

MODELING SYNERGISTIC RELATIONSHIPS BETWEEN WORDS AND IMAGES

Chee Wee Leong, B.Eng., M.Sc.

Dissertation Prepared for the Degree of
DOCTOR OF PHILOSOPHY

UNIVERSITY OF NORTH TEXAS

December 2012

APPROVED:

Rada Mihalcea, Major Professor
Paul Tarau, Committee Member
Miguel Ruiz, Committee Member
Ted Pedersen, Committee Member
Barrett Bryant, Chair of the Department of
Computer Science and Engineering
Costas Tsatsoulis, Dean of the College of
Engineering
Mark Wardell, Dean of the Toulouse Graduate
School

Leong, Chee Wee. Modeling Synergistic Relationships between Words and Images. Doctor of Philosophy (Computer Science and Engineering), December 2012, 117 pp., 24 tables, 12 illustrations, 90 numbered references.

Texts and images provide alternative, yet orthogonal views of the same underlying cognitive concept. By uncovering synergistic, semantic relationships that exist between words and images, I am working to develop novel techniques that can help improve tasks in natural language processing, as well as effective models for text-to-image synthesis, image retrieval, and automatic image annotation. Specifically, in my dissertation, I explore the interoperability of features between language and vision tasks. In the first part, I show how it is possible to apply features generated using evidence gathered from text corpora to solve the image annotation problem in computer vision, without the use of any visual information. In the second part, I address research in the reverse direction, and show how visual cues can be used to improve tasks in natural language processing. Importantly, I propose a novel metric to estimate the similarity of words by comparing the visual similarity of concepts invoked by these words, and show that it can be used further to advance the state-of-the-art methods that employ corpus-based and knowledge-based semantic similarity measures. Finally, I attempt to construct a joint semantic space connecting words with images, and synthesize an evaluation framework to quantify cross-modal semantic relationships that exist between arbitrary pairs of words and images. I study the effectiveness of unsupervised, corpus-based approaches to automatically derive the semantic relatedness between words and images, and perform empirical evaluations by measuring its correlation with human annotators.

Copyright 2012
by
Chee Wee Leong

ACKNOWLEDGEMENTS

Growth requires time, patience and many opportunities. I am deeply indebted to my research supervisor, Rada Mihalcea, for grooming me from a student who had never heard of Natural Language Processing, to an independent researcher who can contribute to original research in NLP. The journey would be tough, but she remained a strong pillar of support throughout. Particularly, I am grateful to her for providing us with financial support all these years, so that I may continue to pursue my ambition while caring for the needs of my family. On almost every issue, she has been encouraging, optimistic, and exceedingly tolerant of my shortcomings. Her unique, open-minded supervision style and willingness to engage in conversation on just about anything is rare among supervisors, but also welcoming to students in our lab. Indeed, she believed in my abilities and brought out the best in me.

I would like to thank Paul Tarau, Ruiz Miguel and Ted Pedersen, who accepted the invitation to sit on the committee, and provided insightful comments during my proposal defense. With their critique, I was able to see the loopholes in my work and patched them, which ultimately led to a much improved dissertation.

My wife accompanied me in this journey willingly, and out of love. She is the reason why I started it so earnestly, and concluded it without regrets. We are also pleased that our newborn daughter jumped on the bandwagon in the last minute, and made the ending all the more stressful, yet exciting and much memorable.

Though thousands of miles away, my family also supported me through consistent prayers and in spirit. To my mum especially, who used nearly all her life savings just to finance my first year of graduate studies before I receive any scholarship and fellowship. I hope to convince myself that she made the right bet, and that I would continue to grow in stature, strength and wisdom.

Finally, thanks and glory be to God, for You and You alone made all these possible.

TABLE OF CONTENTS

	Page
LIST OF TABLES	vi
LIST OF FIGURES	ix
CHAPTER 1. “MEN ARE FROM MARS, WOMEN ARE FROM VENUS”: INTRODUCTION TO LANGUAGE AND VISION.....	1
CHAPTER 2. “ARE MEN REALLY FROM MARS, AND WOMEN REALLY FROM VENUS ?”: SYNERGISM BETWEEN LANGUAGE AND VISION	6
2.1 Cognitive Science, Psychology and Neuroscience	6
2.2 Computer Vision and Image Processing.....	8
2.3 Natural Language Processing and Computational Linguistics	10
2.4 Visual Languages, Alternative and Augmentative Communication and Education	11
CHAPTER 3. BACKGROUND ON RESOURCES	13
3.1 Wordnet.....	13
3.2 ImageNet.....	16
3.3 Bag of Visual Codewords using Shift-Invariant Feature Transform	18
CHAPTER 4. “MEN CAN BE FROM VENUS, WOMEN CAN BE FROM MARS”: INTEROPERABILITY OF FEATURES BETWEEN LANGUAGE AND VISION TASKS (I) – TEXT MINING FOR AUTOMATIC IMAGE TAGGING	21
4.1 Motivation.....	21
4.2 Dataset.....	23
4.3 A New Evaluation Framework: Image Tagging as Lexical Substitution	24
4.4 Extractive Image Annotation	27
4.5 Supervised Learning	33
4.6 Experiments and Evaluations.....	34
4.7 Related Work	35

CHAPTER 5. “MEN CAN BE FROM VENUS, WOMEN CAN BE FROM MARS”: INTEROPERABILITY OF FEATURES BETWEEN LANGUAGE AND VISION TASKS (II) – AN IMAGE-BASED APPROACH FOR MEASURING WORD RELATEDNESS.....	39
5.1 Motivation.....	39
5.2 Dataset.....	40
5.3 Experiments	41
5.4 Discussion	44
5.5 Further Investigation.....	46
5.6 Related Work	58
CHAPTER 6. “MEN ARE FROM MARS AND VENUS, SO ARE WOMEN”: BUILDING A MULTIMODAL SEMANTIC SPACE USING WORDS AND IMAGES.....	60
6.1 Motivation.....	60
6.2 Semantic Vector Models.....	62
6.3 Semantic Relatedness between Words and Images	63
6.4 Dataset.....	65
6.5 Experiments	68
6.6 Discussion	70
6.7 Extended Study: Using Image Attributes for Measuring Cross-Modal Semantic Relatedness.....	74
6.8 Related Work	82
CHAPTER 7. CONCLUSION.....	84
7.1 Research Questions Revisited.....	84
APPENDIX A. EVALUATION DATASETS	86
APPENDIX B. PICTURABILITY SCORES.....	99
BIBLIOGRAPHY	112

LIST OF TABLES

3.1	Words and synsets in Wordnet 3.0.	14
4.1	Two sample images. The number besides each label indicates the number of human annotators agreeing on that label. Note that the mode image has a tag (i.e. “train”) in the gold standard set most frequently selected by the annotators	23
4.2	Example of tags provided by 5 independent annotators for an image depicting the dog “Chihuahua”	27
4.3	Candidate labels obtained for a sample text using the Flickr model	30
4.4	Directional similarity scores for some words	32
4.5	Results obtained on the Web dataset	35
4.6	Results obtained on the BBC dataset used in [26]	36
5.1	Pairwise combination functions of outputs from any two metrics M_1 and M_2 . We set $\beta = 0.5, 1, 1.5$ for three further variants of the F-measure function.	44
5.2	Table showing Spearman correlation of similarity scores generated using different metrics with human judgements, repeated for each of the 6 trimmed datasets	46
5.3	Results obtained with individual knowledge-based and corpus-based text-based measures, with our image measure, and with two combination functions (COMBSUM and F, with 3 variants). The bold correlation numbers represents the highest among all metrics per text-based measure per dataset.	46
5.4	Table showing the word pairs coverage statistics and synset pairings per word pair for each of the three standard evaluation datasets. Each number reported is cumulative i.e. +glosses is the expansion method using glosses + synset	49

5.5	LexExpGloss : Table showing the word pairs coverage statistics and synset pairings per word pair for WS353 dataset, plotted against the percentage of dropped words (reverse sorted using <i>tfidf</i>). The expanded bag of words is formed by synsets + glosses only	51
5.6	LexExpAll : Table showing the word pairs coverage statistics and synset pairings per word pair for WS353 dataset, plotted against the percentage of dropped words (reverse sorted using <i>tfidf</i>). The expanded bag of words is formed by synsets + glosses + hype/hypo and glosses + Wikipedia abstracts	51
5.7	Image difference detection using the Absolute Error (AE) metric in ImageMagick. Given two images, the algorithm gives an absolute count of pixels that are different, with the fuzz factor set at 20%. Only pixels changed by more than the fuzz factor are considered different. These different pixels are marked in red in the newly composed 'difference' image formed by overlaying the two images.	52
5.8	Results obtained for WS353 dataset by dropping the lowest 30% (reverse sorted by <i>tfidf</i>) of expanded set of words formed by synsets/glosses. Figures in bold represent hybrid image-text metrics that are score better correlation than the individual standalone text-based metrics	55
5.9	Results obtained for WS353 dataset by dropping the lowest 60% (reverse sorted by <i>tfidf</i>) of expanded set of words formed by synsets/glosses/hype/hypo/Wikipedia. Figures in bold represent hybrid image-text metrics that are score better correlation than the individual standalone text-based metrics.	55
5.10	Table showing word pairs with their averaged human ratings of similarity from S2	56
5.11	Spearman correlation figures obtained for S1 and S2 using the text-based and image-based methods	57
6.1	A table showing statistical information on our joint semantic space model	65
6.2	A sample of test images with their synset words and glosses : The number in parenthesis represents the numerical association of the word with the image (0-10).	

	Human annotations reveal different degree of semantic relatedness between the image and words in the synset or gloss.	66
6.3	Correlation of automatically generated scores with human annotations on cross-modal semantic relatedness, as performed on the ImageNet test dataset of 2004 pairs of word and image. Correlation figures scoring the highest within a weighting scheme are marked in bold, while those scoring the highest across weighting schemes and within a visual vocabulary size are underlined.	70
6.4	Table showing high-level categories and the corresponding attributes	77
6.5	A sample of test images with their synset words and glosses, human annotations : The number in parenthesis represents the numerical association of the word with the image (0-10). All content words are considered. The visual attributes agreeable by all annotators are listed for each image	79
6.6	Spearman correlation performance of augmented VSM-LSAs using visual attributes and their integration within high level categories	80

LIST OF FIGURES

3.1	A Wordnet “is a” relationship for <i>dog</i> .	15
3.2	A subset of images associated with a node in ImageNet. The WordNet synset illustrated here is { <i>Dog, domestic dog, Canis familiaris</i> } with the gloss: <i>A member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; “the dog barked all night”</i>	17
3.3	Illustrations of illustrated synsets in within two trees in ImageNet.	18
3.4	An illustration of the process of generating “Bag of Visual Codewords”	20
4.1	Flickr picturability labels	30
5.1	Graph showing differential correlation against different metrics, grouped by different datasets and combination functions. Any bar below zero indicates worse performance after combination with the image-based metric.	47
5.2	Schematic diagram depicting our proposed system	48
5.3	Scatter plot of picturability scores of all 447 vocabulary words in the three datasets	57
6.1	Spearman correlation performance with human annotations against number of synsets used	72
6.2	Classification accuracy, as more data is added to construct the semantic space model.	73
6.3	Examples of object attributes taken directly from ImageNet website, which are present in our dataset	75
6.4	Spearman correlation performance of VSMs augmented with individual attributes	82

CHAPTER 1

“MEN ARE FROM MARS, WOMEN ARE FROM VENUS”: INTRODUCTION TO LANGUAGE AND VISION

Both the Martians and Venusians forgot that they were from different planets and were supposed to be different. In one morning everything they had learned about their differences was erased from their memory. And since that day men and women have been in conflict.

John Gray

Since the time of existence, human beings communicate with one another by making sense of perceiving their surrounding environment. Arguably, among the five prevalent senses of sight, hearing, touch, smell and taste, it is vision that we afford the most attention. We see at all times except when we are at rest, at which point we do not consciously hear, nor feel, nor smell, nor taste. Where we perceive, we gain information in relation to how we think, speak and behave. It affects our decision, our mood and at the crux of it all, determines how we live.

The ability to make sense of what we see, coupled with the fact that we can visualize concepts and reproduce them in pictorial forms, bears huge promises as an effective medium for communication. Perhaps not surprisingly, the coexistence of linguistic and pictorial elements in human communication was dated as far back as one could imagine. In fact, before any verbal communication was established, man had used proto-linguistic or non-linguistic means to make himself understood. This evidence suggests that visual representations of information is very helpful to a non-negligible extent, and they require minimal learning in most instances. In addition, studies [3] have shown that only up to an estimated ten percent of our communication is verbal. The unspoken language is our body language. This refers to our behaviors while making the speech, such as smile, gaze, attention span, attitude, arms movement, head shaking etc, all perceived by our listener through their visual system.

