THESIS

IMAGE FEATURE ASSOCIATIONS VIA LOCAL SEMANTIC STRUCTURE

Submitted by

Nicholas James Parrish

Department of Computer Science

In partial fulfillment of the requirements

for the Degree of Master of Science

Colorado State University

Fort Collins, Colorado

Fall 2010

COLORADO STATE UNIVERSITY

September 16, 2010

WE HEREBY RECOMMEND THAT THE THESIS PREPARED UNDER OUR SUPERVISION BY NICHOLAS JAMES PARRISH ENTITLED IMAGE FEATURE ASSOCIATIONS VIA LOCAL SEMANTIC STRUCTURE BE ACCEPTED AS FULFILLING IN PART REQUIREMENTS FOR THE DEGREE OF MASTER OF SCIENCE.

Committee on Graduate Work

J. Ross Beveridge

Lucy Troup

Adviser - Bruce Draper

Department Head - L. Darrell Whitley

ABSTRACT OF THESIS

IMAGE FEATURE ASSOCIATIONS VIA LOCAL SEMANTIC STRUCTURE

Research in the field of object recognition suffers from two distinct weaknesses that limits its effectiveness in natural environments. The first is that this research tends to rely on labeled training images, or other forms of supervision, to learn object models and recognize these models in novel images, thus preventing the learning of objects that are not labeled by humans. The second is that such systems tend to assume that the goal is to recognize a single, dominant foreground object.

This research implements a different method of object recognition that learns, without supervision, which object(s) are in natural scenes. This approach uses the semantic co-occurance information of local image features to form object models from groups of image features, which shall be called percepts. These percepts are then used to recognize objects in novel images. It will be shown that this approach is capable of learning object categories without supervision and recognition in complex multi-object scenes. It will also be shown that this approach out-performs a nearest-neighbor scene recognition approach.

Nicholas James Parrish
Department of Computer Science
Colorado State University
Fort Collins, CO 80523
Fall 2010

TABLE OF CONTENTS

1 Introduction **1**

2 Background **6**
2.1 Local Image Feature Background 6
2.1.1 Scene Recognition 7
2.1.2 Supervised Object Recognition 7
2.2 Algorithm Background 8
2.2.1 Term-document Matrix 8
2.2.2 Term Frequency Inverse Document Frequency (TFIDF) 9
2.2.3 Latent Semantic Analysis (LSA) 9
2.3 SeeAsYou System Background 11
2.3.1 Local Image Feature Detection 11
2.3.2 Local Image Feature Grouping 12
2.3.3 Familiarity 12
2.4 Data Set 14

3 Methods **15**
3.1 Local Percepts 16
3.2 Global Percepts 19

4 Results **20**
4.1 Evaluation of Percept Creation 20
4.2 Evaluation of Recognition 23

5 Conclusions **31**
5.1 Future Work 32

References **33**

LIST OF FIGURES

1.1 A complex natural scene that would be difficult to segment based on image information alone. Example image taken from the The Ponce Group's Object Recognition Database. 2
1.2 A complex natural scene with extracted local image features, shown with red targets. Example image taken from the The Ponce Group's Object Recognition Database. 3

2.1 Latent Semantic Analysis (LSA) decomposes the input term-document matrix using Singular Value Decomposition (SVD). 10
2.2 Examples from The Ponce Group's Object Recognition Database. 13

3.1 High level system architecture. The bottom path (*TFIDF Retrieval*) illustrates the system configuration before this work. The top path (*Local Percept Extraction* and *Percept Grouping and Retrieval*) shows the additional modules added by this work. 16
3.2 Example local percepts from a single probe. Note that each of the training images was either a stop sign without a sky background or a picture of sky without a stop sign. The system was able to separate these two concepts without any supervision or labeled training data. 17

4.1 Example clean and dirty percepts. A and B are each made up a single object, so they are clean. C is a mix of objects in the image and D is a mix of objects both in the image and not, so both are dirty. 21
4.2 Distribution of percepts after training on the Ponce Group's Object Recognition Database. Each bar represents the number of clean percepts found for a given category, except the last, which is the number of percepts that are not clean. 22
4.3 Example clean percepts generated from the The Ponce Group's Object Recognition Data Set training images. 24
4.4 Example dirty percepts generated from the The Ponce Group's Object Recognition Data Set training images. 25
4.5 An example of recognition results for a probe by both systems. Notice that both systems successfully retrieve examples of an object that was in the scene (the apple). 26

4.6 Another example of recognition results for a probe by both systems. Notice that both systems successfully retrieve examples the shoe, but that only one of them retrieves an additional percept for the apple. 26
4.7 A third example of recognition results for a probe by both systems. In this case, the TFIDF system successfully retrieves a set of bears and the Percepts system retrieves two dirty percepts, one of which contains bears. 27
4.8 A fourth example of recognition results for a probe by both systems. In this case, the TFIDF system successfully retrieves one of the objects in the image (the Rubble) and the Percepts system retrieves two percepts of the other object (the Spiderman). 27
4.9 A fifth example of recognition results for a probe by both systems. In this case, the TFIDF system retrieves images of a shoe, which is not in the image and the Percepts system retrieves three percepts (Apple, Shoe, and Salt), two of which are in the image. 28
4.10 Results of recognition using Percepts versus TFIDF on the Ponce Group's Object Recognition Database. 29
4.11 Example test images from the Ponce Group's Object Recognition Database that SeeAsYou has trouble matching. 29

Chapter 1

Introduction

One of the greatest challenges in the automatic interpretation of images is to break images into the objects they contain. Humans look at images and effortlessly recognize the different objects in them, yet many attempts in computer vision to automatically segment images into objects have largely failed [BSU04]. Humans are able to segment images into objects because of their experience with the world; they know what objects look like and how to differentiate one from another. Since this problem is so difficult for computers, many computer vision systems are either trained for specific object classes or avoid segmentation completely and attempt to match whole images to whole images.

Figure 1.1 shows a photograph of a complex natural scene. Any simple method of segmenting this image into objects will fail. The objects themselves are multi-colored and in many instances, the boundaries between objects have lower contrasts than their internal edges. Nevertheless, humans can see the teddy bear, the truck, the vase, the bowl, etc. and are not confused about how to parse the image into objects.

The goal of this research is to develop a system that can extract previously seen objects from complex scenes, such as Figure 1.1. This is not a novel goal, but most other systems that recognize objects in complex scenes use models learned from supervision, such as the work by Rothganger, et al on the data set in Figure 1.1 [RLSP06]. Many other vision systems learn to recognize objects under supervision, relying on hand-labeled

Figure 1.1: A complex natural scene that would be difficult to segment based on image information alone. Example image taken from the The Ponce Group's Object Recognition Database.

training images ([FPZ03], [AT06], [GD05], [ZYZS05], [OPFA06], etc). This thesis presents a system that learns to match individual objects across complex scenes *without* supervision.

In particular, this thesis investigates object recognition within the broader context of computer vision based on local mage features (a.k.a. interest points, keypoints, or focus of attention windows). Recognition through local image features has been studied since at least the mid 1980's [KU85], and has steadily gained in popularity since the publication of SIFT [Low04].

Local image features have many desirable properties. As small image patches, they are less susceptible to changes in viewpoint and illumination. They are more easily traced across images of a video sequence than other points in images ([HS88], [SL04]) and tend to repeat predictably across still frames of a single object [MTS$^+$05]. They may also be analogous to mechanisms in the human vision system.

Figure 1.2: A complex natural scene with extracted local image features, shown with red targets. Example image taken from the The Ponce Group's Object Recognition Database.

