'scrub')                       (1, 'sing')
(3, 'clean the shower')            (1, 'showing')
(2, 'take bath')                   (1, 'shower oneself')
(2, 'rinse')                       (1, 'shave')
(2, 'poop')                        (1, 'putting in new glass')
(2, 'clean oneself')               (1, 'pee')
```

```
(1, 'look out window')
(1, 'getting clean')
(1, 'get washed')
(1, 'enter shower')
(1, 'dry yourself')
(1, 'dry off with a towel')
(1, 'cleanse')
(1, 'clean yourself')
(1, 'clean up')
(1, 'clean shower dooer')
(1, 'bathing')
(1, 'bath')
(1, 'adjust water')
```

Most diverse: 'veterinarians_office' with 253 phrases:

```
(11, 'examine')
(8, 'pet dog')
(6, 'help animals')
(5, 'pet animals')
(4, 'wait')
(4, 'talk to vet')
(4, 'pet the dog')
(4, 'pet')
(4, 'operate')
(4, 'check')
(3, 'take care of dog')
(3, 'take care of animals')
(3, 'pet cat')
(3, 'heal')
(3, 'get pet examined')
(3, 'examine dog')
(3, 'examine a cat')
(3, 'exam')
(3, 'care for pet')
(2, 'worry')
(2, 'watch the vet')
(2, 'visiting the vet')
(2, 'taking care of animals')
(2, 'pet animal')
(2, 'pay a bill')
(2, 'nursing')
(2, 'helping animals')
(2, 'heal animals')
(2, 'get checkup')
(2, 'get check up')
(2, 'examine the dog')
(2, 'examine pets')
(2, 'examine pet')
(2, 'draw blood')
(2, 'comfort the dog')
(2, 'comfort dog')
(2, 'checkup')
(2, 'caring')
(2, 'care for dog')
(1, 'work as assistant')
(1, 'work')
(1, 'watch the vet examine my dog')
(1, 'wash dog')
(1, 'waiting')
(1, 'visit the veterinarian')
(1, 'visit the vet')
(1, 'visit')
(1, 'vaccinating')
(1, 'vaccinate pet')
(1, 'treating animals.')
(1, 'treating a pet')
(1, 'treat their pets')
(1, 'treat sick animals.')
(1, 'treat pet')
(1, 'treat dog')
(1, 'treat animals')
(1, 'treat animal')
(1, 'treat an animal')
(1, "treat a pet's illness")
(1, 'thank the veternarian')
(1, 'test')
(1, 'tend to pet')
(1, 'talk to the doctor')
(1, 'taking the animal to the vet')
(1, 'taking pet to the vet')
(1, 'taking care of the dog')
(1, 'taking care of pets')
(1, 'taking care of a dog')
(1, 'taking animal for medical care')
(1, 'take x-rays')
(1, 'take vitals')
(1, 'take temperatures')
(1, 'take pet to the vet')
(1, "take dog's temperature")
(1, 'take dog to vet')
(1, 'take care of pets')
```

```
(1, 'take animal to vet')
(1, 'take a sit pet for care')
(1, 'take a sick pet')
(1, 'take a pet')
(1, 'spay or neuter a dog')
(1, 'spay')
(1, 'soothe kitty')
(1, 'sit in waiting room')
(1, 'sit down')
(1, 'shake hands')
(1, 'seeing the health of the animal')
(1, 'say "good boy"')
(1, 'saving an animal')
(1, 'run tests')
(1, 'rescue animals')
(1, 'rescue an animal')
(1, 'request')
(1, 'remove tick')
(1, 'receive treatment')
(1, 'put on leash')
(1, 'protect your pets')
(1, 'play with the dog')
(1, 'play with cats')
(1, 'petting dog')
(1, 'petting animal')
(1, 'pet your cat.')
(1, 'pet vaccinations')
(1, 'pet the kitty')
(1, 'pet the iguana')
(1, 'pet the cat')
(1, 'pet the animal')
(1, 'pet puppy')
(1, 'pet pup')
(1, 'pet kitty')
(1, 'pet getting shot')
(1, 'pet checkup')
(1, 'pet an animal')
(1, 'pet a cat')
(1, 'perform surgery')
(1, 'perform medical examination')
(1, 'perform animal surgery')
(1, 'perform a medical exam')
(1, 'operate on animals in need')
(1, 'neuter cat')
(1, 'medically care for animals')
(1, 'measuring')
(1, 'making the cat feel better')
(1, 'maintain your cats heath')
(1, 'listening to the cat heartbeat')
(1, 'listen to the vet')
(1, "listen to a cat's heart")
(1, 'listen')
(1, 'learn to take care of animals')
(1, 'keep your pet healthy')
(1, 'keep pet calm')
(1, 'keep animals healthy')
(1, 'inspect animal for injuries')
(1, 'inject')
(1, 'hug pet')
(1, 'hope for the best')
(1, 'hold their breath')
(1, 'hold the dog')
(1, 'hold dog')
(1, 'hold an animal')
(1, 'helping an animal')
(1, 'helping a golden retriever')
(1, 'help with the kitten')
(1, 'help the veterinarian')
(1, 'help the dog')
(1, 'help pet')
(1, 'help my pet')
(1, 'help animials')
(1, 'help an animal')
(1, 'help a cat')
(1, 'help')
(1, 'health check')
(1, 'healing pets')
(1, 'healing')
(1, 'heal the tiger')
(1, 'heal cat')
(1, 'heal a dog')
(1, 'have pet looked at')
(1, 'groom dog')
(1, 'going to the veterinarian.')