Do we then argue that visual representations of information are vastly better than their corresponding linguistic counterparts ? Take, for instance, the image of an apple. The concept, i.e., “fruit with red or yellow or green skin and sweet to tart crisp whitish flesh”¹ is clearly understood and conveyed effectively by a single pictorial representation, as opposed to a multitude of linguistic representations that might have been needed, where one is required for each language available in the world. Enabling images for the purpose of communicating such concrete, picturable concepts pose many advantages. Essentially, it is universal, minimally supervised and effective.

However, our interaction with others on a daily basis deals with more perplexities than just communicating simple nouns. Fostering a deeper understanding of a topic requires more substantive expression of meanings that is better achieved by using languages. Overall, is language better then ? If a picture speaks a thousand words, how many pictures is a word worth ? Mastering a language may take years of effort, not just in absorbing a significant part of the available lexicon, but also in acquiring grammar rules that allow one to express herself clearly, by grounding the proper semantics on legal syntactic structure of uttered words. But it is exactly this learned permutation of a limited set of vocabulary words that accounts for a language’s rich and accurate expressive power of the speaker’s intentions.

For years, the supplementary and complementary advantages of using images with text have been recognized and exploited in virtually every domain. In fact, much less would remain without the introduction of visually stimulating imagery entities such as icons, cliparts, signs, logos, pictures and the like. The existence of them in newspapers, books, television, web pages, and all other media help us to understand the communicated subject quicker and better. Whatever our preference for it, they exist ubiquitously around us.

However, from a computational point of view, visual cues promise much but delivers little. Despite the large amount of information captured by our visual system, automatic translation of visual content to knowledge-bases for yielding innovation applications has been stifled by two factors, namely, the ineffectiveness of computer-based systems to see naturally like humans see, and a lack of unified framework for grounding image meanings. While word meanings can be acquired

¹1st sense of the noun in WordNet 3.1

and disambiguated using dictionaries, the meaning of an image in isolation is not well-defined and it is mostly task-specific. A given image, for instance, may be simultaneously labelled by a set of words using an automatic image annotation algorithm, or classified under a different set of semantic tags in the image classification task, or simply draw its meaning from a few representative regions following image segmentation performed in an object localization framework.

This dissertation attempts to revisit an old problem that is gaining daily importance, by providing computational approaches to explore and exploit the synergistic relationships between the visual and textual modalities, and to further stretch the boundaries of their applicability in generating novel solutions to existing problems. Specifically, I seek to find answers to the following research questions :

1. Is it possible to decode information in one modality to help tasks existing in the other modality ?

Traditionally, tasks in Computer Vision or Natural Language Processing (NLP) are performed using features obtained exclusively within the domain without regard for cues present in the other modality. Given that languages and vision are both manifestations of human cognitive concepts, is it possible to use information encoded exclusively in the visual modality to benefit tasks in the textual modality, and vice versa?

2. Considering the supplementary and complementary advantages of each modality over the other, can we integrate both image and word features into a unified framework for the construction of a richer semantic space ?

To date, a sizeable number of solutions to problems in NLP are based on models constructed using lexical semantic ontologies such as WordNet, FrameNet or VerbNet, or corpus-based resources such as Wikipedia. Can we augment the usefulness of these resources by adding a layer of visual modality ? Specifically, are we able to generate a richer semantic space that encodes not only knowledge induced by words, but also leverage on information available in images ?

3. Can we formalize the meaning of images by using words in languages?

One of the long-standing problems in computer vision is to formalize ground-truth semantics of a given image. Given the subjective nature of interpreting the *meaning* of an image, is there a way to close this semantic gap by obtaining closer correspondences to the the meaning of words in the multimodal semantic space that we seek to construct?

The dissertation is organized as follows. Chapter 2 provides an overview of the state-of-the-art research in domains concerned with words and images, including fields related, but not limited to, natural language processing and computer vision. Throughout the thesis, a number of recurrent resources are used in experiments, and the introduction to each of them is given in Chapter 3.

In alignment with my pursuit for answers to the first research question, I choose a representative task in Computer Vision, namely, the task of automatic image annotation, and support our hypothesis that information drawn exclusively from textual knowledge-bases can be used to construct an image annotation model that is competitive to the sate-of-the-art, using both image and text features. The experiments and findings are reported in Chapter 4.

In a similar spirit, I perform an investigation in the reverse direction, by using visual information as a clue to devise a new metric for measuring semantic relatedness between words and texts. Through the use of this new metric, either standalone or in combination with other textual metrics, it is possible to achieve the state-of-the-art in the relatedness task as evaluated on standard datasets used by the research community. The details of the experiments and discussions are contained in Chapter 5.

In response to the second and third research questions, I create unsupervised methods to model the synergistic relationships between words and images in a unified semantic framework. I establish a set of guidelines, based on known heuristics, to measure cross-modal semantic relationship between pairs of word and image, and perform a comparative study between automatic derivations of such a relatedness metric in correlation to that of human judgements, as measured on a standard dataset. The details of the experiments and discussions are contained in Chapter 6.

Finally, in Chapter 7, the conclusions for our work are summarized before I draw up some final thoughts for additional work that builds on the current foundations and findings.

CHAPTER 2

“ARE MEN REALLY FROM MARS, AND WOMEN REALLY FROM VENUS ?”:

SYNERGISM BETWEEN LANGUAGE AND VISION

The Venusians welcomed the Martians with open arms. They had intuitively known that this day would come. Their hearts opened wide to a love they had never felt before.

John Gray

This chapter serves to provide a sufficiently broad-based overview of previous work that simultaneously address both words and images¹, with the intention to situate our work in context. The literature review provided herein is of a generic nature. More related work specific to ideas we proposed, if relevant, would be detailed individually in the ensuing chapters.

2.1. Cognitive Science, Psychology and Neuroscience

The meaning of a sentence is encoded in each of its component words. In order to understand a sentence, all meanings of the individual words must be retrieved and combined. Do humans achieve this retrieval of meaning through a lexicon that is part of a linguistic system, or is the meaning stored as part of a general conceptual system in the brain ?

Early research efforts in cognitive science and psychology [67] have proposed two such similar hypotheses to empirically determine how word meanings are processed. The authors used rebus sentences in which a concrete noun is replaced by a pictured object. Such sentences consists of 10 to 15 words each, picture(s) inclusive. They are shown using rapid serial visual presentation (RSVP) to forty subjects. The rationale of using RSVP at a rate of 10 or 12 words per second is to present sentences so quickly that a delay in encoding the picture into a required form (e.g. silently naming it to help establish semantic link between the word before and after) would be highly disruptive to the subjects, and hence producing results that bias towards either of the hypotheses.

¹Throughout the dissertation, the terms *image* and *picture* are used interchangeably, and may jointly refer to any visible representation of an entity, such as person, plant, animal, building, landscape etc, or an event depicting interaction between them.

In all experiments performed, there was no significant delay in understanding of rebus sentences compared to all-words sentences. Accuracy wise, there was no consistent deficit in their interpretations. In fact, the speed of understanding and accuracy of comprehension or immediate recall remains the same regardless of the position of the picture (front, middle or end of the sentence), nor did it matter whether there was one picture or two pictured object replacing concrete nouns.

The conclusive evidence from these experimental results do not support theories that suggest word meanings are placed in a specialized lexical entry. On the contrary, the lexical representation of a noun merely points to a general, non-linguistic conceptual system in human brains where the meaning of a sentence is constructed. This finding points to a truth that our comprehension of real-world entities is not restricted by information encoded in any language system. It also opens up a possibility that humans can communicate with one another through non-linguistic means. These findings have also recently found support in cross-cultural studies which showed that children from different countries, not speaking each other's language, were able to communicate about children's stories just by using drawings and pictures [43].

In neuroscience, concerns were raised that hypothesized whether the human brain process sensory inputs differently, or there exists a general programming of the cerebral cortex that allows for a unilateral manner of information processing through different senses. In a revelatory study [74, 82], researchers produced a "rewired" version of the brain in ferrets to explore the influences of environmental factors on brain cortical development, where visual signals perceived from the retina are directed to the auditory cortex instead of the visual cortex, and therefore enabling animals to "see" with their "hearing" cortex. This discovery not only highlights the adaptability of the cerebral cortex in its development course to achieve specialization in processing a particular sensory input, but also bears important implications, such as the ability to perform constructive restructuring of neural prosthesis for restoring functions in the brain after it suffers from a damaging encounter, such as a stroke.

Importantly, given that the cerebral cortex in mammals is central to memory, thought, language and maintaining consciousness, this study provides an indirect evidence for the interoperability between different cortical signal processing, yet lends a strong motivation to suggest that

our brain is highly capable to re-conciliate discrepancies in information representation by different modalities, such as words and images.

2.2. Computer Vision and Image Processing

Currently, the best search engines on the Internet (Google, Bing, Yahoo) thrive on a high precision and recall of information requested by Web users. However, this success has been largely limited to Information Retrieval (IR) in the natural language domain, while searching for pictures still gives relatively low precision and poor recall. The reason behind this limitation is that most pictures are processed using the captions that label them. Since these captions are almost always packed with a high degree of noise, better searches can only result from searching beyond the captions. The content of each picture must be analysed meaningfully, then tagged with correct words that describe it, before it can be retrieved using traditional textual IR methods.

To alleviate this problem, several Web-based projects attempt to bridge concepts and images through manual image annotations. The first project, known as PicNet [9], is a knowledge-base consisting of dual visual-linguistic representations for words and phrases: seen as the smallest units of communication that carry meaning. Starting with a machine-readable dictionary, i.e. WordNet [63], that defines the words in the common vocabulary and their possible meanings, PicNet seeks to add visual representations to the dictionary entries, with the aim of building a knowledge-base that combines both verbal and visual representations of these basic concepts. PicNet relies on a Web-based system for validating associations between the words and pictures. Given a word and its possible meanings, as defined by a comprehensive dictionary, Web users participate in a variety of game-like activities targeting the association of pictures with words.

The second project, called the ESP game [81], is an online system that collects labels associated with images. The system is set up as a game, where the goal is to assign as many labels as possible to a given image. When two players concurrently assign a label, the label is considered correct and stored in the set of tags associated with the image. Unlike PicNet, which targets the assignment of pictures to words, the ESP game collects words (labels) for pictures. Most of the pictures labelled by the ESP game consist of entire scenes, which often refer to several concepts. For instance, an ESP-annotated image could have the following label assignments: car, person,

tree, house, road. While it is possible that a multiple-object scene could be used to describe a unique concept, the ESP game does not have any constraints concerning the number of concepts associated with an image, and there are often multiple salient concepts associated with an image in the ESP database.

Yet another related project is the Google Image Quiz². Provided with a set of images returned by a search performed against the Google Image search engine, the goal is to guess the keyword that was used in the search. Similar to the ESP game, the Google Image Quiz assigns labels (words) to images, and not images to words, and thus it can often be the case that an image will refer to several salient concepts that are associated with it.

There is also a large body of work concerning with automatic relation of words to pictures. In [4], the authors presented a novel approach for modeling multi-modal sets, using segmented images with associated text. By learning from the joint distribution of image regions and words, many applications can be yielded. These include predicting words associated with whole image (auto-annotation) and corresponding to particular image regions (region naming). Due to the difficult nature of applying data mining methods to collections of images, learning the relationships between image regions and its semantic correlates (words) proves to be an alternative method of multi-modal data mining. This initiative has arguably been followed with growing interest and enthusiasm, leading to several state-of-the-art generative models for automatic image annotation and classification. [12, 20, 24, 25, 26, 28, 33, 38, 51, 56]

In another effort to improve object recognition [1], researchers successfully showed that by automatic captioning using a novel graph-based approach (GCap), words may be reliably predicted from images. The assessment was done on the “standard” Corel image database where GCap outperforms recent, successful automatic captioning methods by up to 10 percentage points in captioning accuracy. This method is fast and scales well with its training and testing with time linear to the data set size. Besides, it requires no user-defined parameters, nor other tuning, which is in contrast to linear/polynomial/ kernel SVMs, k-means clustering etc. These advances in words and

²<http://www.gamesforthebrain.com/game/imagequiz>

images research can lead to better image recognition, and hence produces higher image retrieval accuracy.

2.3. Natural Language Processing and Computational Linguistics

Traditionally, word sense disambiguation (WSD) has been a well-studied problem in computational linguistics. Given a word in a sentence, the task is to determine which sense of the word (with multiple senses) is used in the context of that sentence. Consider the word *bank*. Examples of different senses include *piggy bank* (a container for keeping money), *river bank* (a slope of land besides a body of water), *Wells Fargo Bank* (a financial institution), *snow bank* (a long ridge or pile) etc. Picking the correct sense can be potentially challenging because of metaphorical or metonymic meanings that makes discrimination of closely related senses difficult. Also, there is the issue of inter-annotator variance. WSD systems are usually compared to a benchmark sense-tagged corpora by humans. However, even when creating this benchmark, decisions to arrive on which sense to use for a given word varies across human judges.