Local image features have the disadvantage, however, that they divide up the image into many small pieces. Figure 1.2 shows the same image as Figure 1.1, but this time with local image features extracted (shown as red circles with cross-hairs). These image features capture many (but not all of) the most interesting points in the image, but do not cluster on a single object. As a result, feature-based recognition creates a challenge in knowing which features should be part of a single object.

Feature based recognition has been successfully used in whole image matching systems, where the goal is to match the entire contents of one image to the entire contents of another. Such applications arise in the context of image retrieval [SZ08], and when labeling simple scenes with a single, dominant object ([AT06], [GD05], [ZYZS05], [OPFA06]), such as systems that use the Cal-Tech Database [FP98]. Whole scene matching is not always a satisfactory approach. Whole scene matching systems cannot determine which objects are shared by two images or whether they are matching

based on foreground or background similarities.

As a result, systems that aim to perform a more detailed analysis of complex scenes are forced to find some mechanism for grouping image features by object, and for separating the features of one object from the features of another. The goal of this project is to develop a feature-based computer vision system that learns to group image features into objects based on experience, but with no human supervision.

We define a percept as a group of local image features that, when bound together, represent an object viewed from a certain perspective. As such, percepts are viewpoint dependent representations of objects, similar to aspects [DPR92]. For example, a can of soda is an object, but a group of image features representing a view of the top of the can is a percept. This research proposes a semantics-based method for combining feature points into percepts. This research is different from other object recognition systems in that the system is completely unsupervised and at no point is given labeled images, and yet identifies percepts within images instead of matching whole images.

This work uses co-occurence semantic information under the hypothesis that local image features that tend to appear together are often parts of the same percept. To this end, it is possible to treat an image as a document and its local image features as words and apply text-based information retrieval techniques to identify co-occurences. Given a document, the first step is to retrieve similar images within the corpus. Similarity of two documents (images) is measured by counting feature overlaps between them and weighting them using Term Frequency Inverse Document Frequency (TFIDF) weights which favor uncommon words (features) in the corpus that are in the two. The system applies Latent Semantic Analysis (LSA) to this smaller set of documents by applying SVD to the TFIDF-weighted term-document matrix. The left singular vectors stored in the resulting term-concept matrix represent axes of covariance among terms of the retrieved documents. This thesis proposes that these axes can be used as instances of

percepts in single images.

Percepts are then clustered to form more general percepts across images where local image features have co-occured through time. Once a set of global percepts has been gathered, a new probe image's local features are projected onto each percept, and the percepts with the highest projections represent percepts that the system hypothesizes are in this probe image.

This resulting system will takes unlabeled training images as input, extracts local image features from them, and extracts local percepts from each frame. It then groups these percepts to determine global percepts that it has seen repeatedly. Finally, it uses the global percepts to match against the local image features of novel test images in order to recognize objects that the system has seen before. It will be shown that the unsupervised extraction and clustering of percept instances results in useful global percepts that clearly represent different objects. It will also be shown that using these percepts improves recognition over nearest neighbor whole image matching and has the additional advantage of being able to recognize multiple distinct percepts in a single image.

Chapter 2

Background

This project takes place in the context of the SeeAsYou vision system [Dra07], but the technique is applicable to a large set of vision systems that use local image features for recognition. This chapter first explores the background on local image features and their use in recognition, then describes the information retrieval algorithms that are used with these local image features, and finally, describes SeeAsYou as an example of a recognition system that uses local image features.

2.1 Local Image Feature Background

Local image features are regions of an image that are somehow important. As mentioned previously, local image features have desirable properties such as robustness to minor changes in viewpoint and illumination and their repeatability across still frames of a single object. These properties allow local image features to be very useful in spatial tasks such as tracking [YJS06, SL04], image alignment [Sze06], and 3D object modeling [RLSP06], as well as recognition tasks such as scene recognition and unsupervised object recognition - both of which closely relate to this work and will be discussed below. Many researchers have explored different methods for finding local image features in images including popular methods such as corner detection ([HS88]), Difference of Gaussians ([Low04]), and entropy measures([KB01]).

2.1.1 Scene Recognition

Scene recognition focuses on recognizing areas with similar environments or objects without requiring images to be from the same perspective within the scene. This can be illustrated with scenes of a movie where similar areas are reused at different times during the plot, but rarely from the exact same perspective. A scene can be described as a composition of local image features and scene matching can use this decomposition to match scenes with robustness to viewpoint and illumination changes [GD05], [SL04].

Many systems perform object recognition, but require the object to be the only object in, or the primary focus of, the image. For example, Fei-Fei, et al. present a system that learns object categories from a small set of labeled training images, as long as the object to be recognized takes up the majority of pixels in those images [FFFP07]. Any system that expects a single object per image is essentially performing scene recognition, since it uses both foreground and background information to make decisions. See [PBE+06] for further discussion. This research uses local image features for matching, but does not aim for whole scene recognition, but rather for subsets of the scene that represent objects or parts of objects that may occur in different environments and with different backgrounds.

2.1.2 Supervised Object Recognition

Object Recognition is a difficult problem that has been well studied, yet remains unsolved despite many advancements. The most common strategy for object recognition is through supervised training. In this strategy, a system is first presented with images that have labeled and possibly localized objects that it learns from and then attempts to formulate prototypes of these objects using the local image features that describe them ([DSL05]). Csurka, et al. use sets (”bags”) of features to describe a scenes and compares two supervised clustering algorithms, a support vector machine and a Bayesian neural

network, to match entire scenes [CDF+04]. Fergus, et al. claim to use an unsupervised approach and use a probability density function model generated from unlabeled training images. This approach is essentially supervised because all of the training images are from a single class of object and the goal is to decide whether or not that object is in subsequent test images that may or may not contain it [FPZ03].

These systems are presented to contrast them with the system implemented by this thesis, which takes a broader approach. Rather than learning from human supervision or labeled data, the system is simply presented with images and attempts to find parts of the image that are similar to parts it has seen before and to use semantic models of these parts (a.k.a. percepts) to recognize them in novel test images.

2.2 Algorithm Background

This thesis borrows heavily from the field of text-based information retrieval by treating images as documents and categorized local image features as terms. This will be described later in more detail, but this section presents relevant background from the field of information retrieval.

2.2.1 Term-document Matrix

A term-document matrix is a common representation used in information retrieval for a corpus of documents. It is essentially an index, or table, that stores occurrence information for each term of the corpus. It can be used to quickly determine which documents contain a given term and which terms are in a given document, both without performing a search of the corpus.

This matrix is common in natural language processing because it allows fast lookup of semantic information. It is common to weight the terms of this matrix, emphasizing terms that better discriminate among documents. A common technique for weighting

these terms is discussed below.

2.2.2 Term Frequency Inverse Document Frequency (TFIDF)

In information retrieval, the goal is often to match documents by comparing their terms. Some terms are more distinguishing and interesting than others, and are therefore more important for matching documents. Ideally, one of these terms would appear repeatedly in the matched documents, but rarely over the entire corpus. One technique to focus on using important terms is to assign each term a weight based on occurrence information in the corpus. One such weighting scheme is Term Frequency Inverse Document Frequency (TFIDF) [BYRN+99].

Given a term-document matrix, it is common to apply TFIDF weights to this matrix before using it for retrieval. TFIDF weights are calculated using the following formula:

$$weight_{t,d} = tf_{t,d} * log(\frac{N}{df_t})$$

The term frequency term, $tf_{t,d}$, is the number of occurrences of the term, t, within the document, d, divided by the length of d. It measures how important a term is in the current document. The document frequency term, df_t, is the number of documents in the corpus that contain t. The inverse document frequency term, $log(\frac{N}{df_t})$, measures how important the term is in the entire corpus. Thus, TFIDF favors terms that do not appear in very many documents and that are common in the current document.