(1, 'go to vet')
(1, 'go to the vet')
(1, 'giving the cat a shot')
(1, 'give vetrinary care')
(1, 'give the dog medicine')
(1, 'give shots')
(1, 'give shot to pet')
(1, 'give shot')
(1, 'give dog shots')
(1, 'give cat a shot')
```

```
(1, 'give animals a shot')
(1, 'give a shot')
(1, 'give  injection')
(1, 'getting pet healthy')
(1, 'getting medical care for pet')
(1, 'getting better')
(1, 'get your pet checked up')
(1, 'get your dog checked')
(1, 'get your cat checked')
(1, 'get treatment')
(1, 'get their cat checked out')
(1, 'get shots')
(1, 'get scratched')
(1, 'get results')
(1, 'get puppy shots')
(1, 'get puppy checked up')
(1, 'get pet well')
(1, 'get pet checkup')
(1, 'get pet checked up')
(1, 'get pet checked')
(1, 'get pet a healthcheck')
(1, 'get exam')
(1, 'get dog exam')
(1, 'get checkup for pets')
(1, 'get checkup for pet')
(1, 'get care for your pet')
(1, 'get bath')
(1, 'get animal examined')
(1, 'get a checkup')
(1, 'fix animals')
(1, 'fix')
(1, 'examining animals')
(1, 'examining a collie')
(1, 'examining')
(1, 'examine the cat')
(1, 'examine cat')
(1, 'examine animals')
(1, 'euthanize dog')
(1, 'euthanize animals')
(1, 'do surgery')
(1, 'diagnose dog')
(1, 'diagnose cat')
(1, 'cut nails')
(1, 'converse')
(1, 'comfort the puppy')
(1, 'comfort the animal')
(1, 'comfort')
(1, 'checkup for pet')
(1, 'checkup dog')
(1, 'checking up on the dog')
(1, 'checking up on dog')
(1, 'checking up')
(1, 'checking pet')
(1, 'checking out the dog')
(1, 'checking on cat')
(1, "checking dog's health")
(1, 'checking')
(1, 'check vitals')
(1, 'check up on your pets')
(1, 'check the kitty')
(1, 'check the heartrate')
(1, 'check out pets')
(1, 'check on your pet')
(1, 'check on pets')
(1, 'check on pet')
(1, 'check health')
(1, 'check dog')
(1, 'check cat.')
(1, 'check animal')
(1, 'care for animals')
(1, 'care for animal')
(1, 'care for an animal')
(1, 'care for')
(1, 'care')
(1, 'calm the dog')
(1, 'buy meds')
(1, 'bring pet')
(1, 'bring in pet')
(1, 'be a vet')
(1, 'ask the vet questions')
(1, 'ask questions')
(1, 'ask question')
(1, 'animal checkup')
```

Stem diversity

Least diverse: "shoe shop" with 41 stems;

```
(255, 'shoe')
(128, 'buy')
(92, 'tri')
(67, 'shop')
(19, 'purchas')
(13, 'brows')
(10, 'look')
(6, 'find')
(5, 'new')
(5, 'boot')
(4, 'pick')
(4, 'pair')
(2, 'work')
(2, 'wear')
(2, 'sneaker')
(2, 'size')
(2, 'sell')
(2, 'run')
(2, 'make')
(2, 'go')
(2, 'fit')
(2, 'decis')
(2, 'around')
(1, 'winter')
(1, 'window')
(1, 'touch')
(1, 'talk')
(1, 'take')
(1, 'shp')
(1, 'sensibl')
(1, 'select')
(1, 'right')
(1, 'organ')
(1, 'nike')
(1, 'measur')
(1, 'loafer')
(1, 'like')
(1, 'hoe')
(1, 'get')
(1, 'cloth')
(1, 'admir')
```

Most diverse: "slum" with 181 stems:

```
(63, 'live')
(29, 'clean')
(21, 'walk')
(16, 'help')
(11, 'surviv')
(11, 'food')
(9, 'slum')
(8, 'water')
(8, 'peopl')
(8, 'cloth')
(7, 'poverti')
(7, 'look')
(7, 'hang')
(7, 'beg')
(6, 'trash')
(6, 'feed')
(5, 'visit')
(5, 'take')
(5, 'ride')
(5, 'needi')
(5, 'laundri')
(5, 'hungri')
(5, 'garbag')
(5, 'build')
(4, 'sleep')
(4, 'shop')
(4, 'search')
(4, 'repair')
(4, 'pick')
(4, 'move')
(4, 'feel')
(4, 'explor')
(4, 'donat')
(4, 'car')
(4, 'boat')
(4, 'barter')
(4, 'around')
(3, 'work')
(3, 'wish')
(3, 'wash')
(3, 'villag')
(3, 'swim')
(3, 'struggl')
(3, 'shelter')
(3, 'resid')
(3, 'poor')
(3, 'money')
(3, 'hous')
```

```
(3, 'home')
(3, 'go')
(3, 'find')
(3, 'drive')
(3, 'catch')
(3, 'carri')
(3, 'buy')
(3, 'aid')
(2, 'way')
(2, 'tri')
(2, 'town')
(2, 'talk')
(2, 'street')
(2, 'starv')
(2, 'sorri')
(2, 'run')
(2, 'road')
(2, 'restor')
(2, 'rebuild')
(2, 'play')
(2, 'pictur')
(2, 'mess')
(2, 'market')
(2, 'local')
(2, 'litter')
(2, 'leav')
(2, 'hut')
(2, 'good')
(2, 'get')
(2, 'diseas')
(2, 'damag')
(2, 'cri')
(2, 'away')
(2, 'avoid')
(1, 'young')
(1, 'werent')
(1, 'view')
(1, 'upset')
(1, 'travel')
(1, 'toss')
(1, 'time')
(1, 'suppli')
(1, 'sun')
(1, 'suffer')
(1, 'subsist')
(1, 'stuff')
(1, 'stick')
(1, 'squat')
(1, 'squalor')
(1, 'spend')
(1, 'someth')
(1, 'soccer')
(1, 'sightse')
(1, 'shanti')
(1, 'sell')
(1, 'seek')
(1, 'scrap')
(1, 'sad')
(1, 'rummag')
(1, 'rubbag')
(1, 'row')
(1, 'rob')
(1, 'river')
(1, 'renov')
(1, 'relax')
(1, 're')
(1, 'purchas')
(1, 'provid')
(1, 'prepar')
(1, 'poverish')
(1, 'poorli')
(1, 'pollut')
(1, 'photo')
(1, 'pail')
(1, 'paddl')
(1, 'occup')
(1, 'observ')
(1, 'neighbor')
(1, 'neighboorhood')
(1, 'neigbor')
(1, 'necess')
(1, 'mud')
(1, 'missionari')
(1, 'mingl')
(1, 'meat')
(1, 'make')
(1, 'line')
(1, 'ladder')
(1, 'kid')
(1, 'juryrig')
(1, 'junk')
(1, 'inspect')
(1, 'insid')
(1, 'inhabit')
```

(1, 'hill')	(1, 'dig')
(1, 'hepat')	(1, 'die')
(1, 'happi')	(1, 'death')
(1, 'ground')	(1, 'cook')
(1, 'greet')	(1, 'convers')
(1, 'give')	(1, 'commut')
(1, 'gather')	(1, 'collect')
(1, 'friend')	(1, 'climb')
(1, 'fix')	(1, 'cleanup')
(1, 'fish')	(1, 'citi')
(1, 'fight')	(1, 'chariti')
(1, 'empath')	(1, 'builld')
(1, 'eat')	(1, 'box')
(1, 'dri')	(1, 'bike')
(1, 'distribut')	(1, 'ball')
(1, 'dispos')	(1, 'bad')
(1, 'disea')	(1, 'area')
(1, 'disast')	(1, 'aorund')
(1, 'dirti')	(1, 'age')
(1, 'dirt')	

Table A.1: Cluster members for an intermediate level of hierarchical clustering.