With the same spelling for each word to be disambiguated, it is hard to adopt a purely natural language based approach and expect a very good result over the “most common sense” method (which selects the most frequently used sense), usually used as a base line. In an innovative effort [5], experiments revealed that using pictures can actually help disambiguate words, while the reverse is also true. Starting from a learned set of pictures associated with words, co-constructed meanings can be established from these two different representations of the same entity. The images are then combined with sophisticated text based word sense disambiguation methods to perform disambiguation tasks over a subset of Corel image database with three to five keywords per image. The results show that this technique is superior to using pictures or text based methods alone for disambiguation. The hypothesis governing this observation is that properties implicit in one representation may be more explicit in another and therefore, more extractable. Given a large corpora for training, the relationships between these two can be learned, and hence pictures can be used to provide a non-negligible improvement over WSD tasks.

Some interest has also been followed in language processing to automatically generate corresponding pictorial representations. The WordsEye project [17] targets the generation of scenes

starting with an input text, where the system gradually builds a scene by adding objects identified from the text. Essentially, it is built as a support tool for graphic designers. In fact, although WordsEye's database consists of thousands of object models, the system works only for descriptive sentences of collated objects, and so it cannot generate scenes for prototypical sentences such as "The house has four bedrooms and one kitchen".

Other related projects along similar lines are SPRINT [86], where geometric models are created from natural language descriptions of a scene, using spatial constraints extracted from the text; Put [15], which identifies the placement of objects in a scene using an interactive natural language interface; and CarSim [41]³, which converts narratives about car accidents into 3D scenes by using techniques for information extraction coupled with a planning and a visualization module. More recently, the Text-to-Picture (TTP) system for augmentative communication [89] was used to synthesize a picture from natural language text by finding the important concepts in the text and merging the pictorial representations of these concepts.

2.4. Visual Languages, Alternative and Augmentative Communication and Education

Visual Language is an expression system involving the use of visual objects to express our thinking and feeling. It stems from the pioneering work of Rudolf Arnheim to studies by Robert Horn and includes the use of a visualizing method called *active imagination* developed by Carl Jung [80]. Built on the proposition that we can "draw" our thinking as well as verbalize it, a Visual Language may contain words, images, and shapes.

Particularly, there is a branch of visual languages called *iconic languages*, whose visual sentences consist of a spatial permutation of icons. Each icon bears a unique (or sometimes multiple) meaning in the vocabulary set of icons used in Iconic Language. In human-computer interaction, the iconic language normally has a limited vocabulary set with specific application domain such as database access, form manipulation and image processing. To facilitate the design of such iconic languages, a design methodology was devised [14] based on upon the theory of icon algebra, allowing for a flexible derivation of the meaning of iconic sentences.

³<http://www.carsim.com>

In the human-human interaction, there are also iconic languages used, especially in augmentative communication by people with speech disabilities. Much work also has been done in the area of augmentative and alternative communication regarding the use of visual-graphic symbol acquisition by pre-school age children with developmental and language delays. In their findings, the authors [2] concluded that the acquisition of a language requires an individual to organize the world into a system of symbols and referents. However, learning the relationship between a symbol and referent can be difficult for a child with serious intellectual disability and language delays. The complexity and iconicity of a symbol becomes an important issue in the decision of what medium to use for teaching languages. By using an observational experiential language intervention, they are able to study the effects of four pre-schoolers with developmental and language delays to acquire the meanings of Blissymbols⁴ and lexigrams. The results confirmed the findings that even children with such disabilities are able to acquire language skills through visual representations, although performance varies according to the participants' comprehension skills.

Commercial products with visual interfaces⁵ have also been marketed with success. They are used for augmentative communication for people with physical limitations or speech impediments, with iconic keyboards that can be touched to produce a voice output for communication augmentation. Also related to some extent is the work done in visual programming languages, where visual representations such as graphics and icons are added to programming languages to support visual interactions and to allow for programming with visual expressions. Additionally, there are also many pictorial dictionaries⁶ that boost the quick acquisition of a new language skill, through the use of word/image associations. Specific pictorial references based on medical domains⁷ are also available; these are excellent learning aids for the various professionals in their fields.

⁴Blissymbols is a symbolic, graphical language that is currently composed of over 3,000 symbols

⁵<http://www.amdi.net/>

⁶<http://www.pdictionary.com/>, <http://web.mit.edu/21f.500/www/vocab-photo/>

⁷<http://www.nlm.nih.gov/medlineplus/encyclopedia.html>

CHAPTER 3

BACKGROUND ON RESOURCES

In this chapter, we introduce several resources that are recurrently employed for the development and evaluation of our proposed ideas throughout the dissertation. We also explain the “Bag of Visual Codewords” algorithm, which is a widely adopted method to characterize image effectively for the purpose of image or scene classification and retrieval [87].

3.1. Wordnet

Wordnet¹ [61] is an online semantic lexicon for English language, its creation being inspired by current psycholinguistic theories of human lexical memory. Wordnet is distinctively different from a dictionary or a thesaurus. In a dictionary, words are ordered according to an alphabetical order, while their meanings are scattered randomly throughout. In a thesaurus, words are grouped together semantically at the expense of their alphabetical order. Wordnet attempts to combine the best of both worlds with the construction of a highly searchable lexical list with each entry belonging to a *synset*, which is a set of synonymous words or collocations (a collocation is a sequence of words often used together to show a specific meaning e.g. “car pool”). The meaning of a synset is further clarified with a short defining *gloss*. Each synset with its set of words and gloss represents a single conceptual entity and forms the most basic constructing unit for Wordnet.

In reality, Wordnet is not just seen as a vast collection of synsets. Semantically, there exists meaningful links between synsets. One example is the important *antonym* relationship which denotes a synset as having opposite meaning to another. These relationships are modeled in a way that reflects the organization of a lexicon in the human memory. Together, their existence form a web of semantics where there is a pointer from each synset (a meaning, or a single conceptual entity) to another governing the type of relationship held. This richness of information and its semantic links imply the suitability of Wordnet for use in NLP and AI applications.

¹Wordnet online lexical reference system, Princeton University, NJ, <http://wordnet.princeton.edu/>

WordNet also provides the *polysemy* count of a word, which is the number of synsets that contain the word. When a word appears in more than one synset (i.e. more than one sense, or meaning, or conceptual entity), it implies that some senses are much more common than others. Wordnet uses *frequency scores* to quantify this phenomenon. In a sample corpus, all words are semantically tagged with the corresponding synset, after which a count was given on how often a word appeared in a specific sense.

As of 2011, Wordnet database contains more than 150,000 words organized over 115,000 synsets for a total of 206,941 word-sense pairs. Table 3.1 shows the number of nouns, verbs, adjectives, and adverbs defined in Wordnet 3.0, and the number of synsets for each of these parts of speech.

Part of Speech	Words	Synsets
Noun	117,798	82,115
Verb	11,529	13,767
Adjective	21,479	18,156
Adverb	4,481	3,621
TOTAL	155,287	117,659

TABLE 3.1. Words and synsets in Wordet 3.0.

3.1.1. Semantic Relationships between Noun Synsets

Of two important relationships among noun synsets are the inheritance and part-whole relationships. The inheritance relationship simply means features are inherited from one word to the other. Figuratively speaking, words with inheritance links are ordered on a hierarchical basis, with the lower levels inheriting from the higher levels. Wordnet classify this type of relationship into the *hypernym* relationship, which states that X is a kind of Y if Y is a hypernym of X, and conversely, the *hyponym* relationship, which states that Y is a kind of X if Y is a hyponym of X. When two words share a common hypernym, we call them *coordinate terms*. Hence Inheritance can also be thought of as “IS A” relationship. For instance, a “dog” is a “canine”, a “canine” is in turn a “carnivore”, a “carnivore” is in turn a “placental” and so on. This is shown in Figure 3.1.


```

dog
⇒ canine, canid
  ⇒ carnivore
    ⇒ placental, placental mammal, eutherian, eutherian mammal
      ⇒ mammal, mammalian
        ⇒ vertebrate, craniate
          ⇒ chordate
            ⇒ animal, animal being, beast, brute, creature, fauna
              ⇒ organism, being
                ⇒ living thing, animate thing
                  ⇒ object, physical object
                    ⇒ physical entity
                      ⇒ entity

```

FIGURE 3.1. A Wordnet “is a” relationship for *dog*.

Note that the hypernym-hyponym relationship is transitive, meaning that a “dog” inherits from “mammal” i.e. a “dog” is also a “mammal” if Y is a hyponym of X. Also, the relationship is one-to-many, meaning that a “dog” can only be a “mammal”, not a “reptile” at the same time, but besides “dog”, a “cat”, a “pig”, a “duck” can all be “mammals”. They are coordinate terms. Part-whole relationship indicates a “PART OF” relationship. Intuitively, a “hand” is a part of a “body”, and hence qualify for the part-whole relationship. We call the “hand” a *meronym* of the “body”, and conversely, the “body” is a *holonym* of the “hand”.

Part-whole relationships are similar to Inheritance relationships in their hierarchical structure and transitivity. The two type of relationships can be combined to form a composite relationship, as in if X is a hyponym of Y, and W is meronym X, and Z is a meronym of Y, then W can be a hyponym of Z.

3.1.2. Semantic Relationships between Verb Synsets

Verb synsets also exhibit hypernym-hyponym relationships between them. A clear instance of such a relationship is the verb “walk.” “stagger,” “trudge,” “stride” are all hyponyms of the hypernym “walk.”

A relationship exclusive to verb synsets would be *troponym*. To understand troponymy, we first visit *entailment*, a concept well-defined for propositional logic. When X entails Y, we state that under no conceivable state of affairs, there exists a situation of X is true Y is false, and vice-versa. Now, if we say “snore” entails “sleep,” there is no way whatsoever to state confidently that “sleep” entails “snore” too, and hence the relationship is unilateral. Troponymy, thus, specify every verb X entails a more general verb Y, if X is a troponym of Y. The causative relation relates the “cause” (in the word “display”) to “effect” (in the word “see”). This type of relation is transitive; if X causes Y, Y causes Z, then we conclude that X causes Z.

Besides having semantic relations in a category, there are also semantic relations connecting different categories. For instance, an adjective modifies an attribute, resulting in a link between the adjective to the synset containing the attribute. An adverb may link to an adjective from which it is derived.

3.2. ImageNet

The ImageNet database [18] is an ontology of images developed for advancing content-based image search algorithms, and serving as a benchmarking standard for various visual recognition tasks, such as object recognition, image classification and object recognition. At its core, ImageNet exploits the hierarchical structure of WordNet by attaching relevant images to each , hence providing pictorial illustrations of the concept associated with the synset. On average, each synset contains 500-1000 images that are carefully audited through a stringent quality control mechanism, one in which annotators from Amazon Mechanical Turk² are asked to verify the inclusion of objects from synsets in candidate images.

²<https://www.mturk.com/mturk/welcome>



FIGURE 3.2. A subset of images associated with a node in ImageNet. The WordNet synset illustrated here is $\{\text{Dog, domestic dog, Canis familiaris}\}$ with the gloss: *A member of the genus Canis (probably descended from the common wolf) that has been domesticated by man since prehistoric times; occurs in many breeds; “the dog barked all night”*

The long standing goal of ImageNet project is to construct a *large-scale, accurate* and *diverse* image database, one based on a well-established semantic ontology. Starting from 27 high-level categorical synsets, images are grouped according to the hierarchical structure of WordNet via the hypernym-hyponym semantic relationship. The end result is a large network of trees, each connecting pictorially-illustrated sysnets. The current 27 subtrees consists of 14 million images spread over 21841 categorical synsets. Figure 3.2 shows a example of illustrated synset in ImageNet, while Figure 3.3 depicts snapshots of two root-to-leaf branches of two ImageNet trees.

To date, ImageNet is the *largest* image database to be constructed, in terms of the number of categories, number of images per category and the total number of images. As an illustration, at the time of its creation, no other image dataset offers illustration of 147 categories [18]. A comparative study made between ImageNet and the ESP data [81] reveals the former to possess larger and denser network of interconnected nodes, even when observing the same sub-trees of images.

Due to its image collection policy which is not biased toward any specific positioning, pose, appearance, background clutter or occlusions, there exists a rich diversity of images qualifying under the same sysnset. This is confirmed by measuring the lossless JPG file size of the average of

each image, where a smaller file size corresponds to a blurrier image of higher gray scaled patches, which is consistent and characteristic of commonly pooled images covering a greater diversity in their appearances.