2.2.3 Latent Semantic Analysis (LSA)

Another technique borrowed from the field of information retrieval is Latent Semantic Analysis (LSA). LSA extracts concepts from a term-document matrix by factoring it through applying singular value decomposition (SVD), similar to Principle Components Analysis (PCA) [BYRN+99]. The use of LSA is also motivated by the hypothesis that it simulates human cognition phenomena when it comes to recognition [DDF+90].

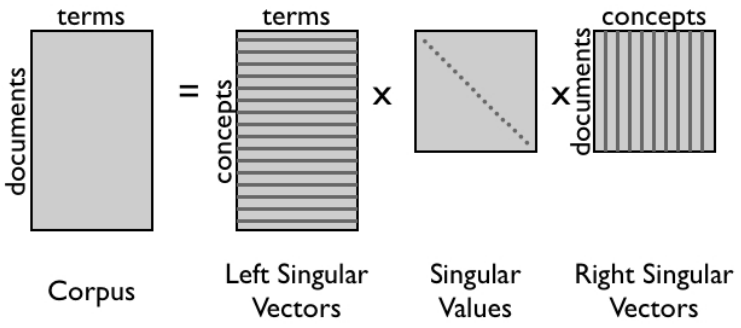

Figure 2.1: Latent Semantic Analysis (LSA) decomposes the input term-document matrix using Singular Value Decomposition (SVD).

This process extracts singular vectors from the term-document matrix, called concepts, and results in three output matrices: one orthonormal matrix relates terms to these concepts, another diagonal matrix contains the singular values, and the third is orthonormal and relates concepts to documents, see Figure 2.1. These concepts are often used to deal with the natural language problems of synonymy, where multiple terms share a meaning, and polysemy, where one term has multiple meanings. For an in-depth discussion on LSA, see [DDF$^+$90].

A variant on LSA limits the space to a subset of documents and terms that are related to a given document. This is called Local LSA. Local LSA focuses on finding local concepts that define the differences among similar images rather than global axis of variation in the entire corpus. One motivation for a localized LSA is computation time, since SVD is inherently an n^3 operation and is therefore only feasible for relatively small copra. Local LSA is not a new concept, it has been used by Wu in language translation to find similarities in small sets of documents and then apply them to a larger corpus, in order to reduce the computational complexity of LSA [WO08] and by Aseervatham to find useful similarities within k-means clusters as prototypes for object recognition [Ase08].

2.3 SeeAsYou System Background

This thesis extends the SeeAsYou vision system, but the approach could be used in similar recognition systems. SeeAsYou follows a common template for systems that use local image features: It begins by extracting local image features, then creates a descriptor for each one based on its local properties, and then matches them with other local image features from other images, through clustering or some other form of categorization, to achieve recognition. This same structure is found in systems from Csurka, et al. [CDF+04], Fegus, et al. [FPZ03], and Sivic, et al. [SRE+05] to name a few. Nonetheless, in order to put our new algorithm in context, we will briefly describe SeeAsYou in more detail. Readers already familiar with SeeAsYou may choose to skip to section 2.4.

SeeAsYou is a biomimetic vision system. The goal of biomimetic vision research is to develop vision systems that, through biological inspiration, implement aspects of the human visual system in order to accomplish visual tasks. It is hypothesized that by mimicking biology, computer vision systems will be more robust and adaptable to various tasks, rather than focused on a single problem domain. In order to mimic the human vision system, a level of abstraction must be chosen. SeeAsYou has chosen the level of Regional Functional Anatomy, which treats the brain as a set of regions separated by function. The primary goal of the SeeAsYou system is recognition through emulation of the ventral path of the human vision system. This path begins with local image feature detection, then feature extraction and clustering of these image features, which are then used to determine familiarity.

2.3.1 Local Image Feature Detection

The Saliency Module of SeeAsYou models covert spatial attention and is modeled after SIFT [Low04]. This is implemented using Difference of Gaussian convolution based on the theory that points which are on-center off-surround and vice-versa are interesting.

In order to detect salient regions at multiple scales, the image must be convolved in three dimensions, which is potentially expensive. To avoid this cost, SeeAsYou uses a pyramid of scales and finds optima in all three dimensions by comparing maxima in a pyramid level to the surrounding levels. Though SIFT was chosen, this project is applicable to other types of interest point operators.

2.3.2 Local Image Feature Grouping

In order to group local image features, a set of feature descriptors must be decided upon to give a relationship among points. In SIFT, Lowe uses a 128-dimension vector based on orientation histograms, but SeeAsYou instead creates descriptors in four different feature spaces (Hough, hue saturation value, edge orientation magnitude, and raw pixels) to be clustered independently. In each of these channels, SeeAsYou generates histograms using a channel-code voting mechanism [JF08], that can be used as feature vectors to be clustered.

The clustering algorithm used by SeeAsYou is based on the work of R. Granger, et. al. with Thalamo-Cortical Loops [RWG04]. It is hypothesized that these loops function as hierarchical clustering mechanisms in the human brain. SeeAsYou's implementation clusters local image features in each of the four different channels with a fixed hierarchy depth to allow both general and specific feature clusters. Each cluster is then assigned a unique id number to be used as a label. Since each local image feature is clustered in each channel separately, and in each channel there a varying levels in the cluster hierarchy, it may be given many different labels.

2.3.3 Familiarity

The familiarity module treats all of the labels from all of the local image features of an image as terms in a document. The module creates a term-document matrix from these labels and weights them using TFIDF weighting. In this matrix, each document is

(a) Train: Apple (b) Train: Bear (c) Train: Rubble (d) Train: Salt

(e) Train: Shoe (f) Train: Spidy (g) Train: Truck (h) Train: Vase

(i) Test (j) Test (k) Test (l) Test

Figure 2.2: Examples from The Ponce Group's Object Recognition Database.

equivalent to a high dimensional vector, where each term of the corpus is a dimension of the space. When this module receives a probe, which has already had local image features extracted and labeled, it scores each other document in the corpus as the sum of its TFIDF-weighted shared terms. Thus, the best match is an image that shares a lot of high scoring terms with the probe.

It should be noted that this familiarity module is never provided with labels of the images, and has no information regarding the number of potential object categories. This project expands this module to use a stronger unsupervised recognition system, but the system described here will be used for comparison so that both strategies receive the exact same input.

2.4 Data Set

To evaluate this research, SeeAsYou is presented with a data set from the Ponce Group at the University of Illinois at Urbana-Champaign. This data set presents labeled objects against very plain backgrounds for training and then presents these same objects, or combinations of them in different environments for testing. See Figure 2.2 for example training and test images.

This data set was designed for a different type of object recognition strategy that uses 3D models of objects to help with recognition [RLSP06]. The models were created by using labeled images of each class to find matching local image features and infer spatial relationships among the features. It will be shown that the results of this thesis are very promising even though the strategy presented does not rely on the labels and does not create 3D spatial models of objects.

Chapter 3

Methods

As described in the previous section, SeeAsYou uses unlabeled training images as its initial input. It begins by extracting local image features from these images, creating descriptors (a.k.a. feature vectors) from them and using these descriptors to cluster local image features into categories. For the purpose of this thesis, these earlier parts of SeeAsYou are treated as a ”black box” that delivers categorized local image features for each input image, and are represented by the *Local Image Feature Extraction & Grouping* box of Figure 3.1. Therefore, this portion of the system could be replaced by any system that performs similar tasks and that can provide the same output.