Cluster members
'firing_range/indoor'
'martial_arts_gym'
end cluster
'amusement_arcade'
'athletic_field/outdoor'
'auditorium'
'badminton_court/indoor'
'ball_pit'
'baseball_field'
'basketball_court/outdoor'
'batters_box'
'boxing_ring'
'bullring'
'classroom'
'conference_center'
'conference_room'
'courtroom'
'driving_range/outdoor'
'fairway'
'game_room'
'golf_course'
'kindergarden_classroom'
Continued on next page

Table A.1 – continued from previous page
Cluster members
'lecture_room'
'movie_theater/indoor'
'playground'
'playroom'
'podium/indoor'
'poolroom/establishment'
'poolroom/home'
'putting_green'
'recreation_room'
'sandbox'
'squash_court'
'stadium/baseball'
'stadium/football'
'stage/indoor'
'tennis_court/indoor'
'tennis_court/outdoor'
'theater/indoor_procenium'
'theater/indoor_seats'
'volleyball_court/indoor'
'volleyball_court/outdoor'
'wrestling_ring/indoor'
end cluster
'assembly_line'
'auto_factory'
'biology_laboratory'
'brewery/indoor'
'cheese_factory'
'chemistry_lab'
'clean_room'
'computer_room'
'construction_site'
'control_room'
'cubicle/office'
'electrical_substation'
'engine_room'
'excavation'
'factory/indoor'
'home_office'
'industrial_area'
'nuclear_power_plant/outdoor'
'office'
'oil_refinery/outdoor'
'oilrig'
'physics_laboratory'
Continued on next page

Table A.1 – continued from previous page
Cluster members
'power_plant/outdoor'
'server_room'
'trench'
'wind_farm'
'wine_cellar/barrel_storage'
'wine_cellar/bottle_storage'
end cluster
'art_gallery'
'museum/indoor'
'observatory/outdoor'
end cluster
'balcony/exterior'
'balcony/interior'
'bow_window/indoor'
'bow_window/outdoor'
'gazebo/exterior'
'hot_tub/outdoor'
'jacuzzi/indoor'
'living_room'
'parlor'
'patio'
'pavilion'
'picnic_area'
'sauna'
'veranda'
end cluster
'archive'
'attic'
'bakery/shop'
'banquet_hall'
'bar'
'basement'
'bazaar/indoor'
'bazaar/outdoor'
'bistro/indoor'
'bookstore'
'butchers_shop'
'cafeteria'
'candy_store'
'clothing_store'
'coffee_shop'
'delicatessen'
'diner/indoor'
'diner/outdoor'
Continued on next page

Table A.1 – continued from previous page
Cluster members
'dinette/home'
'dinette/vehicle'
'dining_car'
'dining_room'
'drugstore'
'fastfood_restaurant'
'florist_shop/indoor'
'food_court'
'galley'
'gas_station'
'general_store/indoor'
'general_store/outdoor'
'gift_shop'
'ice_cream_parlor'
'jewelry_shop'
'kitchen'
'kitchenette'
'library/indoor'
'market/indoor'
'market/outdoor'
'music_store'
'pharmacy'
'pub/indoor'
'restaurant'
'restaurant_kitchen'
'restaurant_patio'
'shed'
'shoe_shop'
'shopfront'
'shopping_mall/indoor'
'slum'
'supermarket'
'sushi_bar'
'thriftshop'
'toyshop'
'videostore'
'warehouse/indoor'
'wet_bar'
end cluster
'fountain'
end cluster
'ticket_booth'
end cluster
'toll_plaza'
Continued on next page

Table A.1 – continued from previous page
Cluster members
end cluster
'abbey'
'alley'
'amphitheater'
'apartment_building/outdoor'
'apse/indoor'
'aquarium'
'aqueduct'
'arch'
'atrium/public'
'badlands'
'bamboo_forest'
'barn'
'barndoor'
'basilica'
'bayou'
'beach'
'boardwalk'
'boat_deck'
'boathouse'
'botanical_garden'
'bridge'
'building_facade'
'burial_chamber'
'butte'
'cabin/outdoor'