Due to the varying degrees of specificity of synsets under a tree, obtaining a clean dataset overall is challenging. Such an intuition is justified by the observation that images of a “working dog” is harder to collect than a “dog”. To quantify the accuracy of images present in ImageNet, a total of 80 synsets are randomly sampled at different levels tree depth and judged by an independent group of subjects. On average, a precision of 99.7% is achieved for each synset.

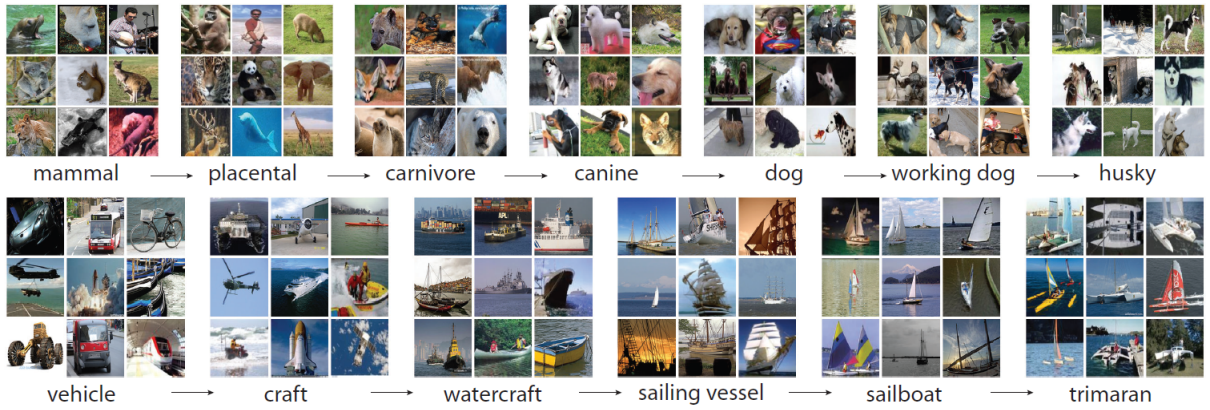


FIGURE 3.3. Illustrations of illustrated synsets in within two trees in ImageNet.

3.3. Bag of Visual Codewords using Shift-Invariant Feature Transform

Inspired by the bag-of-words approach employed in information retrieval, the “bag of visual codewords” is a similar technique used mainly for scene classification [87]. Starting with an image collection, visual features are first extracted as data points from each image, characterizing its appearance. By projecting data points from all the images into a common space and grouping them into a large number of clusters such that similar data points are assigned to the same cluster, we can treat each cluster as a “visual codeword” and express every image in the collection as a “bag of visual codewords”. This representation enables the application of methods used in text retrieval to tasks in image processing and computer vision.

Typically, the type of visual features selected can be *global* – suitable for representing an entire image, or *local* – specific to a given region in the image, depending on task requirement.

For a global representation, features are often described using a continuous feature space, such as a color histogram in three different color spaces (RGB, HSV and LAB), or textures using Gabor and Haar wavelets [56]. Likewise, local descriptors such as key points [24] can also adopt such a representation. Regardless of the features used, visual codeword generation involves the following three important phases.

- (1) **Feature detection:** The image is divided into partitions of varying degrees of granularity from which features can be extracted and represented. Typically, we can employ normalized cuts to divide an image into irregular regions, or apply uniform segmentation to break it into smaller but fixed grids, or simply locate information-rich local patches on the image using interest point detectors.
- (2) **Feature description:** A descriptor is selected to represent the features that are being extracted from the image. Typically, feature descriptors (global or local) are represented as numerical vectors, with each vector describing the feature extracted in each region. This way, an image is represented by a set of vectors from its constituent regions.
- (3) **Visual codeword generation:** Clustering methods are applied to group vectors into clusters, where the center of each cluster is defined as a visual codeword, and the entire collection of clusters defines the visual vocabulary for that image collection. Each image region or patch abstracted in feature detection is now represented by the visual codeword mapped from its corresponding feature vector.

The process of visual codeword generation is illustrated in Figure 3.4. [24] has shown that, unlike most previous work on object or scene classification that focused on adopting global features, local features are in fact extremely powerful cues. In this dissertation, we use the Scale-Invariant Feature Transform (SIFT) introduced by [55] to describe distinctive local features of an image in the feature description phase. SIFT descriptors are selected for their invariance to image scale,

Feature detection on overlapping patches

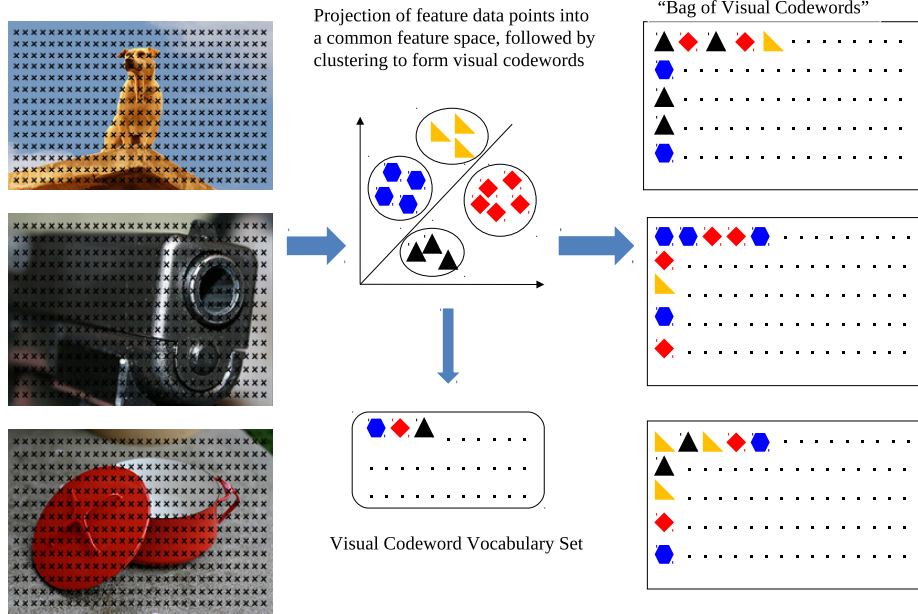


FIGURE 3.4. An illustration of the process of generating "Bag of Visual Codewords"

rotation, differences in 3D viewpoints, addition of noise, and change in illumination. They are also robust across affine distortions.

CHAPTER 4

“MEN CAN BE FROM VENUS, WOMEN CAN BE FROM MARS”: INTEROPERABILITY OF FEATURES BETWEEN LANGUAGE AND VISION TASKS (I) – TEXT MINING FOR AUTOMATIC IMAGE TAGGING

Martians value power, competency, efficiency, and achievement. They are always doing things to prove themselves and develop their power and skills. Their sense of self is defined through their ability to achieve results. They experience fulfilment primarily through success and accomplishment.

John Gray

In this chapter¹, we attempt to bridge the gap between textual features engineering and vision tasks, by showing that it is possible to apply features generated using evidence gathered from text corpora directly for solving a selected task in computer vision. We first outline the motivation for making automatic image tagging as the selected vision task. Next, we propose a new evaluation framework for image tagging, which is based on an analogy drawn between the tasks of image labeling and lexical substitution. We then present three extractive approaches for the task of image annotation. The methods proposed are based only on the text surrounding an image, without the use of image features. Finally, by combining several orthogonal methods through machine learning, we show that it is possible to achieve a performance that is competitive to a state-of-the-art image annotation system that relies on visual and textual features, thus demonstrating the effectiveness of text-based extractive annotation models.

4.1. Motivation

With continuously increasing amounts of images available on the Web and elsewhere, it is important to find methods to annotate and organize image databases in meaningful ways. Tagging images with words describing their content can contribute to faster and more effective image search and

¹published in [50]

classification. In fact, a large number of applications, including the image search feature of current search engines (e.g., Yahoo!, Google) or the various sites providing picture storage services (e.g., Flickr, Picasa) rely exclusively on the tags associated with an image in order to search for relevant images for a given query.

However, the task of developing accurate and robust automatic image annotation models entails daunting challenges. First, the availability of large and correctly annotated image databases is crucial for the training and testing of new annotation models. Although a number of image databases have emerged to serve as evaluation benchmarks for different applications, including image annotation [20], content-based image retrieval [51] and cross language information retrieval [32], such databases are almost exclusively created by manual labeling of keywords, requiring significant human effort and time. The content of these image databases is often restricted only to a few domains, such as medical and natural photo scenes [32], and specific objects like cars, airplanes, or buildings [28]. For obvious practical reasons, it is important to develop models trained and evaluated on more realistic and diverse image collections.

The second challenge concerns the extraction of useful image and text features for the construction of reliable annotation models. Most traditional approaches relied on the extraction of image colors and textures [51], or the identification of similar image regions clustered as blobs [20] to derive correlations between image features and annotation keywords. In comparison, there are only a few efforts that leverage on the multitude of resources available for natural language processing to derive robust linguistic-based image annotation models. One of the earliest efforts involved the use of captions for face recognition in photographs through the construction of a specific lexicon that integrates linguistic and photographic information [77]. More recently, several approaches have proposed the use of WordNet as a knowledge-base to improve content-based image annotation models, either by removing noisy keywords through semantic clustering [40] or by inducing a hierarchical classification of candidate labels [78].

In this chapter, we explore the use of several natural language resources to construct image annotation models that are capable of automatically tagging images from unrestricted domains with



	Normal Image	Mode Image
		
Gold standard	czech (5), festival (5), Oklahoma (4), yukon (4), october (4), web page (2), the first (2), event (2), success (1), every (1), year (1)	train (5), station (4), steam (4), trans siberian (4), steam train (4), travel (3), park (3), siberian (3), old (3), photo (1), trans (2), yekaterinburg (2), the web (2), photo host (1)

TABLE 4.1. Two sample images. The number besides each label indicates the number of human annotators agreeing on that label. Note that the mode image has a tag (i.e. “train”) in the gold standard set most frequently selected by the annotators

good accuracy. Unlike traditional image annotation methodologies that generate tags using image-based features, we propose to extract them in a manner analogous to keyword extraction. Given a target image and its surrounding text, we extract those words and phrases that are most likely to represent meaningful tags. More importantly, we are interested to investigate the potential of such linguistic-based models on image annotation accuracy and reliability. Our work is motivated by the need for annotation models that can be efficiently applied on a very large scale (e.g. harvesting images from the web), which are required in applications that cannot afford the complexity and time associated with current image processing techniques.

4.2. Dataset

As the methods we propose are extractive, standard image databases with no surrounding text such as Corel [20] are not suitable, nor are they representative for the challenges associated with raw data from unrestricted domains. We thus create our own dataset using images randomly extracted from the Web.

To avoid sparse searches, we use a list of the most frequent words in the British National Corpus as seed words, and query the web using the Google Image API. A webpage is randomly selected from the query results if it contains a single image in the specified size range (width and

height of 275 to 1000 pixels²) and its text contains more than 10 words. Next, we use a Document Object Model (DOM) HTML parser³ to extract the content of the webpage. Note that we do not perform manual filtering of our images except where they contain undesirable qualities (e.g. porn, corrupted or blank images).

In total, we collected 300 image-text pairs from the web. The average image size is 496 pixels width and 461 pixels height. The average text length is 278 tokens and the average document title length is 6 tokens. In total, there are 83,522 words and the total vocabulary is 8,409 words.

For each image, we also create a gold standard of manually assigned tags, by using the labels assigned by five human annotators. The image annotation is conducted via Amazon Mechanical Turk, which was shown in the past to produce reliable annotations [76]. For increased annotation reliability, we only accept annotators with an approval rating of 98%.

Given an image, an annotator extracts from the associated text a minimum of five words or collocations. Annotators can choose words freely from the text, while collocation candidates are restricted to a fixed set obtained from the n-grams ($n \leq 7$) in the text that also appear as article names or surface forms in Wikipedia. Moreover, when interpreting the image, the annotators are instructed to focus on both the denotational and conotational attributes present in the image⁴.

4.3. A New Evaluation Framework: Image Tagging as Lexical Substitution

While evaluations of previous work in image annotation were often based on labels provided with the images, such as tags or image captions, in our dataset such annotations are either missing or unreliable. A random sampling of 15 images reveal that 7 lack captions. Some images have empty ALT tag, while others are labeled with non-useful information such as DSN0001.jpg, 4x400.png etc. We rely instead on human-produced extractive annotations (as described in the previous section), and formulate a new evaluation framework based on the intuition that an image can be substituted with one or more tags that convey the same meaning as the image itself. Ideally, there

²Empirically determined to filter advertisements, banners and undersized images.