This research implements an addition to the SeeAsYou system that attempts to form percepts from local image features and then to recognize instances of these percepts in new images (*Local Percept Extraction* and *Percept Grouping and Retrieval* boxes of Figure 3.1). A percept has already been defined as a group of local image features that, when bound together, represent an object viewed from a certain perspective. As an example of percept extraction and motivation for their use, see Figure 3.2. Given the training images on the left, containing sky and stop sign images, and then probing the system with an image that contains both sky and a stop sign, two distinct percepts are extracted. Local percepts are found using local LSA centered on each image and are derived from the left singular vectors of the term-document matrix. Later, local per-

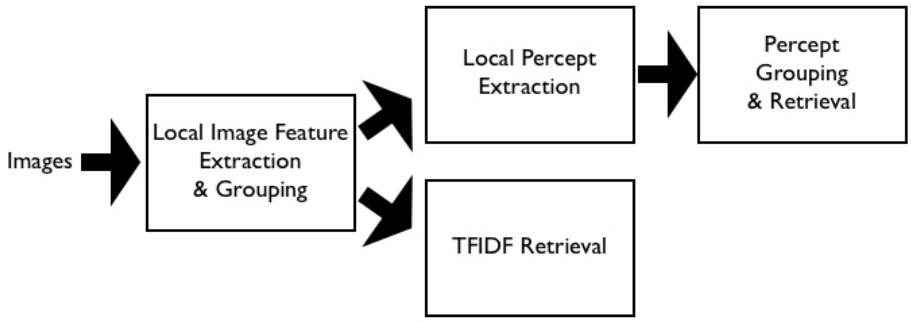

Figure 3.1: High level system architecture. The bottom path (*TFIDF Retrieval*) illustrates the system configuration before this work. The top path (*Local Percept Extraction* and *Percept Grouping and Retrieval*) shows the additional modules added by this work.

cepts from multiple images are clustered together to form more concrete global percepts which are used for recognition.

3.1 Local Percepts

After grouping local image features, the system effectively has a list of indices (numeric labels) for each image it processes. These indices can be treated as words in documents. Thus, a term-document matrix is formed and grows in one dimension as more images are perceived and in the other dimension as new categories are discovered.

Then, for each new input image, the system finds a set of images that is most similar to the probe using the TFIDF matching strategy described in Section 2.3.3 as the similarity measure. This local neighborhood is chosen because it is assumed that these images share features that can be exploited to derive percepts. If LSA was performed on the entire corpus, it would find axes of variance among all documents, which are not likely to correspond to single categories within the given image.

Now, given this neighborhood of similar images, a new term-document matrix is created that contains only these images that were retrieved by TFIDF and whose weights

Figure 3.2: Example local percepts from a single probe. Note that each of the training images was either a stop sign without a sky background or a picture of sky without a stop sign. The system was able to separate these two concepts without any supervision or labeled training data.

are those provided by applying the TFIDF formula to each vector in this local space. There are two motivations for using this local space rather than the entire matrix. The first is that the global term-document matrix is growing unbounded with the number of images that have been perceived which will inhibit scalability, but this local space is bounded by a constant - the number of images retrieved through the TFIDF step above, which is a parameter that can be set by the user. The second is that it is assumed that this local neighborhood will capture percepts that are focused on the current probe and images that are very similar to it.

SVD is then applied to this local space, effectively applying a local LSA to the data. This gives us a factorization of the local data matrix into three matrices: the term-concept matrix, the singular value diagonal matrix, and the concept by document matrix. The term-concept matrix relates the terms of the input matrix to a set of concepts

that correspond to each left singular vector of this space. The singular value matrix is a diagonal matrix that contains one singular value for each left singular vector and represent a measure of the magnitude of the axes of variance. The concept by document matrix relates the same concepts from the term-concept matrix to the set of documents from the input matrix. Using the first matrix, concepts can be represented as vectors of terms - these are used as percepts. As an example, given an image with object A and background B, it is assumed that since the TFIDF matching is essentially matching whole scenes, that some of its top matches will also contain object A and others will contain background B. These two sets of images are likely to contribute to two different singular vectors: one with high weights of object A's local image features and another with high weights of background B's local image features.

These percepts can be ranked within this image by their singular values, which give a measure of how much the data varies along the vector. Since a singular value describes the relative importance of a percept, it can also be used to ignore very low scoring percepts. This is similar to a strategy sometimes used in PCA where only the top eigenvectors are used to compress data. It is common to keep the first eigenvectors until the cumulative sum of their eigenvalues is about 90% of the total sum of the eigenvalues [Jol02]. This strategy is used in this system in order to ignore less important axes and therefore reduce noise.

Since percepts are vectors in a high dimensional space, they are difficult to visualize. In order to visualize a percept, SeeAsYou projects each of the other documents in the corpus onto the vector and displays some fixed number of the top projections. This gives the user an idea of what features the percept is using.

3.2 Global Percepts

Once a set of local percepts is formed for each image, the system attempts to merge them together with similar percepts from other images and to recognize these percepts in other images. If two images produce singular vectors that are similar measured in terms of the angle between them in the global term-document space, the system combines them using a weighted average in each dimension.

As percepts are combined, they emphasize similarities and de-emphasize differences, and we hypothesize that they will converge to find sets of local image features that represent percepts common to multiple images. These global percepts are visualized using the same strategy described above for local percepts by projecting images onto them to find the strongest representations of the percepts.

One problem with this approach is that each image will most likely have a different set of images selected for its local LSA than other images and therefore have singular vectors from different spaces, so comparing them is potentially difficult. To avoid this problem, each vector is represented using only its non-zero dimensions, and it assumed that it contains zero magnitude in all others. Thus, all of these percepts are treated as vectors in the union of all of the local spaces that are used, and are compared in this global space that contains all of them.

Chapter 4

Results

The goal of SeeAsYou is to form percepts from re-occuring objects or parts of objects so that they can be recognized in future novel stimuli. This is two separate tasks, percept learning and percept recognition, so each will be evaluated separately. We begin by evaluating the quality of the global percepts formed from the data described in section 2.4 and then we proceed to evaluate recognition performance when using the percepts on the testing data from the same set and compare these result to a TFIDF nearest neighbor approach.

4.1 Evaluation of Percept Creation

The first task of the system is to form percepts from images it has seen. The eight objects shown in Figure 2.2 are very different from each other, and we will evaluate the system based on how well it performed at creating percepts from unlabeled training samples of these objects. Evaluation of percepts is difficult because they are abstract vectors in a high dimensional space, so there is no intuitive measure for a ”good” or a ”bad” percept. Therefore, we define a *clean* percept as one whose top projections are all from the same object class. Ideally, this would be defined in terms of some threshold of magnitude along the percepts' axis, but since each axis is initially from a different local LSA space, these magnitudes cannot be compared directly. Instead, we have empirically decided