'campus'
'canal/natural'
'canal/urban'
'canyon'
'car_interior/frontseat'
'castle'
'catacomb'
'cathedral/indoor'
'cathedral/outdoor'
'cavern/indoor'
'chalet'
'church/indoor'
'church/outdoor'
'cliff'
'cloister/indoor'
'coast'
'corn_field'
'corridor'
Continued on next page

Table A.1 – continued from previous page
Cluster members
'cottage_garden'
'courthouse'
'courtyard'
'covered_bridge/exterior'
'creek'
'crevasse'
'crosswalk'
'dam'
'desert/sand'
'desert/vegetation'
'dock'
'doorway/outdoor'
'field/cultivated'
'field/wild'
'fishpond'
'forest/broadleaf'
'forest/needleleaf'
'forest_path'
'forest_road'
'formal_garden'
'greenhouse/indoor'
'greenhouse/outdoor'
'harbor'
'hayfield'
'herb_garden'
'highway'
'hill'
'hospital'
'hot_spring'
'hotel/outdoor'
'house'
'hunting_lodge/outdoor'
'ice_floe'
'ice_shelf'
'iceberg'
'igloo'
'inn/outdoor'
'islet'
'kasbah'
'labyrinth/outdoor'
'lake/natural'
'library/outdoor'
'lido_deck/outdoor'
'lift_bridge'
'lighthouse'
Continued on next page

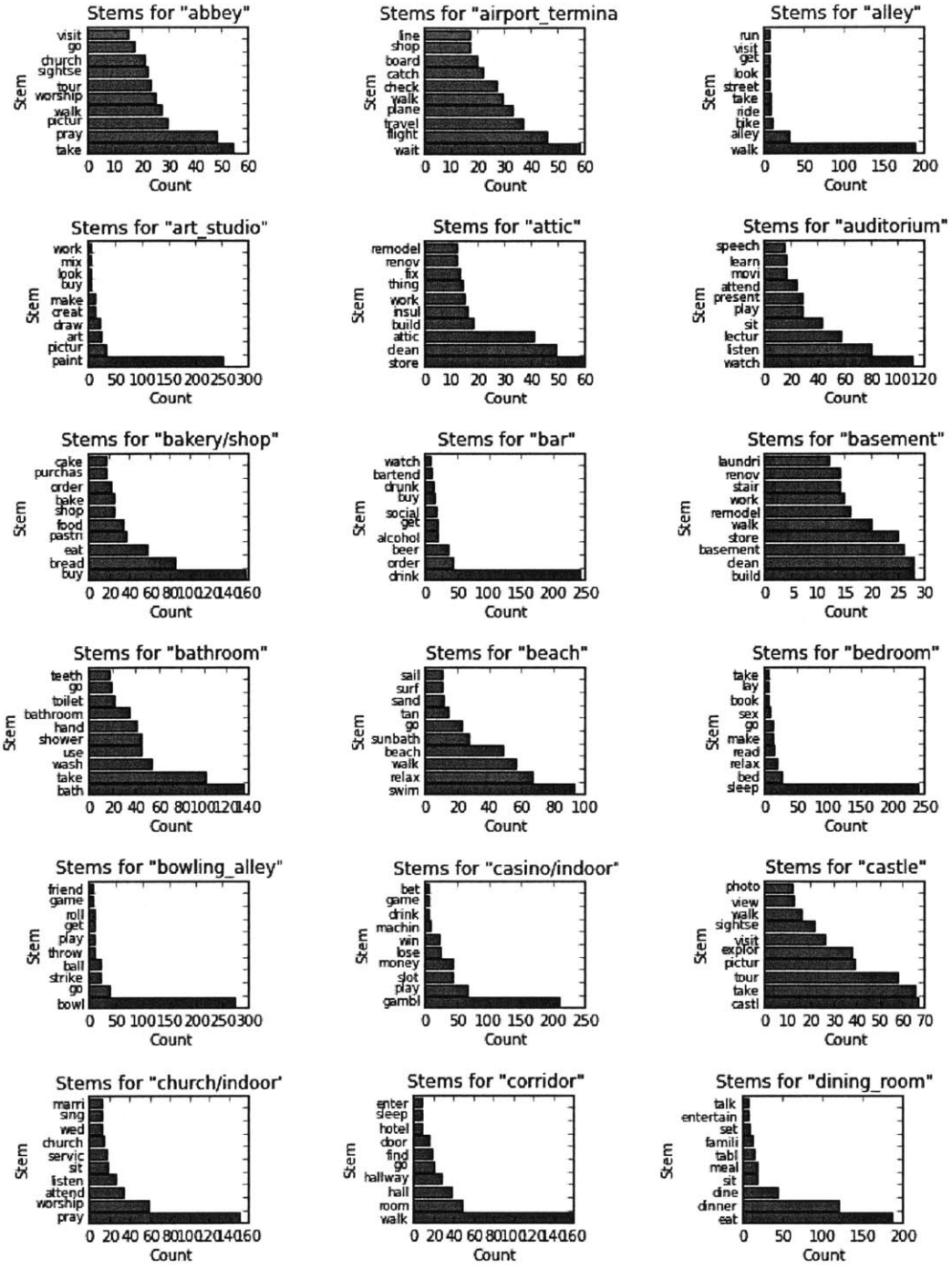
Table A.1 – continued from previous page
Cluster members
'lock_chamber'
'mansion'
'manufactured_home'
'marsh'
'medina'
'moat/water'
'monastery/outdoor'
'mosque/indoor'
'mosque/outdoor'
'mountain'
'mountain_snowy'
'oast_house'
'ocean'
'orchard'
'pagoda'
'palace'
'park'
'planetarium/outdoor'
'plaza'
'pond'
'promenade_deck'
'pulpit'
'rainforest'
'residential_neighborhood'
'rice_paddy'
'river'
'rock_arch'
'rope_bridge'
'ruin'
'sandbar'
'schoolhouse'
'sea_cliff'
'ski_lodge'
'ski_resort'
'ski_slope'
'sky'
'snowfield'
'street'
'swamp'
'swimming_pool/indoor'
'swimming_pool/outdoor'
'synagogue/indoor'
'synagogue/outdoor'
'temple/east_asia'
'temple/south_asia'
Continued on next page

Table A.1 – continued from previous page
Cluster members
'throne_room'
'topiary_garden'
'tower'
'tree_farm'
'underwater/coral_reef'
'valley'
'vegetable_garden'
'viaduct'
'village'
'vineyard'
'volcano'
'water_tower'
'waterfall/block'
'waterfall/fan'
'waterfall/plunge'
'watering_hole'
'wheat_field'
'windmill'
end cluster
'jail/indoor'
end cluster
'airplane_cabin'
'airport_terminal'
'arrival_gate/outdoor'
'bus_interior'
'car_interior/backseat'
'cockpit'
'control_tower/outdoor'
'hangar/indoor'
'hangar/outdoor'
'heliport'
'landing_deck'
'limousine_interior'
'railroad_track'
'runway'
'subway_interior'
'subway_station/platform'
'train_railway'
'train_station/platform'
'van_interior'
end cluster
'amusement_park'
'beauty_salon'
'carrousel'
Continued on next page

Table A.1 – continued from previous page
Cluster members
'corral'
'dentists_office'
'elevator/door'
'elevator/interior'
'escalator/indoor'
'hospital_room'
'kennel/indoor'
'kennel/outdoor'
'operating_room'
'pasture'
'riding_arena'
'stable'
'veterinarians_office'
end cluster
'gymnasium/indoor'
end cluster
'racecourse'
'raceway'
end cluster
'chicken_coop/indoor'
'chicken_coop/outdoor'
end cluster
'office_building'
'skyscraper'
end cluster
'laundromat'
'utility_room'
end cluster
'closet'
'locker_room'
end cluster
'ballroom'
'discotheque'
end cluster
'track/outdoor'
end cluster
'music_studio'
end cluster
'television_studio'
end cluster
'fire_escape'