³<http://search.cpan.org/dist/HTML-ContentExtractor/>

⁴Annotation instructions, dataset and gold standard can be downloaded at <http://lit.csci.unt.edu/index.php/Downloads>

is a single tag that “best” describes the image overall (i.e. the gold standard tag agreed by the majority of human annotators), but there are also multiple tags that describe the fine-grained concepts present in the image. Our evaluation framework is inspired by the lexical substitution task [57], where a system attempts to generate a word (or a set of words) to replace a target word, such that the meaning of the sentence is preserved.

Given this analogy, the evaluation metrics used for lexical substitution can be adapted to the evaluation of image tagging. Specifically, we measure the precision and the recall of a tagging method using four subtasks: **best normal**: provides precision and recall for the top-ranked tag returned by a method; **best mode**: provides precision and recall only if the top-ranked tag by a method matches the tag in the gold standard that was most frequently selected by the annotators; **out of ten (oot) normal**: provides precision and recall for the top ten tags by the system; and **out of ten (oot) mode**: similar to best mode, but it considers the top ten tags returned by the system instead of one. Table 4.1 show examples of a normal and a mode image.

Formally, let us assume that H is the set of annotators, namely $\{h_1, h_2, h_3, \dots\}$, and I , $\{i_1, i_2, i_3, \dots\}$ is the set of images for which each human annotator provide at least five tags. For each i_j , we calculate m_j , which is the most frequent tag for that image, if available. We also collect all r_j^k , which is the set of tags for the image i_j from the annotator h_k .

Let the set of those images where there is a tag agreed upon by the most annotators (i.e. the images with a mode) be denoted by IM , such that $IM \subseteq I$. Also, let $A \subseteq I$ be the set of images for which the system provides more than one tag. Let the corresponding set for the images with modes be denoted by AM , such that $AM \subseteq IM$. Let $a_j \in A$ be the set of system’s extracted tags for the image i_j .

Thus, for each image i_j , we have the set of tags extracted by the system, and the set of tags from the human annotators. As the next step, the multiset union of the human tags is calculated, and the frequencies of the unique tags is noted. Therefore, for image i_j , we calculate R_j , which is $\sum r_j^k$, and the individual unique tag in R_j , say res , will have a frequency associated with it, namely $freq_{res}$.

Given this setting, the precision (P) and recall (R) metrics we use are defined below.

Best measures:

$$(1) \quad P = \frac{\sum_{a_j: i_j \in A} \frac{\sum_{res \in a_j} freq_{res}}{|a_j|}}{|A|}$$

$$(2) \quad R = \frac{\sum_{a_j: i_j \in I} \frac{\sum_{res \in a_j} freq_{res}}{|a_j|}}{|I|}$$

$$(3) \quad modeP = \frac{\sum_{bestguess_j \in AM} (1if_best_guess = m_j)}{|AM|}$$

$$(4) \quad modeR = \frac{\sum_{bestguess_j \in IM} (1if_best_guess = m_j)}{|IM|}$$

Out of ten (oot) measures:

$$(5) \quad P = \frac{\sum_{a_j: i_j \in A} \frac{\sum_{res \in a_j} freq_{res}}{|R_j|}}{|A|}$$

$$(6) \quad R = \frac{\sum_{a_j: i_j \in I} \frac{\sum_{res \in a_j} freq_{res}}{|R_j|}}{|I|}$$

$$(7) \quad modeP = \frac{\sum_{a_j: i_j \in AM} (1if_any_guess \in a_j = m_j)}{|AM|}$$

$$(8) \quad modeR = \frac{\sum_{a_j: i_j \in IM} (1_{if_any_guess \in a_j = m_j})}{|IM|}$$

As a simplified example (with less tags), consider i_j showing a picture of a Chihuahua being labeled by five annotators with the following tags, shown in Table 4.2 :

Annotator	Tags
1	dog,pet
2	chihuahua
3	animal,dog
4	dog,chihuahua
5	dog

TABLE 4.2. Example of tags provided by 5 independent annotators for an image depicting the dog “Chihuahua”

In this case, $r_j^1 = \{\text{dog,pet}\}$, $r_j^2 = \{\text{chihuahua}\}$, $r_j^3 = \{\text{animal,dog}\}$ and so on. The tag “dog” appears the most frequent among the five annotators, hence $m_j = \{\text{dog}\}$. $R_j = \{\text{dog, dog, dog, dog, chihuahua, chihuahua, animal, pet}\}$. The res with associated frequencies would be dog 4, chihuahua 2, animal 1, pet 1. If the system’s proposed tag for i_j is $\{\text{dog, animal}\}$, then the numerator of P and R for best subtask would be $\frac{4+1}{8} = 0.313$. Similarly, the numerator of P and R for oot subtask is $\frac{4+1}{8} = 0.625$.

4.4. Extractive Image Annotation

The main idea underlying our work is that we can perform effective image annotation using information drawn from the associated text. Following [26], we propose that an image can be annotated with keywords capturing the denotative (entities or objects depicted) and connotative (semantics or ideologies interpreted) attributes in the image. For instance, a picture showing a group of athletes and a ball may also be tagged with words like “soccer,” or “sports activity.” Specifically, we use a combination of knowledge sources to model the denotative quality of a word as its picturability, and the connotative attribute as its saliency. The idea of visualness and salience as textual features

for discovering named entities in an image was first pursued by [19], using data from the news domain. In contrast, we are able to perform annotation of images from unrestricted domains using content words (nouns, verbs and adjectives). In the following, we first describe three unsupervised extractive approaches for image annotation, followed by a supervised method using a re-ranking hypothesis that combines all the methods.

Flickr Picturability

Featuring a repository of four billion images, Flickr (<http://www.flickr.com>) is one of the most comprehensive image resources on the web. As a photo management and sharing application, it provides users with the ability to tag, organize, and share their photos online. Interestingly, an inspection of Flickr tags for randomly selected images reveal that users tend to describe the denotational attributes of images, using concrete and picturable words such as *cat*, *bug*, *car* etc. This observation lends evidence to Flickr’s suitability as a resource to model the picturability of words.

Algorithm 1 Flickr Picturability Algorithm

```

Start :  $L[] = \phi$  ,  $TF[] = \text{tf of each word in } T$ 
for each word in  $T$  do
  if  $\text{length}(\text{word}) \geq \alpha$  then
     $\text{RelatedTags} = \text{getRelatedTags}(\text{word});$ 
    if  $\text{size}(\text{RelatedTags}) > 0$  then
       $L[\text{word}] += \beta * TF[\text{word}]$ 
      for each tag in  $\text{RelatedTags}$  do
        if  $\text{exists } TF[\text{tag}]$  then
           $L[\text{tag}] += TF[\text{tag}]$ 
        end if
      end for
    end if
  end if
end for

```

Given the text (T) of an image, we can use the *getRelatedTags* API to retrieve the most frequent Flickr tags associated with a given word, and use them as corpus evidence to filter or promote words in the text. In the filtering phase we ignore any words that return an empty list

of Flickr’s related tags, based on the assumption that these words are not used in the Flickr tags repository. We also discard words with a length that is less than three characters ($\alpha=3$). In the promotion phase, we reward any retrieved tags that appear as surface forms in the text. This reward is proportional to the term frequency of these tags in the text. Additionally, we also include in the final label set any word that returns a non-empty related tags set with a discounted weight ($\beta=0.5$) of its term frequency, to the end of enriching our labels set while assuring more credit are given to the picturable words. The algorithm is described in Algorithm 1.

To extract multiword labels, we locate all n-grams formed exclusively from our extracted set of possible labels. The subsequent score for each of these n-grams is:

$$(9) \quad L[w_i..w_{i+k}] = \left(\sum_{j=i}^{j=i+k} L[w_j] \right) / k$$

By reverse sorting the associative array in L , we can retrieve the top K words to label the image. For illustration, let us consider the following text snippet.

On the Origin of Species, published by Charles Darwin in 1859, is considered to be the foundation of evolutionary biology.

After removing stopwords, we consider the remaining words as candidate labels. For each of these candidates w_i (i.e. *origin, species, published, charles, darwin, foundation, evolutionary, and biology*), we query Flickr and obtain their related tag set R_i . *origin, published, and foundation* return an empty set of related tags and hence are removed from our set of candidate labels, leaving *species, charles, darwin, evolutionary, and biology* as possible annotation keywords with the initial score of 0.5. In the promotion phase, we score each w_i based on the number of votes it receives from the remaining w_j (Figure 4.1). Each vote represents an occurrence of the candidate

tag w_i in the related tag set R_j of the candidate tag w_j . For example, *darwin* appeared in the Flickr related tags for *charles*, *evolutionary*, and *biology*, hence it has a weight of 3.5. The final list of candidate labels are shown in Table 4.3.

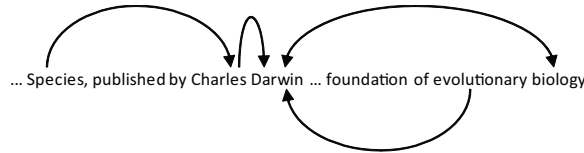


FIGURE 4.1. Flickr picturability labels

Label	$S(w_i)$
darwin	3.5
charles darwin	2.5
charles	1.5
biology	1.5
evolutionary biology	1.0
evolutionary	0.5
species	0.5

TABLE 4.3. Candidate labels obtained for a sample text using the Flickr model

Wikipedia Salience

We hypothesize that an image often describes the most important concepts in the associated text. Thus, the keywords selected from a text could be used as candidate labels for the image. We use a graph-based keyword extraction method similar to [60], enhanced with a semantic similarity measure. Starting with a text, we extract all the candidate labels and add them as vertices in the graph. A measure of word similarity is then used to draw weighted edges between the nodes. Using the PageRank algorithm, the words are assigned with a score indicating their salience within the given text.

To determine the similarity between words, we use a directed measure of similarity. Most word similarity metrics provide a single-valued score between a pair of words w_1 and w_2 to indicate their semantic similarity. Intuitively, this is not always the case, as w_1 may be represented

by concepts that are entirely embedded in other concepts, represented by w_2 . In psycholinguistics terms, uttering w_1 may bring to mind w_2 , while the appearance of w_2 without any contextual clues may not associate with w_1 . For example, *Obama* brings to mind the concept of *president*, but *president* may trigger other concepts such as *Washington*, *Lincoln*, *Ford* etc., depending on the existing contextual clues. Thus, the degree of similarity of w_1 with respect to w_2 should be separated from that of w_2 with respect to w_1 . Specifically, we use the following measure of similarity, based on the Explicit Semantic Analysis (ESA) vectors derived from Wikipedia [30]:

$$(10) \quad DSim(w_i, w_j) = \frac{C_{ij}}{C_i} * Sim(w_i, w_j)$$

where C_{ij} is the count of articles in Wikipedia containing words w_i and w_j , C_i is the count of articles containing words w_i , and $Sim(w_i, w_j)$ is the cosine similarity of the ESA vectors representing the input words. The *directional weight* (C_{ij}/C_i) amounts to the degree of association of w_i with respect to w_j . Using the directional inferential similarity scores as directed edges and distinct words as vertices, we obtain a graph for each text. The directed edges denotes the idea of “recommendation” where we say w_1 recommends w_2 if and only if there is a directed edge from w_1 to w_2 , with the weight of the recommendation being the directional similarity score. Starting with this graph, we use the graph iteration algorithm from [60] to calculate a score for each vertex in the graph. The output is a sorted list of words in decreasing order of their ranks, which are used as candidate labels to annotate the image. This is achieved by using C_j instead of C_i for the denominator in the directional weight. Table 4.4 shows the directional similarity for some words.

To illustrate with an example, consider the text snippet :

Microsoft Corporation is a multinational computer technology corporation that develops, manufactures, licenses, and supports a wide range of software products for

w_i	w_j	$\text{DSim}(w_i, w_j)$
broadband	Internet	0.797
Internet	broadband	0.032
Ipod	apple	0.792
apple	Ipod	0.076
Bush	president	0.385
president	Bush	0.072
Microsoft	software	0.550
software	Microsoft	0.231

TABLE 4.4. Directional similarity scores for some words

computing devices

after stopword removal, the list of nouns extracted is *Microsoft, computer, corporation, devices, products, technology, software*. Note that the top-ranked word must infer some or all of the words in the text. In this case, the word *Microsoft* infers the terms *computer, technology* and *software*.

To calculate the semantic relatedness between two collocations, we use a simplified version of the text-to-text relatedness technique proposed by [36] and [58] that incorporate the directional inferential similarity as an underlying semantic metric.