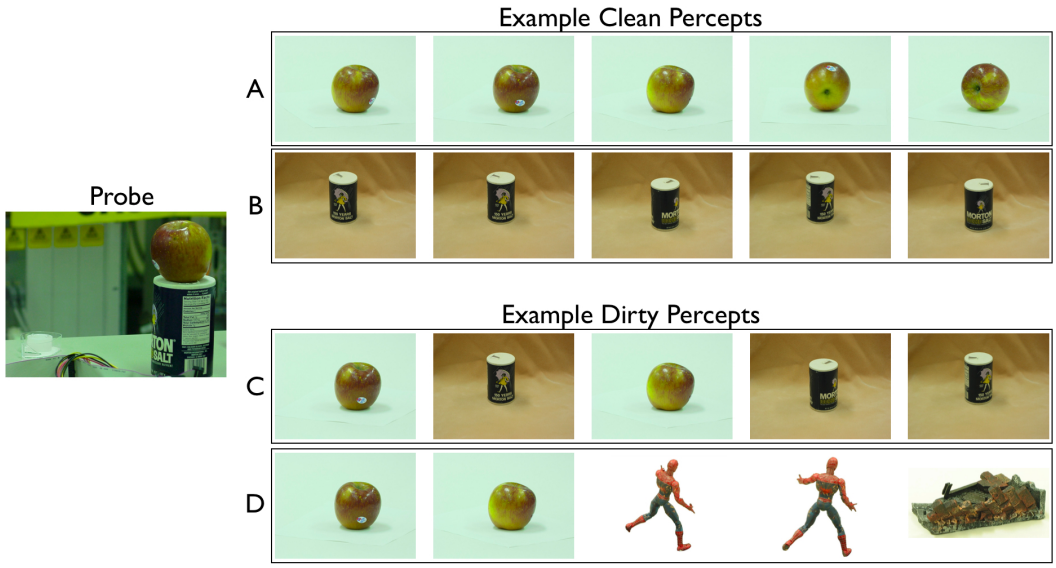

Figure 4.1: Example clean and dirty percepts. A and B are each made up a single object, so they are clean. C is a mix of objects in the image and D is a mix of objects both in the image and not, so both are dirty.

that a percept is clean if it's top 5 projections are from the same category. The number 5 was chosen because each category has at least 5 training samples and the probability of a random vector achieving this is $(\frac{1}{8})^5$ or approximately $3x10^{-5}$.

For the system to successfully recognize these eight objects, there should clean percepts for each object. There may be multiple percepts for each object, since they may represent different aspects of the object or may be using different descriptors from different channels or different hierarchy levels.

It turns out that approximately 64% of the percepts generated from the training data set were clean (representing a single object - the top 5 projections were all from the same class); 36% were mixed. The percepts that were clean were spread relatively evenly across the eight object categories, except for the truck and the apple, which had more training samples than the other categories and, as a result, had more clean percepts. See Figure 4.2 for more details on their distribution. See Figures 4.3 and 4.4 for examples of

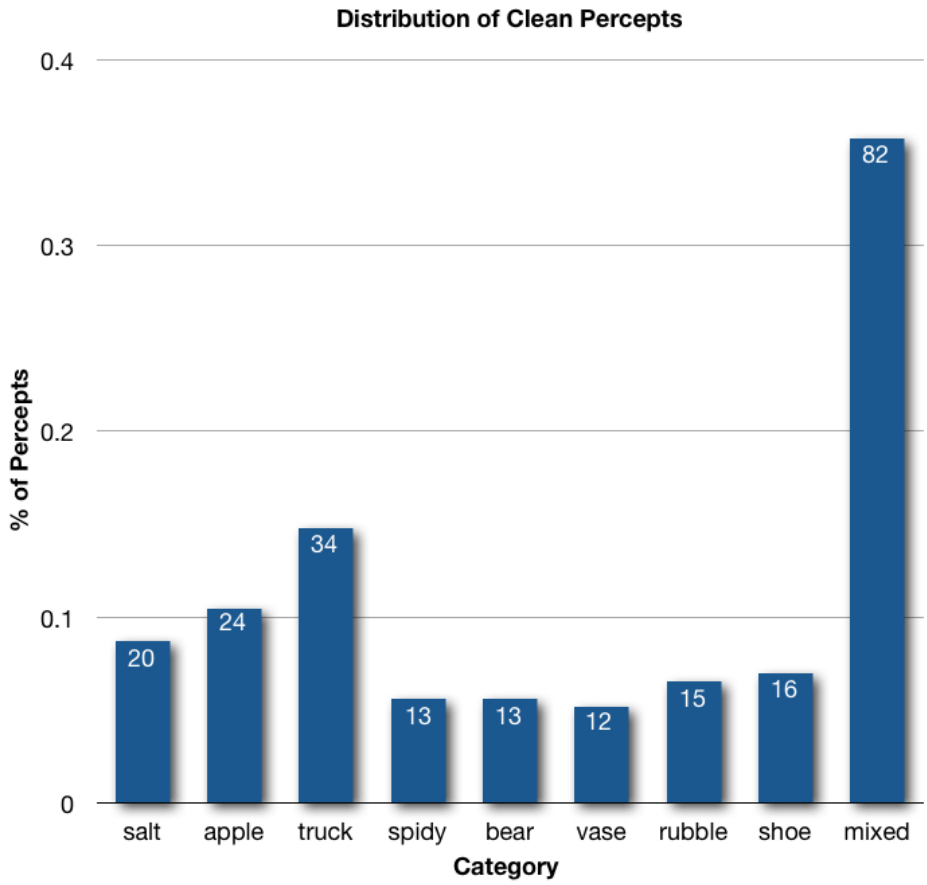

Figure 4.2: Distribution of percepts after training on the Ponce Group's Object Recognition Database. Each bar represents the number of clean percepts found for a given category, except the last, which is the number of percepts that are not clean.

both clean and dirty percepts formed during the same experiment. It should be noted that the mixed percepts are not necessarily bad, since they are products of other similarities among images in terms of color, shape, texture, etc. It is expected that the system will create percepts that do not correspond to human-defined objects since there are many similarities among images beyond which object(s) are in them.

4.2 Evaluation of Recognition

The next step is to evaluate recognition using these percepts. One issue with this evaluation is that this percept-based system is dependent on the other modules of SeeAsYou: the attention system may not choose useful attention windows, the categorizer might incorrectly group different features, etc. Since these percepts depend heavily of other parts of the SeeAsYou system, we will evaluate performance when compared to using just TFIDF nearest-neighbor matching. Both percepts and the TFIDF matching receive the exact same input, so this is now an even playing field. This comparison is also justified by the popularity of TFIDF in the information retrieval community, and its use in other object recognition systems [SZ08].

We will evaluate both systems using performance on image retrieval in the same category as the probe - even though neither system received labels with the training data. We will compare the two systems by comparing their rates of true positives and rates of false positives. For TFIDF, we use the the best match according to similarity score and if that score is above the threshold, it is considered a positive prediction, which can then be either true or false. For percepts, we use the percept onto which the probe projects the highest, and if it is a clean percept, we look at the category of the percept and use it as a prediction. For both algorithms, the prediction generated is a true positive if the prediction was correct (in the same class) and a false positive if it was not.

Figure 4.2 shows the true positive rate compared to the false positive rate for this