'staircase'
end cluster
'tree_house'
Continued on next page

Table A.1 – continued from previous page
Cluster members
end cluster
'booth/indoor' **end cluster**
'campsite' 'tent/outdoor' **end cluster**
'bedroom' 'berth' 'childs_room' 'dorm_room' 'hotel_room' 'jail_cell' 'motel' 'nursery' 'youth_hostel' **end cluster**
'elevator_shaft' **end cluster**
'phone_booth' **end cluster**
'baggage_claim' **end cluster**
'reception' **end cluster**
'waiting_room' **end cluster**
'garbage_dump' 'landfill' **end cluster**
'lobby' **end cluster**
'yard' **end cluster**
'driveway' 'garage/indoor' 'parking_garage/indoor' 'parking_garage/outdoor' 'parking_lot' **end cluster**
'raft' **end cluster**
'bathroom' 'outhouse/outdoor' 'shower'
Continued on next page

Table A.1 – continued from previous page
Cluster members
end cluster
'fire_station' **end cluster**
'anechoic_chamber' **end cluster**
'podium/outdoor' **end cluster**
'bowling_alley' **end cluster**
'ice_skating_rink/indoor' 'ice_skating_rink/outdoor' 'skatepark' **end cluster**
'wave' **end cluster**
'pilothouse/indoor' **end cluster**
'pantry' **end cluster**
'casino/indoor' **end cluster**
'art_school' 'art_studio' **end cluster**
'cemetery' 'mausoleum'


Stems for "abbey"
visit
go
church
sightse
tour
worship
walk
pictur
pray
take
Stem
0 10 20 30 40 50 60
Count
Stems for "airport_termina
line
shop
board
catch
check
walk
plane
travel
flight
wait
Stem
0 10 20 30 40 50 60
Count
Stems for "alley"
run
visit
get
look
street
take
ride
bike
alley
walk
Stem
0 50 100 150 200
Count
Stems for "art_studio"
work
mix
look
buy
make
creat
draw
art
pictur
paint
Stem
0 50 100 150 200 250 300
Count
Stems for "attic"
remodel
renov
fix
thing
work
insul
build
attic
clean
store
Stem
0 10 20 30 40 50 60
Count
Stems for "auditorium"
speech
learn
movi
attend
present
play
sit
lectur
listen
watch
Stem
0 20 40 60 80 100 120
Count
Stems for "bakery/shop"
cake
purchas
order
bake
shop
food
pastri
eat
bread
buy
Stem
0 20 40 60 80 100 120 140 160
Count
Stems for "bar"
watch
bartend
drunk
buy
social
get
alcohol
beer
order
drink
Stem
0 50 100 150 200 250
Count
Stems for "basement"
laundri
renov
stair
work
remodel
walk
store
basement
clean
build
Stem
0 5 10 15 20 25 30
Count
Stems for "bathroom"
teeth
go
toilet
bathroom
hand
shower
use
wash
take
bath
Stem
0 20 40 60 80 100 120 140
Count
Stems for "beach"
sail
surf
sand
tan
go
sunbath
beach
walk
relax
swim
Stem
0 20 40 60 80 100
Count
Stems for "bedroom"
take
lay
book
sex
go
make
read
relax
bed
sleep
Stem
0 50 100 150 200 250
Count
Stems for "bowling_alley"
friend
game
roll
get
play
throw
ball
strike
go
bowl
Stem
0 50 100 150 200 250 300
Count
Stems for "casino/indoor"
bet
game
drink
machin
win
lose
money
slot
play
gambl
Stem
0 50 100 150 200 250
Count
Stems for "castle"
photo
view
walk
sightse
visit
explor
pictur
tour
take
castl
Stem
0 10 20 30 40 50 60 70
Count
Stems for "church/indoor"
marri
sing
wed
church
servic
sit
listen
attend
worship
pray
Stem
0 20 40 60 80 100 120 140 160
Count
Stems for "corridor"
enter
sleep
hotel
door
find
go
hallway
hall
room
walk
Stem
0 20 40 60 80 100 120 140 160
Count