Formally, let T_a and T_b be two text fragments of size a and b respectively. After removing all stopwords, we first determine the number of shared terms (ω) between T_a and T_b . Second, we calculate the semantic relatedness of all possible pairings between non-shared terms in T_a and T_b . We further filter these possible combinations by creating a list φ which holds the strongest semantic pairings between the fragments' terms, such that each term can only belong to one and only one pair.

$$(11) \quad \text{Sim}(T_a, T_b) = \frac{(\omega + \sum_{i=1}^{|\varphi|} \varphi_i) \times (2ab)}{a + b}$$

where ω is the number of shared terms between the collocations and φ_i is the similarity score for the i th pairing.

Topical Modeling

Intuitively, every text is written with a topic in mind, and the associated image serves as an illustration of the text meaning. In this work, we investigate the effect of topical modeling on image annotation accuracy directly. We use the Pachinko Allocation Model (PAM) [53] to model the topics in a text, where keywords forming the dominant topic are assumed as our set of annotation keywords. Compared with previous topic modeling approaches, such as Latent Dirichlet allocation (LDA) or its improved variant Correlated Topic Model (CTM) [8], PAM captures correlations between all the topic pairs using a directed acyclic graph (DAG). It also supports finer-grained topic modeling, and has state-of-the-art performance on the tasks of document classification and topical keyword coherence. Given a text, we use the PAM model to infer a list of *super-topics* and *sub-topics* together with words weighted according to the likelihood that they belong to each of these topics. For each text, we retrieve the top words belonging to the dominant super-topic and sub-topic. We use 50 super-topics and 100 sub-topics as operating parameters for PAM, since these values were found to provide good results in previous work on topic modeling. Default values are used for other parameters in the model.

4.5. Supervised Learning

The three tagging methods target different aspects of what constitutes a good label for an image. We use them as features in a machine learning framework, and introduce a final rank attribute $S(t_j)$, which is a linear combination of the reciprocals of the rank of each tag as given by each method,

$$(12) \quad S(t_j) = \sum_{m \in \text{methods}} \lambda_m \frac{1}{r_{t_j}^m}$$

where $r_{t_j}^m$ is the rank for tag t_j given by method m . The weight of each method λ_m is estimated from the training set using information gain values. Since our predicted variable (*mode* precision or recall) is continuous, we use the Support Vector Algorithm (nu-SVR) implementation of SVM [13] to perform regression analysis on the weights for each method via a radial basis function kernel. A ten-fold cross-validation is applied on the entire dataset of 300 images.

4.6. Experiments and Evaluations

We evaluate the performance of each of the three tagging methods separately, followed by an evaluation of the combined method. Each system produces a ranked list of K words or collocations as tags assigned to a given image. A system can discretionary generate less (but not more) than K tags, depending on its confidence level.

For comparison, we implement three baselines: *tf*idf*, *Doc Title* and *Random*. For *tf*idf*, we use the British National Corpus to calculate the *idf* scores, while the frequency of a term is calculated from the entire text associated with an image. The *Doc Title* baseline is similar, except that the term frequency is calculated based on the title of the document. The *Random* baseline randomly selects words from a co-occurrence window of size K before and after an image as its annotation. Following other tagging methods, we apply a pre-processing stage, where we part-of-speech tag the text (to retain only nouns), followed by stemming. We also determine an upper bound, which is calculated as follows. For each image, the labels assigned by each of the five annotators are in turn evaluated against a gold standard consisting of the annotations of the other four annotators. The best performing annotator is then recorded. This process is repeated for each of the 300 images, and the average precision and recall are calculated. This represents an upper bound, as it is the best performance that a human can achieve on this dataset. Table 4.5 shows our experimental results.

	Best				out-of-ten (oot)			
	Normal		Mode		Normal		Mode	
Models	P	R	P	R	P	R	P	R
Flickr picturability	6.32	6.32	78.57	78.57	35.61	35.61	92.86	92.86
Wikipedia Saliency	6.40	6.40	7.14	7.14	35.19	35.19	92.86	92.86
Topic modeling	5.99	5.99	42.86	42.86	37.13	37.13	85.71	85.71
Combined (SVM)	6.87	6.87	67.49	67.49	37.85	37.85	100.00	100.00
Doc Title	6.40	6.40	75.00	75.00	18.97	18.97	82.14	82.14
$tf * idf$	5.94	5.94	14.29	14.29	38.40	38.40	78.57	78.57
Random	3.76	3.76	3.57	3.57	30.20	30.20	50.00	50.00
Upper bound (human)	12.23	12.07	81.48	81.48	82.44	81.55	100.00	100.00

TABLE 4.5. Results obtained on the Web dataset

Among the individual methods, the method implementing Flickr picturability has the highest individual score for *best* and *oot* modes, yielding a precision and recall of 78.57% and 92.86% respectively. The Wikipedia Saliency method also scores the highest (jointly with Flickr) in the *oot* mode, but for the *best* mode achieves a score only marginally better than the random baseline. A plausible explanation is that it tends to favor “all-inferring” over-specific labels, while the most frequently selected tags in mode pictures are typically more “picturable” than being specific (e.g. “train” for the mode picture in Table 4.1). The topic modeling method has mixed results: its scores for *oot* normal and mode are somewhat competitive with $tf*idf$, but it scores consistently lower than the DocTitle in the *best* subtask, possibly due to the absence of a more sophisticated re-ranking algorithm tailored for the image annotation task other than the intrinsic ranking mechanism in PAM. It is worth noting that the combined supervised system provides the overall best results (6.87%) on the *best* normal, and achieves a perfect precision and recall (100%) for *oot* mode, which means perfect agreement with the human tagging.

4.7. Related Work

We also compare our work against [26] as it allows for a direct comparison with models using both image and textual features under a standard evaluation framework. We obtained the BBC dataset used in their experiments, which consists of 3121 training and 240 testing images. In this dataset, images are implicitly tagged with captions by the author of the corresponding BBC article. The evaluations are run against these captions.

	Top 10			Top 15			Top 20		
Models	P	R	F1	P	R	F1	P	R	F1
<i>tf*idf</i>	4.37	7.09	5.41	3.57	8.12	4.86	2.65	8.89	4.00
DocTitle	9.22	7.03	7.20	9.22	7.03	7.20	9.22	7.03	7.20
Lavrenko03	9.05	16.01	11.81	7.73	17.87	10.71	6.55	19.38	9.79
ExtModel	14.72	27.95	19.82	11.62	32.99	17.18	9.72	36.77	15.39
Flickr picturability	12.13	22.82	15.84	9.52	26.82	14.05	8.23	29.80	12.90
Wikipedia Saliency	11.63	21.89	15.18	9.28	26.20	13.70	7.81	29.41	12.35
Topic Modeling	11.42	21.49	14.91	9.28	26.20	13.70	7.86	29.57	12.42
Combined (SVM)	13.38	25.17	17.47	11.08	31.29	16.37	9.50	35.76	15.01

TABLE 4.6. Results obtained on the BBC dataset used in [26]

In their experiments, Feng and Lapata created four annotation models. The first two (*tf*idf* and Document Title) are the same as used in our baseline experiments. The third model (Lavrenko03) is an application of the continuous relevance model in [38], trained with the BBC image features and captions. Finally, the forth (ExtModel) is an extension of the relevance model using additional information in auxiliary texts. Briefly, the model assumes a multiple Bernoulli distribution for words in a caption, and generates tags for a test image using a weighted combination of the accompanying document, caption and image features learned during training.

The experimental setup is similar to the earlier section, but a few modifications are made for a fair and direct comparison. First, we extend our models coverage to include content words (i.e. nouns, verbs, adjectives) determined using the Tree Tagger [73]. Second, no collocations are used. Third, we adopt the evaluation framework used by Feng and Lapata to extract the top 10, 15 and 20 tags. Note that in our methods, the extraction of tags for a test image is only done on the document surrounding the image, after excluding the caption. As the number of negative examples (words not present in the caption) greatly outnumber the positive instances, we employ an undersampling method [44] to balance the dataset for training.

The results are shown in Table 4.6. Interestingly, all our unsupervised extraction-based models perform consistently above the supervised Lavrenko03 model, indicating that textual features are more informative than captions and image features taken together. Comparing with models using significantly less document information (*tf*idf* and Doc title), our models gain even

greater advantage. Note that the title of any BBC article does not exceed 10 words, hence comparison is only meaningful given the top 10 tags retrieved.

Feng and Lapata used LDA to perform reranking of final candidates in their ExtModel. However, when used as a model alone, the PAM topic model achieved promising scores in all the categories, performing best for top 10 keywords (F1 of 14.91%). Flickr picturability stands out as the best performing unsupervised method, scoring the highest precision (12.13%, top 10), recall (29.80%, top 20) and F1 (15.84%, top 10).

Overall, this comparative evaluation yields some important insights. First, our combined model using SVM is statistically better ($p < 0.1$ for top 10, 15, 20) than the Laverenko03 model, but not statistically different from the ExtModel. This demonstrates the effectiveness of textual-based models over traditional models trained with image features and captions. While it is intuitively clear that image features help in improving tagging performance, we show that mining only the text surrounding an image, where it exists, can yield a performance that is comparable to a state-of-the-art system that uses both textual and visual features. Moreover, an increase in complexity of a model by using more features may hinder its applicability to large datasets, but not necessarily improving annotation performance [56]. On this, text-based annotation models can provide a desirable compromise. For instance, our unsupervised models implementing Flickr picturability and Wikipedia Saliency are able to extract annotations from a BBC article (average 133.85 tokens) in approximately 1 second and 20 seconds respectively.

Other Related Work

Several online systems have sprung into existence to achieve annotation of real world images through human collaborative efforts (Flickr) and stimulating competition [81]. Although a large number of image tags can be generated in short time, these approaches depend on the availability of human annotators and are far from being automatic. Similarly, research in the other direction via text-to-image synthesis [52, 16, 59] has also helped to harvest images, mostly for concrete words, by refining image search engines.

Most approaches to automatic image annotation have focused on the generation of image labels using annotation models trained with image features and human annotated keywords [6, 38, 56, 83]. Instead of predicting specific words, these methods generally target the generation of semantic classes (e.g. vegetation, animal, building, places etc), which they can achieve with a reasonable amount of success. Recent work has also considered the generation of labels for real-world images [51, 26]. To our knowledge, we are unaware of any other work that performs extractive annotation for images from unrestricted domains through the exclusive use of textual features.

CHAPTER 5

“MEN CAN BE FROM VENUS, WOMEN CAN BE FROM MARS”: INTEROPERABILITY OF FEATURES BETWEEN LANGUAGE AND VISION TASKS (II) – AN IMAGE-BASED APPROACH FOR MEASURING WORD RELATEDNESS

Venusians have different values. They value love, communication, beauty, and relationships. They spend a lot of time supporting, helping, and nurturing one another. Their sense of self is defined through their feelings and the quality of their relationships. They experience fulfilment through sharing and relating.

John Gray

This chapter continues to explore the interoperability of features between language and vision tasks, by performing research in the reverse direction to exploit visual information for improving tasks in NLP. We select the semantic relatedness problem as the representative task to evaluate our hypothesis. Traditional approaches to the semantic relatedness task are often restricted to text-based methods with little regard for other multimodal evidence. Here, we propose a novel metric to estimate the similarity of words by comparing visual similarity of concepts invoked by these words. We demonstrate the promise of such a method through comparative evaluations on three standard datasets. Furthermore, by using the correct combination function to integrate image information with other similarity metrics, it is possible to attain the state-of-the-art, confirming the applicability of visual cues as a possible orthogonal information source in measuring similarity between words.

5.1. Motivation

Measuring the semantic relatedness of words is an important task with applications in information extraction and retrieval, query reformulation, word sense disambiguation, plagiarism detection and textual entailment. Owing mainly to the nature of this task, research efforts in the past have typically centered around methodologies employing the use of knowledge-based or corpus-based textual resources, with only little (if any) work paying attention to evidence provided by other

multimodal sources, such as visual cues presented by the images that are associated with a given word. While it can be shown that the human cognitive system is sensitive to visual information, and incorporating a dual linguistic-and-pictorial representation of information can actually enhance knowledge acquisition [66], the use of visual information to improve tasks in natural language processing has been largely unexplored.

In this chapter¹, we hypothesize that the relatedness between the visual representations of a pair of words can be effectively used to gauge their similarity. We employed the “bag of visual words” technique, widely used in computer vision and discussed in 3.3, to show how distinctive features of an image can be harvested. The main resource for our experiments is ImageNet, discussed in 3.2, which is used in our work to bridge the semantic gap between words and images. Finally, we show how a new relatedness metric based exclusively on visual information can be constructed for the semantic relatedness task. We evaluate this metric alongside existing corpus-based [79] and knowledge-based metrics [65] either in a standalone or combined setting and present our findings.