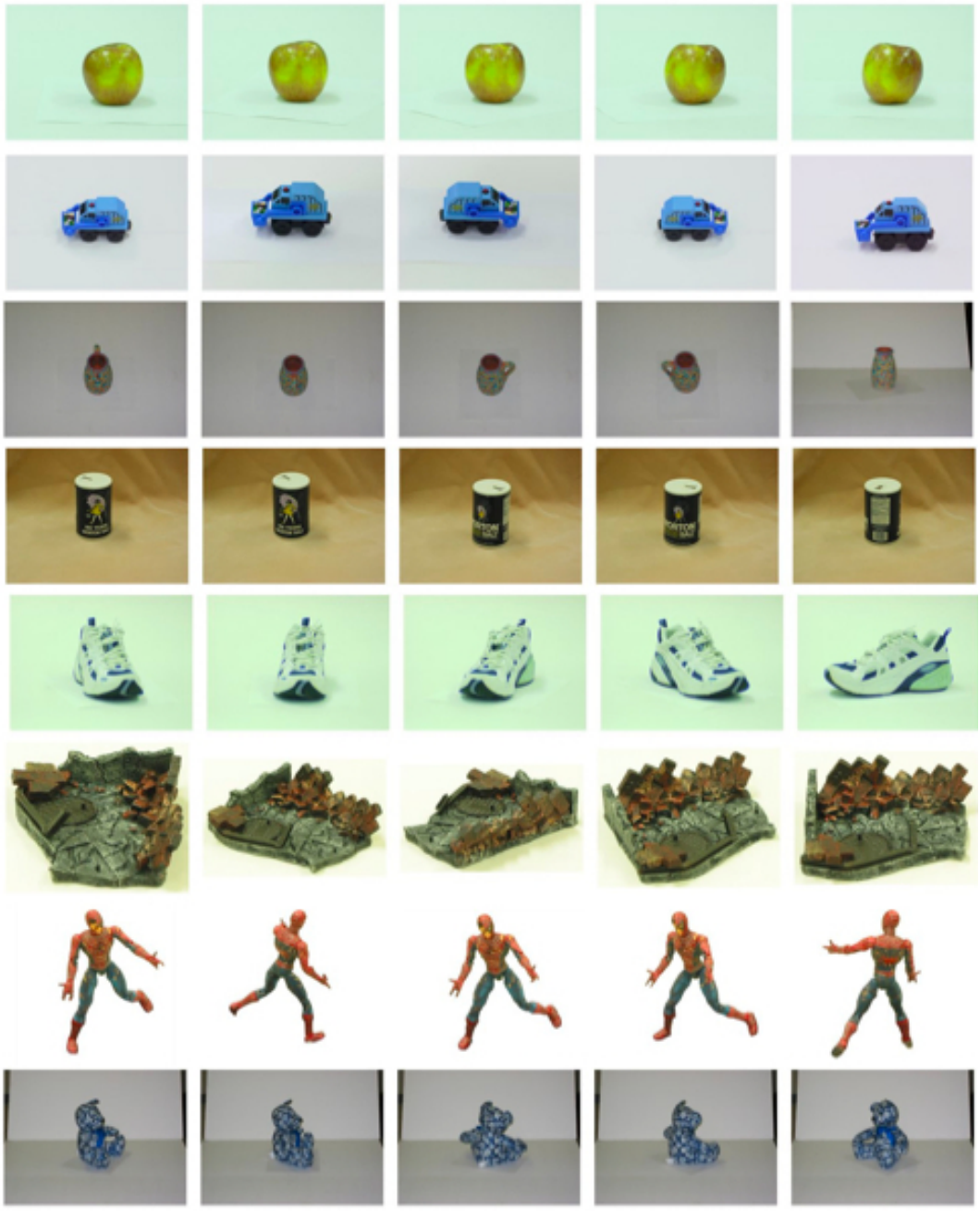

Figure 4.3: Example clean percepts generated from the The Ponce Group's Object Recognition Data Set training images.

Figure 4.4: Example dirty percepts generated from the The Ponce Group's Object Recognition Data Set training images.

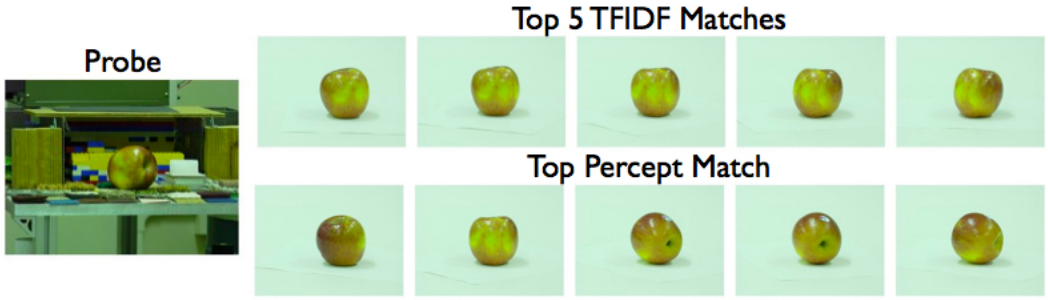

Figure 4.5: An example of recognition results for a probe by both systems. Notice that both systems successfully retrieve examples of an object that was in the scene (the apple).

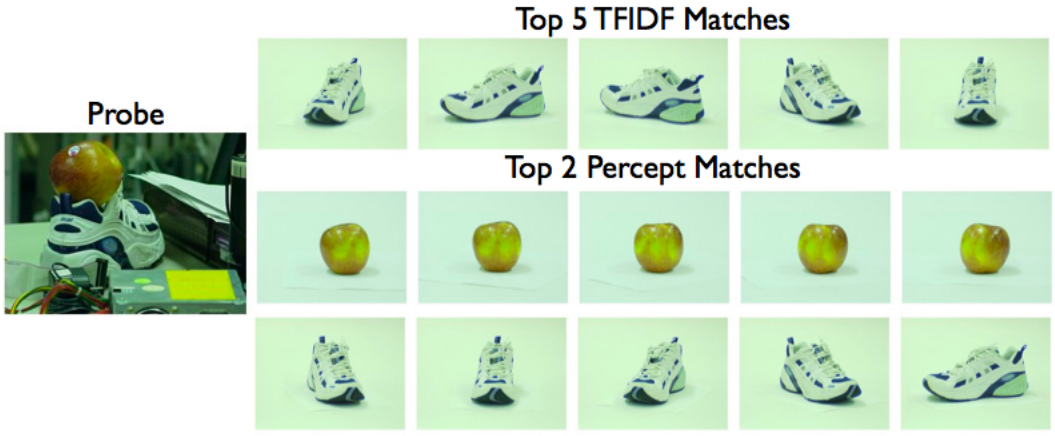

Figure 4.6: Another example of recognition results for a probe by both systems. Notice that both systems successfully retrieve examples the shoe, but that only one of them retrieves an additional percept for the apple.

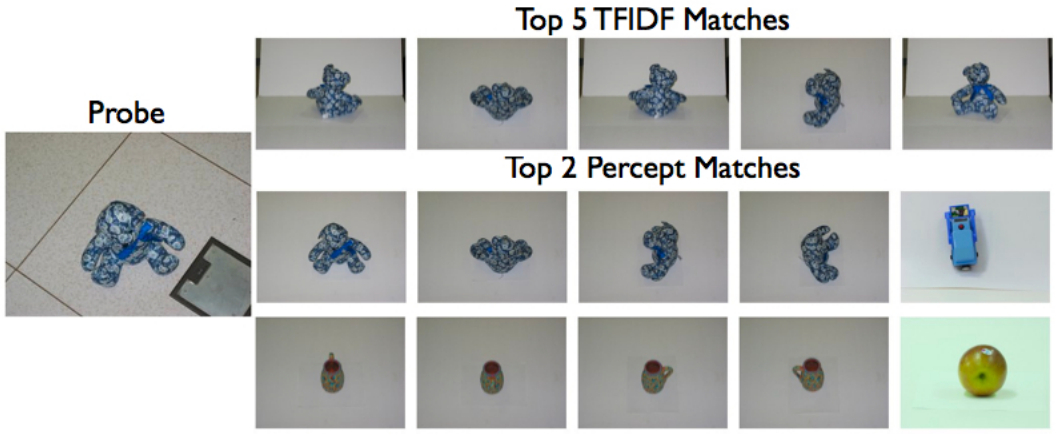

Figure 4.7: A third example of recognition results for a probe by both systems. In this case, the TFIDF system successfully retrieves a set of bears and the Percepts system retrieves two dirty percepts, one of which contains bears.