Stems for "dining_room"
talk
entertain
set
famili
tabl
meal
sit
dine
dinner
eat
Stem
0 50 100 150 200
Count

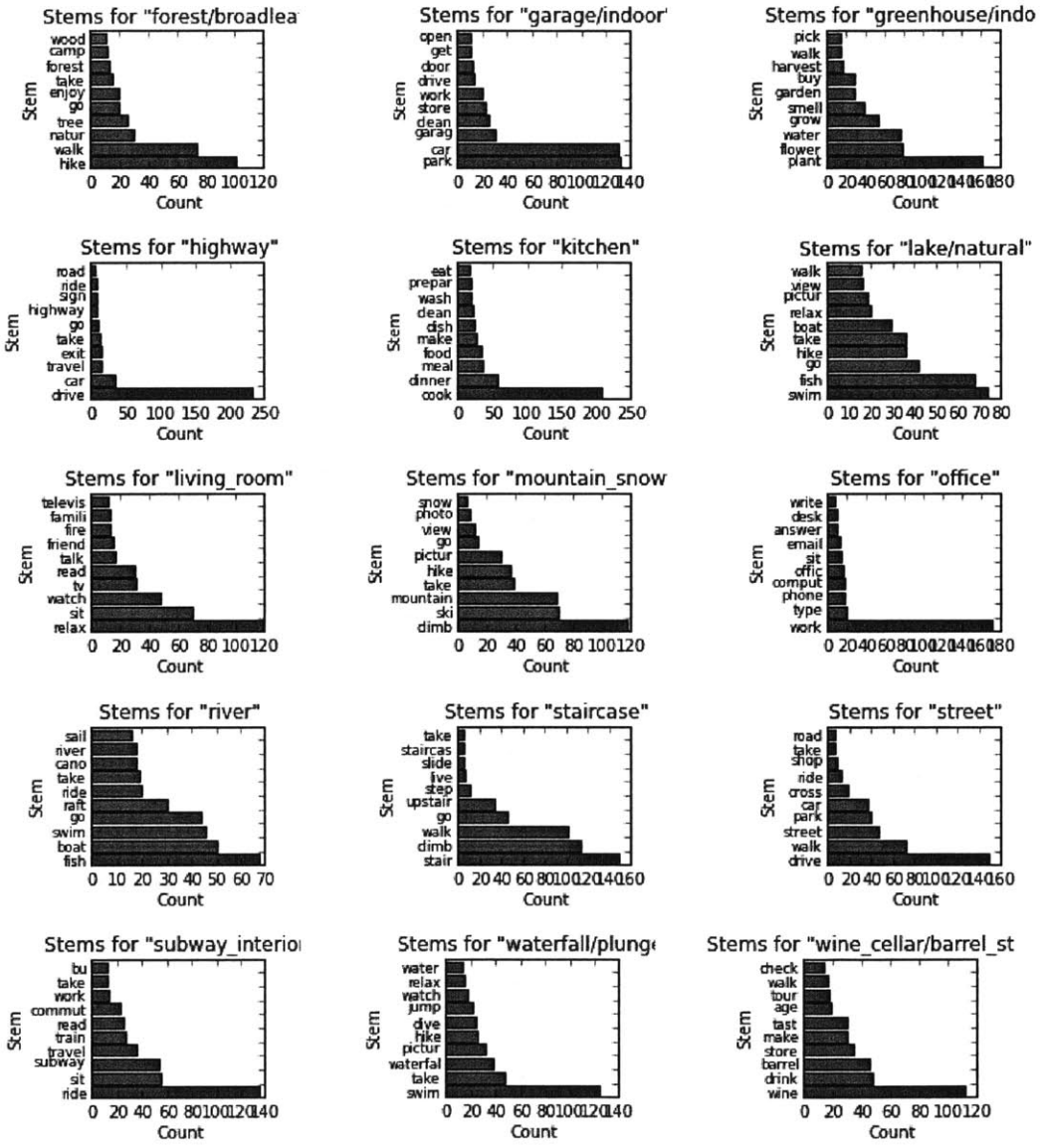


Figure A-1: Top 5 most common stems for every category in the 33-category subset

Appendix B

Mechanical Turk Best Practices

Data collection took place using Amazon Mechanical Turk, an online crowdsourcing platform through which "Requesters" can post small tasks (known as Human Intelligence Tasks, or HITs) for anonymous "Workers" to complete. Each HIT on Mechanical Turk is typically a few seconds to a few minutes long, and workers generally receive a few cents per HIT. Although it is possible to pay rates much below minimum wage, the community standard among workers is that fair payment should correspond to a rate of roughly 10 cents per minute, or $6.00/hour (mturk forum, 2013); the most experienced and reliable workers avoid tasks that pay less than this rate, and warn others to avoid them as well.

Most studies can be run using Amazon's built-in HIT design tools[1]. For more advanced functionality than the standard HIT design interface, the command-line interface offers more control in designing, running, and updating HITs.

Additional best practices uncovered in the course of this work are described here:

- The "best practices" guides contains valuable information on HIT design:
  - `http://mturkpublic.s3.amazonaws.com/docs/MTURK_BP.pdf`
  - `http://mechanicalturk.typepad.com/blog/tips-tricks/`
- High qualifications ensure trustworthy workers
  - At least 95% HIT approval rate
  - At least 100 HITs approved
  - Location is United States
- Running a small trial, with a text area at the bottom asking for comments, can provide useful feedback before running your full batch.
- Workers care about being paid enough, being paid promptly, and not getting any "rejections" (which are very damaging for their ability to continue to work)
  - Paying enough: aim for 10 cents per minute of work.

[1]Instructions are accessible at http://aws.amazon.com/documentation/mturk/

  - Paying promptly and minimizing rejections: even if they're working poorly, your Turkers still did the work for you and you should pay them. Also your IRB probably expects that you pay your subjects regardless of their performance. I recommend setting a short (1-hour) auto-approval timer.

- Building trust with your workers can establish a good reputation in the Turker community
  - Workers will rate requesters on websites such as "Turkopticon."
  - Workers appreciate getting responses to emails, especially if they are confused by your instructions, express concern that they will not be paid for their work, or raise doubts about whether you are paying fairly.
  - Turker forums (such as `http://mturkforum.com/`) can provide additional feedback. Consider creating an account and being available on the forums to answer questions.

References

Aas, Kjersti, & Eikvil, Line. 1999. *Text categorisation: A survey.* Norsk regnesentral.

Adelson, E. H. 2001. On seeing stuff: The perception of materials by humans and machines. *Page 112 of: Proceedings of the SPIE*, vol. 4299.