5.2. Dataset

Given the maturity of techniques used to extract visual content from images, it is possible to study the synergistic relationships between semantic representations of words and images given the availability of a large lexical resource with associated relevant images. For such a resource, we turn to the ImageNet, a visual-linguistic resource outlined and discussed in 3.2.

Compared to other image databases with keyword annotations, we believe that ImageNet is suitable for evaluating our hypothesis for two important reasons. First, by leveraging on reliable semantic annotations in WordNet (i.e., words in the synset), we can effectively circumvent the propagation of errors caused by unreliable annotations, and consequently hope to reach more conclusive results for this study. Second, unlike other image databases, ImageNet consists of millions of images, and it is a growing resource with more images added on a regular basis. This aligns with our long-term goal of extending our image-based similarity metric to cover more words in the lexicon.

¹A part or whole of the work documented here has been published in [48]

To evaluate the effectiveness of our image-based model for measuring word-to-word relatedness, we selected three datasets widely used in the past:

Rubenstein and Goodenough (RG65) consists of 65 word pairs ranging from synonymy pairs (e.g., car - automobile) to completely unrelated terms (e.g., noon - string). The 65 noun pairs were annotated by 51 human subjects. All the nouns pairs are non-technical words scored using a scale from 0 (not-related) to 4 (perfect synonymy).

Miller-Charles (MC30) is a subset of the Rubenstein and Goodenough dataset, consisting of 30 word pairs, whose relatedness was rated by 38 human subjects, using a scale from 0 to 4.

WordSimilarity-353 (WS353), also known as Finkelstein-353, consists of 353 word pairs annotated by 13 human experts, on a scale from 0 (unrelated) to 10 (very closely related). The Miller-Charles set is a subset in the WordSimilarity-353 data set. Unlike the Miller-Charles data set, which consists only of single generic words, the dataset also includes proper names and technical terms, therefore posing an additional degree of difficulty for any relatedness metric.

5.3. Experiments

In our experiments, we seek answers to the following questions. First, what is the effectiveness of our image-based method in measuring word-to-word relatedness, as compared to existing text-based methods? Second, can our image-based method complement these text-based methods via a combination of their outputs ?

Note that as ImageNet is still a resource under development, not all word pairs in the datasets presented in section 5.2 are covered. To level the playing field, in our experiments we only select those pairs of words of which both words would appear as surface forms in the synsets of ImageNet with validated images. Our trimmed dataset consists of 19 word-pairs from the Miller-Charles dataset (**MC19**), 39 word-pairs from the Rubenstein-Goodenough dataset (**RG39**) and 160

word-pairs from the Word Similarity dataset (**WS160**) respectively. However, due to coverage issues, an anomaly exists in situations such as *monk* – *slave*, where both words may appear in single-candidate synsets, i.e., {monk,monastic} and {slave ant} respectively, but are represented using fundamentally different images (person vs animal). Arguably, allowing semantic relatedness comparison of such a word-pair is meaningless. To prevent this, we further constrain the selection of word pairs of which at least a pair of candidate synsets each representing a word in the pair belong to the same high level category, using the ImageNet Applicability Test algorithm, explained in Algorithm 2. Note that both selection steps are performed automatically, and thus the identification of the word pairs that can be used in conjunction with the image-based approach can be effectively applied to any dataset, regardless of size. From this, the further trimmed dataset consists of 10 word-pairs from the Miller-Charles dataset (**MC10**), 18 word-pairs from the Rubenstein-Goodenough dataset (**RG18**) and 56 word-pairs from the Word Similarity dataset (**WS56**).

Algorithm 2 ImageNet Applicability Test

```

Start :  $C_{ImageNet} = \{c_1, \dots, c_n\}$ ,  $W = \{w_1, w_2\}$ 
for each  $w_i$  in  $W$  do
     $S_i = \{\text{all synsets in ImageNet containing } w_i\}$ 
     $H_{S_i} = \{\text{all hypernyms of } s \text{ such that } s \in S_i\}$ 
     $C_i = \{H_{S_i} \cap C_{ImageNet}\}$ 
end for
if  $C_1 \cap C_2 \neq \emptyset$  then
    apply ImageNet as a metric for  $w_1, w_2$ 
end if

```

For each word in a pair, we randomly select 50 images from the validated image pool of its associated synset², and extract all the visual codewords from all such images, using the technique explained in section 3.3. Each image is first pre-processed to have a maximum side length of 300 pixels. Next, SIFT gray-scale descriptors are obtained by densely sampling the image on 20x20 overlapping patches spaced 10 pixels apart using a publicly available image-processing toolkit.³ K-means clustering is applied on a random subset of 10 million SIFT descriptors to derive a visual

²Note that a word may appear as surface forms across multiple synsets. In such cases, we randomly sample 50 images from each of the synsets

³<http://www.image-net.org/challenges/LSVRC/2011>

vocabulary of 1,000 codewords. Each descriptor is then quantized into a visual codeword by assigning it to the nearest cluster. As such, each image J can now be expressed as a vector $\langle tf_i.w_i \rangle$, where $i=1:1000$ and tf_i is the frequency of occurrence of visual codeword w_i in image J . For each synset, we sum the vectors of all 50 images and normalize each w_i by its total frequency in the synset.

Image Metric: Given a word pair w_i and w_j , let $S_i = \{v_k^i\}$ and $S_j = \{v_m^j\}$ be their set of candidate visual vectors respectively. Then, computing the semantic relatedness of two words amounts to finding the maximum visual relatedness between all the possible pairings of synsets representing both words, using the cosine similarity between the visual vectors of the synsets, given below. The dimensionality of the vector, n , is set to 1000, which is the size of the visual codeword vocabulary.

$$(13) \quad Sim_{img}(w_i, w_j) = \max_{v_k \in S_i, v_m \in S_j} \frac{\sum_{p=1}^n v_k^p v_m^p}{\sqrt{\sum_{p=1}^n (v_k^p)^2} \sqrt{\sum_{p=1}^n (v_m^p)^2}}$$

Text Metric: For a comparative study, we evaluate several knowledge-based methods, including Roget and WordNet Edges [37], H&S [35], L&C [47], J&C [39], LIN [54], RES [68], and two corpus-based methods Latent Semantic Analysis (LSA) [46] and Explicit Semantic Analysis (ESA) [30].

Combined Metric: In the combined setting, we attempt to integrate the output of our image-based metric with that of existing text-based metrics in a pairwise manner via two combination functions, which were previously noted for their effectiveness in Information Retrieval systems [29]. Specifically, we combine the text-based and image-based metrics by summing their relatedness figures (COMBSUM) and by calculating their F-measure (F) defined as the harmonic mean of the two input metrics. The combination functions are shown in Table 5.1. Intuitively, the function CombSUM

ID	Name	Combination Function
1	CombSUM	$M_1 + M_2$
2	F-measure	$(1 + \beta^2) \frac{M_1 M_2}{\beta^2 M_1 + M_2}$

TABLE 5.1. Pairwise combination functions of outputs from any two metrics M_1 and M_2 . We set $\beta = 0.5, 1, 1.5$ for three further variants of the F-measure function.

serves to augment evidence lacking in either scores. For F-measure, a lower β value would place a higher weight on M_1 while a higher β value would emphasize M_2 more. In our experiments, we set M_2 to be our image-based metric. Because the similarity scores are differently distributed across various methods, we apply a normalization step within each metric to assert the same lower and upper-bound prior to the combination:

$$(14) \quad Score_{norm} = (Score_{original} - Score_{min}) / (Score_{max} - Score_{min})$$

For each dataset and metric, we obtain the Spearman rank correlation of the automatically generated similarity scores with the ground-truths by human subjects.

5.4. Discussion

The results in Table 5.2 show that our image-based metric can be an effective metric on its own, scoring a competitive Spearman correlation of 0.669 on the MC19 dataset, and reasonably well-correlated (0.547) to human ratings on the RG39 dataset. Perhaps not surprisingly, these two datasets consists mainly of words such as *car*, *forest*, *bird*, *furnace* etc which are *picturable*, concrete entities that possess distinctive and unambiguous visual representations. Its performance, however, degrades on the WS160 dataset with a somewhat low correlation rating of 0.300, mainly due to presence of more broadly defined words lacking a visual identity (e.g. *equipment* in the word pair *phone* – *equipment*), or the word appears as a surface form in the synset but does not constitute a noun entity represented by a validated image pool in ImageNet (e.g. *glass* in the

synset {glass lizard, glass snake, joint snake}). After applying the ImageNet Applicability Test (Algorithm 2), the correlation figures, based on the image-based metric obtained from the second trimmed set of datasets MC10, RG18 and WS56, expectedly show greater improvements. In fact, constraining the MC10 dataset further to a subset consisting of synonymous/near-synonymous word pairs enables our image-based method to outperform all other metrics, suggesting its promise for use as a *high-precision, low coverage* metric for detecting picturable, synonymous word pairs.

Table 5.3 shows the correlation for the combined and single metric settings for the datasets MC10, RG18 and WS56. As a means to show the improvement of adding visual information to the text-based metrics, Figure 5.1 shows the differential correlation figures between the combined and single text-based metric correlation for the datasets MC10, RG18 and WS56.

Regardless of the performance of the individual image-based metric, the hybrid image-text approach improves over the standalone text-based metric in almost all cases, and this holds for both knowledge-based and corpus-based methods, with few exceptions (e.g WNE in WS56 under F-measure, $\beta = 1.5$). The RES metric benefits the most from the combinations e.g. it scores a differential correlation consistently over 0.5 in the MC10 dataset under all combination scenarios, probably because it presents only a rough gauge of similarity [65] that is now supplemented by visual cues. The combination function with the most consistent improvement in all scenarios is F-measure, $\beta = 0.5$, which relatively favors text-based methods more, and in a scenario where the image-based method is orthogonal can lead to state-of-the-art in all except in one case (MC10, H&S).

	MC19	RG39	WS160	MC10	RG18	WS56
WNE	0.625	0.722	0.432	0.846	0.867	0.482
H&S	0.780	0.794	0.434	0.883	0.775	0.453
J&C	0.770	0.816	0.421	0.685	0.828	0.454
L&C	0.625	0.722	0.409	0.846	0.867	0.515
LIN	0.650	0.737	0.403	0.685	0.820	0.496
RES	0.560	0.659	0.392	0.328	0.580	0.469
LSA	0.691	0.582	0.612	0.867	0.546	0.520
ESA	0.632	0.661	0.540	0.515	0.611	0.453
Image Metric	0.669	0.547	0.300	0.851	0.820	0.404

TABLE 5.2. Table showing Spearman correlation of similarity scores generated using different metrics with human judgements, repeated for each of the 6 trimmed datasets

	Text-based measures								Image metric
	WNE	H&S	J&C	L&C	LIN	RES	LSA	ESA	
MC10									
STANDALONE	0.846	0.883	0.685	0.846	0.685	0.328	0.867	0.515	0.851
SUM	0.879	0.927	0.830	0.855	0.806	0.842	0.915	0.842	
F(0.5)	0.855	0.867	0.745	0.879	0.830	0.855	0.891	0.782	
F(1)	0.855	0.855	0.806	0.855	0.842	0.891	0.927	0.782	
F(1.5)	0.855	0.855	0.830	0.855	0.855	0.867	0.915	0.782	
RG18									
STANDALONE	0.867	0.775	0.828	0.867	0.820	0.580	0.546	0.611	0.820
CombSUM	0.887	0.826	0.867	0.887	0.863	0.813	0.728	0.827	
F(0.5)	0.907	0.796	0.833	0.907	0.861	0.732	0.609	0.625	
F(1)	0.893	0.796	0.833	0.907	0.869	0.793	0.607	0.627	
F(1.5)	0.869	0.807	0.842	0.907	0.862	0.850	0.617	0.649	
WS56									
STANDALONE	0.482	0.453	0.454	0.515	0.496	0.469	0.520	0.453	0.404
CombSUM	0.457	0.474	0.471	0.507	0.523	0.524	0.538	0.440	
F(0.5)	0.515	0.561	0.473	0.620	0.565	0.553	0.563	0.469	
F(1)	0.453	0.583	0.513	0.546	0.520	0.570	0.588	0.475	
F(1.5)	0.444	0.578	0.530	0.470	0.479	0.508	0.582	0.485	

TABLE 5.3. Results obtained with individual knowledge-based and corpus-based text-based measures, with our image measure, and with two combination functions (COMBSUM and F, with 3 variants). The bold correlation numbers represents the highest among all metrics per text-based measure per dataset.