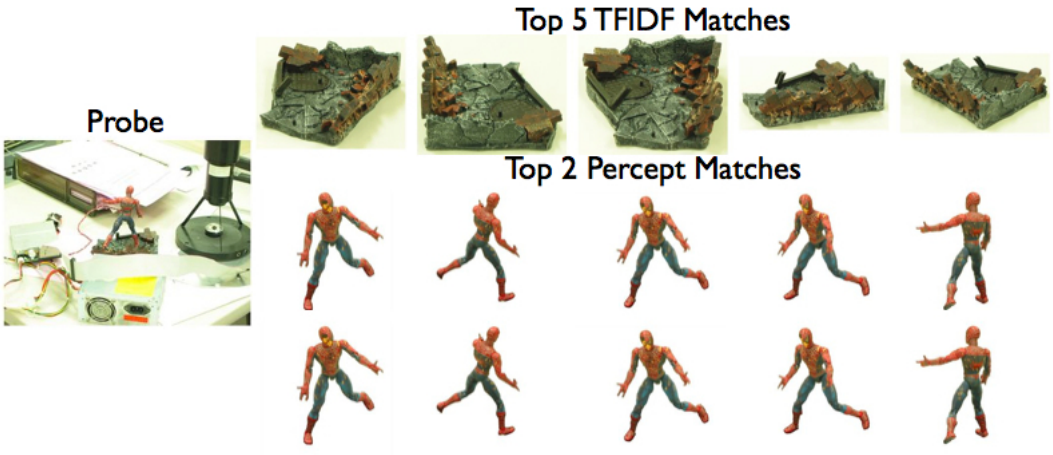

Figure 4.8: A fourth example of recognition results for a probe by both systems. In this case, the TFIDF system successfully retrieves one of the objects in the image (the Rubble) and the Percepts system retrieves two percepts of the other object (the Spiderman).

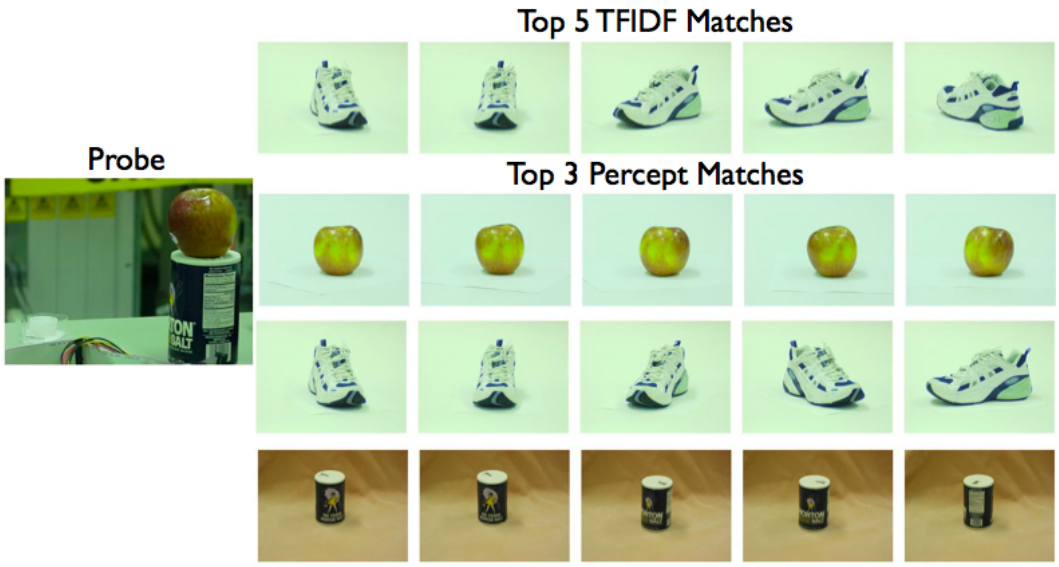

Figure 4.9: A fifth example of recognition results for a probe by both systems. In this case, the TFIDF system retrieves images of a shoe, which is not in the image and the Percepts system retrieves three percepts (Apple, Shoe, and Salt), two of which are in the image.

percepts approach and the TFIDF approach. Each point on the curve is created by modifying a cut-off parameter of the system. For the TFIDF approach, this is the minimum TFIDF score that is accepted to claim that an object is in an image, and for the percepts approach, this is the minimum projection length of an image onto that percept for the system to say the object is in that image. It is expected that the area under the ROC curve of a system that made random predictions would be 0.50. For the TFIDF approach, the area under the ROC curve is 0.759, and for percept-based recognition it is 0.920. It is clear that both approaches successfully recognize object instances, but that the percepts approach recognizes correctly more often. To understand the ROC curve better, let the reader look at the two curves at a false positive rate of 0.2. A false positive rate of 0.2 means that that when the system is presented with an scene, it claims there is an object that is not present 20% of the time. At this false positive rate, the TFIDF approach achieves a true positive rate of 0.5 while the Percepts approach achieves 0.9. The true

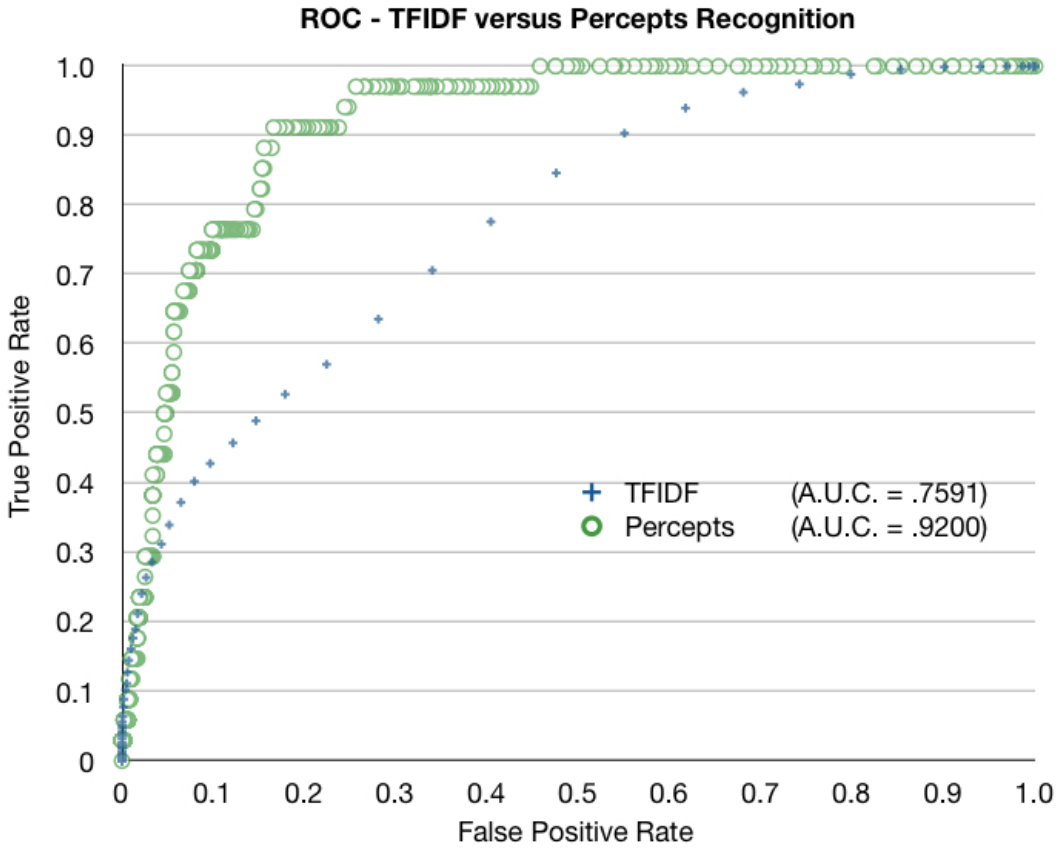

Figure 4.10: Results of recognition using Percepts versus TFIDF on the Ponce Group's Object Recognition Database.

Figure 4.11: Example test images from the Ponce Group's Object Recognition Database that SeeAsYou has trouble matching.

positive rate shows how often the system claims an object is in the image and it is. Different applications have different requirements for true positive and false positive rates, so the trade off between them can be very important. This ROC curve shows that for a given false positive rate, the true positive rate of the percepts approach is often better than the TFIDF approach by 20 to 40 percent.