Barnard, Kobus, & Forsyth, David. 2001. Learning the semantics of words and pictures. *Pages 408–415 of: Computer Vision, 2001. ICCV 2001. Proceedings. Eighth IEEE International Conference on*, vol. 2.

Barnard, Kobus, Duygulu, Pinar, Forsyth, David, De Freitas, Nando, Blei, David M., & Jordan, Michael I. 2003. Matching words and pictures. *The Journal of Machine Learning Research*, **3**, 11071135.

Biederman, Irving. 1987. Recognition-by-components: A theory of human image understanding. *Psychological review*, **94**(2), 115147.

Blank, Moshe, Gorelick, Lena, Shechtman, Eli, Irani, Michal, & Basri, Ronen. 2005. Actions as space-time shapes. *Pages 1395–1402 of: Computer Vision, 2005. ICCV 2005. Tenth IEEE International Conference on*, vol. 2.

Blei, David M., & Jordan, Michael I. 2003. Modeling annotated data. *Page 127134 of: Proceedings of the 26th annual international ACM SIGIR conference on Research and development in informaion retrieval.*

Blei, David M., Ng, Andrew Y., & Jordan, Michael I. 2003. Latent dirichlet allocation. *The Journal of Machine Learning Research*, **3**(Mar.), 993–1022.

Choi, Myung Jin, Lim, Joseph J., Torralba, Antonio, & Willsky, Alan S. 2010. Exploiting hierarchical context on a large database of object categories. *Pages 129–136 of: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on.*

Chow, C., & Liu, C. 1968. Approximating discrete probability distributions with dependence trees. *Information Theory, IEEE Transactions on*, **14**(3), 462467.

Cover, Thomas M., & Thomas, Joy A. 1991. Entropy, relative entropy and mutual information. *Elements of Information Theory*, 1249.

Cover, Thomas M., & Thomas, Joy A. 2012. *Elements of information theory.* Wiley-interscience.

Ehinger, Krista A., Xiao, Jianxiong, Torralba, Antonio, & Oliva, Aude. 2011. Estimating scene typicality from human ratings and image features.

Epstein, Russell, & Kanwisher, Nancy. 1998. A cortical representation of the local visual environment. *Nature*, **392**(6676), 598–601.

Fabre-Thorpe, M., Delorme, A., Marlot, C., & Thorpe, S. 2001. A limit to the speed of processing in ultra-rapid visual categorization of novel natural scenes. *Journal of cognitive neuroscience*, **13**(2), 171180.

Farhadi, Ali, Hejrati, Mohsen, Sadeghi, Mohammad Amin, Young, Peter, Rashtchian, Cyrus, Hockenmaier, Julia, & Forsyth, David. 2010. Every Picture Tells a Story: Generating Sentences from Images. *Pages 15–29 of:* Daniilidis, Kostas, Maragos, Petros, & Paragios, Nikos (eds), *Computer Vision ECCV 2010.* Lecture Notes in Computer Science, no. 6314. Springer Berlin Heidelberg.

Fei-Fei, Li, Iyer, Asha, Koch, Christof, & Perona, Pietro. 2007. What do we perceive in a glance of a real-world scene? *Journal of Vision*, **7**(1).

Fowlkes, E. B., & Mallows, C. L. 1983. A Method for Comparing Two Hierarchical Clusterings. *Journal of the American Statistical Association*, **78**(383), 553–569. ArticleType: research-article / Full publication date: Sep., 1983 / Copyright 1983 American Statistical Association.

Gardenfors, Peter, & Williams, Mary-Anne. 2001. Reasoning about categories in conceptual spaces. *Pages 385–392 of: International Joint Conference on Artificial Intelligence*, vol. 17.

Good, I. J. 1953. The population frequencies of species and the estimation of population parameters. *Biometrika*, **40**, 237–264.

Greene, Michelle. 2013. *Statistics of High-level Scene Context.* unpublished.

Greenshtein, Eitan, & Park, Junyong. 2009. Application of non parametric empirical Bayes estimation to high dimensional classification. *The Journal of Machine Learning Research*, **10**, 16871704.

Hoiem, D., Efros, A.A., & Hebert, M. 2006. Putting Objects in Perspective. *Pages 2137–2144 of: 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, vol. 2.

Horton, John J., Rand, David G., & Zeckhauser, Richard J. 2011. The online laboratory: conducting experiments in a real labor market. *Experimental Economics*, **14**(3), 399–425.

Hubert, Lawrence, & Arabie, Phipps. 1985. Comparing partitions. *Journal of Classification*, **2**(1), 193–218.

Isola, Phillip, Parikh, Devi, Torralba, Antonio, & Oliva, Aude. 2011. *Understanding the intrinsic memorability of images.* Tech. rept. DTIC Document.

Kjellstrm, Hedvig, Romero, Javier, & Kragi, Danica. 2011. Visual object-action recognition: Inferring object affordances from human demonstration. *Computer Vision and Image Understanding*, **115**(1), 8190.

Lazebnik, Svetlana, Schmid, Cordelia, & Ponce, Jean. 2006. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. *Pages 2169–2178 of: Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on*, vol. 2.

Li, Li-Jia, & Fei-Fei, Li. 2007. What, where and who? classifying events by scene and object recognition. *Page 18 of: Computer Vision, 2007. ICCV 2007. IEEE 11th International Conference on.*

Liu, Jingen, Luo, Jiebo, & Shah, M. 2009. Recognizing realistic actions from videos "in the wild". *Pages 1996–2003 of: IEEE Conference on Computer Vision and Pattern Recognition, 2009. CVPR 2009.*

Loper, Edward, & Bird, Steven. 2002. NLTK: The natural language toolkit. *Page 6370 of: Proceedings of the ACL-02 Workshop on Effective tools and methodologies for teaching natural language processing and computational linguistics-Volume 1.*

Lowe, D.G. 1999. Object recognition from local scale-invariant features. *Pages 1150–1157 vol.2 of: The Proceedings of the Seventh IEEE International Conference on Computer Vision, 1999*, vol. 2.

Manning, Christopher D., Raghavan, Prabhakar, & Schtze, Hinrich. 2008. *Introduction to information retrieval.* Vol. 1. Cambridge University Press Cambridge.

Marszalek, M., Laptev, I., & Schmid, C. 2009. Actions in context. IEEE.

Meil, Marina. 2005. Comparing clusterings: an axiomatic view. *Page 577584 of: Proceedings of the 22nd international conference on Machine learning.*

mturk forum. 2013. *Data entry for event database.* http://mturkforum.com/showthread.php?6211-Data-entry-for-event-database.

Mundy, Joseph L. 2006. Object Recognition in the Geometric Era: A Retrospective. *Pages 3–28 of:* Ponce, Jean, Hebert, Martial, Schmid, Cordelia, & Zisserman, Andrew (eds), *Toward Category-Level Object Recognition.* Lecture Notes in Computer Science, no. 4170. Springer Berlin Heidelberg.

Mur, Marieke, Bandettini, Peter A., & Kriegeskorte, Nikolaus. 2009. Revealing representational content with pattern-information fMRIan introductory guide. *Social cognitive and affective neuroscience*, **4**(1), 101109.

Oliva, Aude, & Torralba, Antonio. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International journal of computer vision*, **42**(3), 145175.

Oliva, Aude, & Torralba, Antonio. 2006. Building the gist of a scene: The role of global image features in recognition. *Progress in brain research*, **155**, 23.

Oliva, Aude, & Torralba, Antonio. 2007. The role of context in object recognition. *Trends in cognitive sciences*, **11**(12), 520527.

Oliva, Aude, Mack, Michael L., Shrestha, Mochan, & Peeper, Angela. 2004. Identifying the perceptual dimensions of visual complexity of scenes. *In: Proc. of the 26th Annual Meeting of the Cogn. Sci. Soc.*

Ordonez, Vicente, Kulkarni, Girish, & Berg, Tamara L. 2011. Im2text: Describing images using 1 million captioned photographs. *In: Neural Information Processing Systems (NIPS).*

Paolacci, Gabriele, Chandler, Jesse, & Ipeirotis, Panagiotis. 2010. Running experiments on amazon mechanical turk. *Judgment and Decision Making*, **5**(5), 411419.

Patterson, Genevieve, & Hays, James. 2012. SUN attribute database: Discovering, annotating, and recognizing scene attributes. *Page 27512758 of: Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on.*

Pedersen, Ted, Patwardhan, Siddharth, & Michelizzi, Jason. 2004. WordNet::Similarity: measuring the relatedness of concepts. *Page 3841 of: Demonstration Papers at HLT-NAACL 2004.* HLT-NAACLDemonstrations '04. Stroudsburg, PA, USA: Association for Computational Linguistics.

Porter, Martin F. 1980. An algorithm for suffix stripping. *Program: electronic library and information systems*, **14**(3), 130137.

Princeton University. 2010. *About WordNet.* http://wordnet.princeton.edu.

Resnik, Philip. 1995. Disambiguating noun groupings with respect to WordNet senses. *Pages 54–68 of: Proceedings of the third workshop on very large corpora.*

Rish, Irina. 2001. An empirical study of the naive Bayes classifier. *Page 4146 of: IJCAI 2001 workshop on empirical methods in artificial intelligence*, vol. 3.

Rosch, Eleanor, & Lloyd, Barbara B. 1978. *Cognition and categorization.* Hillsdale, New Jersey.

Russell, Bryan C., Torralba, Antonio, Murphy, Kevin P., & Freeman, William T. 2008. LabelMe: A Database and Web-Based Tool for Image Annotation. *International Journal of Computer Vision*, **77**(1-3), 157–173.

Salton, G., Wong, A., & Yang, C. S. 1975. A vector space model for automatic indexing. *Commun. ACM*, **18**(11), 613–620.

Salton, Gerard, & Buckley, Christopher. 1988. Term-weighting approaches in automatic text retrieval. *Information Processing & Management*, **24**(5), 513–523.

Schuldt, Christian, Laptev, Ivan, & Caputo, Barbara. 2004. Recognizing human actions: a local SVM approach. *Page 3236 of: Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, vol. 3.

Stark, Michael, Lies, Philipp, Zillich, Michael, Wyatt, Jeremy, & Schiele, Bernt. 2008. Functional object class detection based on learned affordance cues. *Pages 435–444 of: Computer Vision Systems.* Springer.

The MathWorks, Inc. 2013. *Statistics Toolbox - MATLAB.* http://www.mathworks.com/products/statistics/.

Torralba, A., Murphy, K.P., Freeman, W.T., & Rubin, M.A. 2003. Context-based vision system for place and object recognition. *Pages 273–280 vol.1 of: Ninth IEEE International Conference on Computer Vision, 2003. Proceedings.*

Torralba, A., Fergus, R., & Freeman, W.T. 2008. 80 Million Tiny Images: A Large Data Set for Nonparametric Object and Scene Recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, **30**(11), 1958–1970.

Torralba, Antonio, Oliva, Aude, Castelhano, Monica S., & Henderson, John M. 2006. Contextual guidance of eye movements and attention in real-world scenes: The role of global features in object search. *Psychological review*, **113**(4), 766–786.

Tversky, Barbara, & Hemenway, Kathleen. 1983. Categories of environmental scenes. *Cognitive Psychology*, **15**(1), 121–149.

Warin, Martin. 2004. Using WordNet and Semantic Similarity to Disambiguate an Ontology. *Retrieved January*, **25**, 2008.

Xiao, Jianxiong, Hays, James, Ehinger, Krista A., Oliva, Aude, & Torralba, Antonio. 2010. Sun database: Large-scale scene recognition from abbey to zoo. *Pages 3485–3492 of: Computer vision and pattern recognition (CVPR), 2010 IEEE conference on.*

Zipf, G.K. 1935. *The psycho-biology of language.* Vol. ix. Oxford, England: Houghton, Mifflin.