Chapter 5

Conclusions

This thesis presented a system with the goal of recognizing objects or parts of objects that it has seen before. Its input was unlabeled training images and it formed percepts from re-occuring sets of image features. This system did not use any supervised training or labeled training images. This approach is different from much of the object recognition community which uses labels to find similarities within classes to form models and uses them to find other instances of the class in novel test images.

From the results, it is clear that this new approach is very promising. It consistently created clean percept models for the objects it was presented and out-performed TFIDF whole image matching, which uses the exact same input. Another interesting note is that percepts were intended to be view-point dependent, but many of the percepts discovered in this data set were able to recognize objects from different perspectives. This is most likely due to the combination of local percepts into more general global percepts that combined these viewpoint dependent aspects.

One potential application of this work is image browsing software where a user can interactively find images that are similar to a given image. For example, the user could present an image of a car in a desert setting next to a camel and would be presented with one set of images of vehicles, a set of desert images, and a set of camel images. This would be a novel way to search for images that are not easily described with words or

where image meta data is unavailable.

5.1 Future Work

Immediate future work in this area will include applying a more advanced grouping mechanism to the global percepts, perhaps utilizing Granger's algorithm that is used earlier in SeeAsYou. Since this research simply finds percepts from semantic information alone, other future work in this area might supplement this approach with spatial information, segmentation, or common motion information. Spatial information can be very useful for supporting or discounting certain combinations of local image features. Segmentation could give another source of information that relates local image features together by using the texture information of the local image features. Common motion is useful in a video domain where local image features that move together, independently from the background, are more likely to be parts of the same object. No one of these approaches will be perfect for identifying and recognizing objects, but their combination could be very powerful.

REFERENCES

[Ase08] S. Aseervatham. A local Latent Semantic Analysis-based kernel for document similarities. In *IEEE International Joint Conference on Neural Networks, 2008. IJCNN 2008.(IEEE World Congress on Computational Intelligence)*, pages 214–219, 2008.

[AT06] A. Agarwal and B. Triggs. Hyperfeatures-multilevel local coding for visual recognition. *Lecture Notes in Computer Science*, 3951:30, 2006.

[BSU04] E. Borenstein, E. Sharon, and S. Ullman. Combining top-down and bottom-up segmentation. In *Computer Vision and Pattern Recognition Workshop, 2004 Conference on*, pages 46–46, 2004.

[BYRN+99] R. Baeza-Yates, B. Ribeiro-Neto, et al. *Modern information retrieval*. ACM press New York, 1999.

[CDF+04] G. Csurka, C. Dance, L. Fan, J. Willamowski, and C. Bray. Visual categorization with bags of keypoints. In *Workshop on Statistical Learning in Computer Vision, ECCV*, volume 2004, 2004.

[DDF+90] S. Deerwester, S.T. Dumais, G.W. Furnas, T.K. Landauer, and R. Harshman. Indexing by latent semantic analysis. *Journal of the American society for information science*, 41(6):391–407, 1990.

[DPR92] SJ Dickinson, AP Pentland, and A. Rosenfeld. 3-D shape recovery using distributed aspect matching. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 14(2):174–198, 1992.

[Dra07] B.A. Draper. A Biomimetic Vision Architecture. In *IEEE International Conference on Computer Vision Systems, 2007 ICVS'07*, Bielefeld, Germany, 2007.

[DSL05] Gyuri Dork, , Cordelia Schmid, and Projet Lear. Object class recognition using discriminative local features. Technical report, IEEE Transactions on Pattern Analysis and Machine Intelligence, 2005.

[FFFP07] L. Fei-Fei, R. Fergus, and P. Perona. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Computer Vision and Image Understanding*, 106(1):59–70, 2007.

[FP98] R. Fergus and P. Perona. The Caltech database, 1998.

[FPZ03] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *2003 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2003. Proceedings*, volume 2, 2003.

[GD05] Kristen Grauman and Trevor Darrell. Efficient image matching with distributions of local invariant features. In *CVPR '05: Proceedings of the 2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'05) - Volume 2*, pages 627–634, Washington, DC, USA, 2005. IEEE Computer Society.

[HS88] C. Harris and M. Stephens. A combined corner and edge detector. pages 147–151, Manchester, UK, USA, 1988.

[JF08] E. Jonsson and M. Felsberg. Efficient computation of channel-coded feature maps through piecewise polynomials. *Image and Vision Computing*, 2008.

[Jol02] IT Jolliffe. *Principal component analysis*. Springer, 2002.

[KB01] Timor Kadir and Michael Brady. Scale, saliency and image description. *IJCV*, 45(2):83–105, November 2001.

[KU85] C. Koch and S. Ullman. Shifts in selective visual attention: towards the underlying neural circuitry. *Human neurobiology*, 4(4):219–227, 1985.

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

[MTS+05] K. Mikolajczyk, T. Tuytelaars, C. Schmid, A. Zisserman, J. Matas, F. Schaffalitzky, T. Kadir, and L.V. Gool. A comparison of affine region detectors. *International Journal of Computer Vision*, 65(1):43–72, 2005.

[OPFA06] A. Opelt, A. Pinz, M. Fussenegger, and P. Auer. Generic object recognition with boosting. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 28(3):416–431, 2006.

[PBE+06] J. Ponce, TL Berg, M. Everingham, DA Forsyth, M. Hebert, S. Lazebnik, M. Marszalek, C. Schmid, BC Russell, A. Torralba, et al. Dataset issues in object recognition. *Lecture Notes in Computer Science*, 4170:29, 2006.

[RLSP06] Fred Rothganger, Svetlana Lazebnik, Cordelia Schmid, and Jean Ponce. 3d object modeling and recognition using local affine-invariant image descriptors and multi-view spatial constraints. *Int. J. Comput. Vision*, 66(3):231–259, 2006.

[RWG04] A. Rodriguez, J. Whitson, and R. Granger. Derivation and analysis of basic computational operations of thalamocortical circuits. *J. Cognitive Neuroscience*, 16(5):856–877, 2004.

[SL04] Iryna Skrypnyk and David G. Lowe. Scene modelling, recognition and tracking with invariant image features. In *ISMAR '04: Proceedings of the 3rd IEEE/ACM International Symposium on Mixed and Augmented Reality*, pages 110–119, Washington, DC, USA, 2004. IEEE Computer Society.

[SRE+05] J. Sivic, BC Russell, AA Efros, A. Zisserman, and WT Freeman. Discovering objects and their location in images. In *Tenth IEEE International Conference on Computer Vision, 2005. ICCV 2005*, volume 1, 2005.

[SZ08] J. Sivic and A. Zisserman. Efficient Visual Search of Videos Cast as Text Retrieval. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 23, 2008.

[Sze06] Richard Szeliski. Image alignment and stitching: a tutorial. *Found. Trends. Comput. Graph. Vis.*, 2(1):1–104, 2006.

[WO08] Y. Wu and D.W. Oard. Bilingual topic aspect classification with a few training examples. In *Proceedings of the 31st annual international ACM SIGIR conference on Research and development in information retrieval*, pages 203–210. ACM New York, NY, USA, 2008.

[YJS06] Alper Yilmaz, Omar Javed, and Mubarak Shah. Object tracking: A survey. *ACM Comput. Surv.*, 38(4):13, 2006.

[ZYZS05] W. Zhang, B. Yu, GJ Zelinsky, and D. Samaras. Object class recognition using multiple layer boosting with heterogeneous features. In *IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2005. CVPR 2005*, volume 2, 2005.