5.5. Further Investigation

5.5.1. Coverage Expansion and Scalability Effectiveness

We have demonstrated the promise of an image-based word relatedness metric through initial evaluation experiments on standard datasets. Since application coverage of this metric for

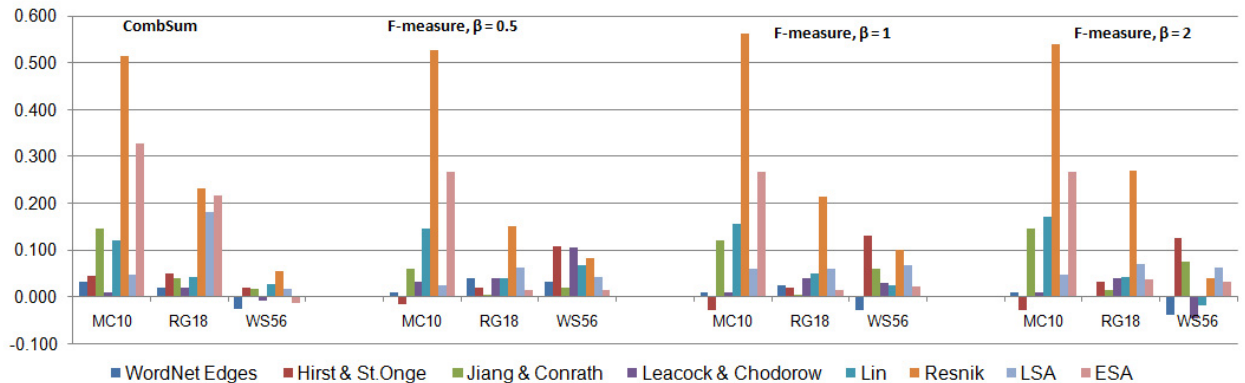


FIGURE 5.1. Graph showing differential correlation against different metrics, grouped by different datasets and combination functions. Any bar below zero indicates worse performance after combination with the image-based metric.

words in those datasets is limited, whether such a metric is simply ideological or feasible in practice requires deeper insights. In this section, there are two questions that we wish to address. First, can we integrate existing lexical resources into our image-based metric so that its application coverage can be improved? Second, in the light of such an improved coverage, can it still yield an accurate semantic relatedness measure between words that correlates well to human ratings, as evaluated on standard datasets?

An important assumption in using our image-based metric is that there is a direct correspondence between the words in a synset and the set of images illustrating that synset. The words in the synset form an equivalent semantic class where entries are semantically synonymous to each other, and each image can be used to pictorially describe the meaning of each of those words. The challenge involved in extending the coverage for more vocabulary implies that expansion of synsets is a necessary step.

This can be achieved in two ways. Firstly, we can relax the assumption that images associated with a synset are good visual descriptors only for the words in the synsets. Words in the associated gloss of the synset are necessarily related to the synset. By transitivity, the corresponding set of images is related to the words contained in the gloss as well. As an initial step, these gloss words are appended to the synset. Next, as previously done in work on word to word similarity [64], we also include words in the direct hypernyms and hyponyms of the target synset, as well

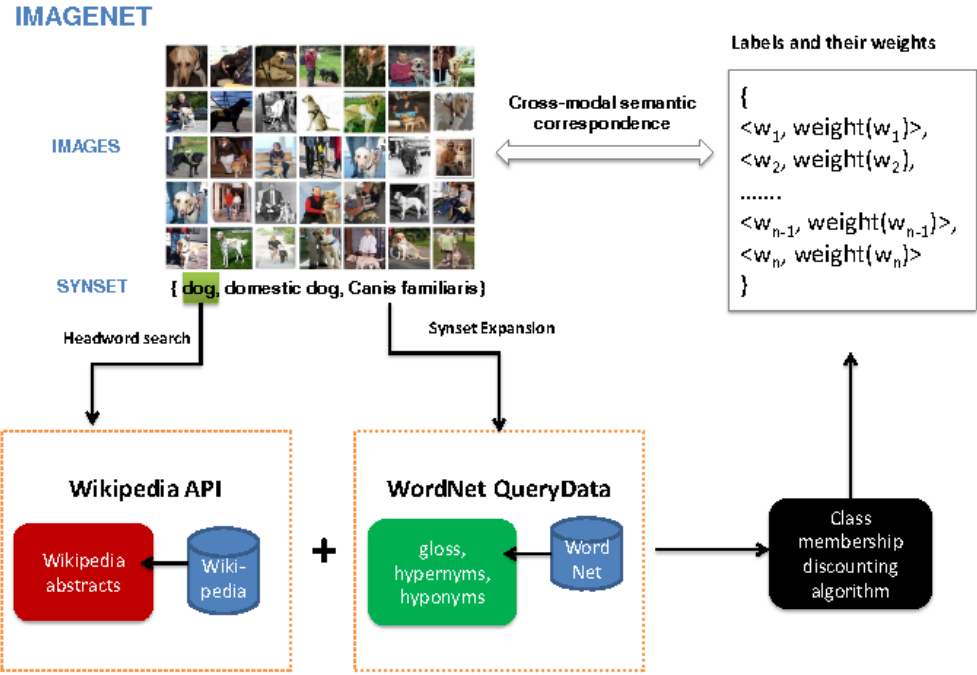


FIGURE 5.2. Schematic diagram depicting our proposed system

as their glosses, in the expanded bag of words. Note that the decision not to pursue other semantic links or those hypernyms/hyponyms not directly related to the synset is motivated on grounds of decreasing semantic relatedness between those candidate words and words in the synset [11], and increased computational complexity due mainly to the large number of associated images.

Secondly, we perform the mapping of synsets to entries in another resource, namely Wikipedia, so that lexical coverage may be increased using another lexical resource both supplementing and complementing Wordnet. Alternatively, efforts to pursue additional visual coverage can also be performed through images associated with named entities and concrete nouns in Wikipedia pages. However, it is not the focus of the work here, and shall be left to future exploits. Using the head word of the target synset, we query the Wikipedia database using an API⁴ that returns a list of related Wikipedia articles that are not necessarily disambiguated, but are related to the head word.

⁴<http://en.wikipedia.org/w/api.php>

	#Word pairs covered			Avg #synset pairings/word pair		
	MC30	RG65	WS353	MC30	RG65	WS353
Synsets	19	39	158	161	106	151
+Glosses	23	51	269	8654	4437	2403
+Hype/Hypo and glosses	24	54	278	21841	10751	5869
+Wikipedia abstracts	29	62	337	32428	16979	29832

TABLE 5.4. Table showing the word pairs coverage statistics and synset pairings per word pair for each of the three standard evaluation datasets. Each number reported is cumulative i.e. +glosses is the expansion method using glosses + synset

The set of words in the abstracts for each of these Wikipedia articles are again included into the expanded bag of words. Our proposed system is shown in Figure 5.2. Lemmatization and stopword removal are both applied in our lexical expansion process using Wordnet and Wikipedia.

Table 5.4 shows the improvement in words coverage when each of the methods outlined above is employed. As observed, there is a significant increase in words coverage for all datasets (+4 word pairs for MC30, +12 pairs for RG65 and +111 pairs for WS353) when all words in the associated gloss of the target synset are considered. Further addition of words from the direct hypernyms and hyponyms and their glosses improves coverage marginally (+1 word pairs for MC30, +3 word pairs for RG65, and +9 word pairs for WS353). A plausible explanation for this observation is that words in direct hypernyms/hyponyms and their glosses tend to have more overlap with those in the target synset and gloss, due to the structure of the concept hierarchy. Finally, the increase in words coverage is also significant (+5 word pairs for MC30, +8 word pairs for RG 65 and +59 word pairs for WS353). Since Wikipedia abstracts are more verbose in their discussion about any subject matter, this is somewhat expected.

An obvious drawback of coverage expansion via the generation of “bag of words” is that some words may be more related than others to the target synset images. A possible way to expand coverage without losing accuracy is to model graded membership of words belonging to a synset, such that the final image-based metric would only consider words that are reasonably related to the original target synset. This task can be perceived differently and solved using different approaches. For instance, through Chi square (χ^2) statistics, we can compute the (in)dependence

relation between any given synset s and a word w using its frequency counts, and any word with $x^2(s, w)$ below a threshold can be discarded from the synset. Alternatively, the association strength of a given word with the target synset can be estimated using pointwise mutual information (PMI) between the word and each of the words in the synset using corpus evidence, and averaged over the size of the synset. As before, words below a threshold would be discounted and removed. Investigation of such term weighting or class membership discounting techniques are well-discussed in the literature [87]. In our work, we employ the use of the simple, yet effective, term-weighting method *tfidf*, with a focus to develop a lightweight system capable of scaling up efficiently.

We are interested to perform further experiments on the WS353 dataset, since it is the most challenging of the three datasets. The main intention of our study is to perform filtering of irrelevant words from each synset so that coverage can be maintained while reducing the average synset pairings. Specifically, *tfidf* is computed for each word in the expanded set of words for each synset. Next, we drop the lowest $k\%$ of words in each synset after the *tfidf scores* are sorted in reverse order. We apply this method in cases where coverage has seen a sudden increase i.e. the expansion of synsets by using its gloss, henceforth referred to as **LexExpGloss**, and the expansion of synsets using all lexical resources (Wordnet and Wikipedia), henceforth referred to as **LexExpAll**. Table 5.5 and 5.6 shows the results of our approach. In each case, a similar trend is observed. As more words are dropped, the number of synset pairings required for the image-based metric decreases. As observed, the maximum word pairs coverage for LexExpGloss is 269 and stays constant up to dropping 30% of the words in each synset. Likewise, the maximum word pairs coverage for WordNetGlosses stays at 337 up to dropping 60% of the words in each synset. Not surprisingly, as we expand more words using more lexical resources, the more we can drop. Since our intention is to maximize coverage and minimize synset comparisons, in an effort to derive an efficient, maximum-coverage similarity metric, we performed re-runs of the experiment on these two subsets of LexExpGloss and WordNetGlosses respectively. These posthoc analysis results are shown in Tables 5.8 and 5.9 respectively.

% of dropped words	#Word pairs covered	Avg #synset pairings/word pair
0%	269	2404
10%	269	904
20%	269	453
30%	269	209
40%	268	104
50%	267	64
60%	245	35
70%	200	23
80%	138	15
90%	102	12

TABLE 5.5. **LexExpGloss** : Table showing the word pairs coverage statistics and synset pairings per word pair for WS353 dataset, plotted against the percentage of dropped words (reverse sorted using *tfidf*). The expanded bag of words is formed by synsets + glosses only

% of dropped words	#Word pairs covered	Avg #synset pairings/word pair
0%	337	29832
10%	337	17091
20%	337	9343
30%	337	5350
40%	337	3041
50%	337	1849
60%	337	1186
70%	331	668
80%	295	400
90%	256	166

TABLE 5.6. **LexExpAll** : Table showing the word pairs coverage statistics and synset pairings per word pair for WS353 dataset, plotted against the percentage of dropped words (reverse sorted using *tfidf*). The expanded bag of words is formed by synsets + glosses + hype/hypo and glosses + Wikipedia abstracts

As it turns out, adding more words to each synset causes the probability of a given word associated with more than one synset to increase. Recall from Equation 13 that the semantic similarity between a pair of words is defined as the maximum visual similarity computed over all pairs of synsets in a complete bipartite graph, where each partition of the graph contains candidate synsets for each word in the word pair. A small increase in number of words coverage tend to




Image Difference Detection		
Image 1	Image 2	Difference Image
		

TABLE 5.7. Image difference detection using the Absolute Error (AE) metric in ImageMagick. Given two images, the algorithm gives an absolute count of pixels that are different, with the fuzz factor set at 20%. Only pixels changed by more than the fuzz factor are considered different. These different pixels are marked in red in the newly composed 'difference' image formed by overlaying the two images.

cause the number of synset pairings between a word pair to increase significantly. This evidence is provided in Table 5.4 in terms of the number of synset pairings per word pair. Computing the visual similarity between a given pair of synsets takes approximately 1 second, which amounts to 116 days just to compute the similarity between 337 word pairs in the WordSim 353 dataset⁵. Fortunately, this step needs to be computed only once and stored in hashes⁶, and future retrieval of visual similarity between the same synset pairing takes constant time. Regardless, assuming that accuracy must be maintained, using the appropriate term weighting techniques to retain only the most relevant keywords per synset is extremely important to scalability of the proposed image-based metric. Future work would investigate the applicability of other measures besides *tfidf* for keyword filtering.

5.5.2. Informed Baselines

An interesting question to ask is whether we can adopt an effective visual-based relatedness metric using other resources and computer vision techniques that are readily available and equally efficient. To this end, we implement two baselines. For the first baseline, instead of using images in ImageNet, we use the Google Image Search API to download the top 50 images for each word in a

⁵Assuming non-parallelized, Perl implementation on a 2.83 GHz quad-core CPU with 8GB RAM. However, using full matrix multiplication for computing synset similarity in MATLAB decreases the time taken significantly to a couple of hours.

⁶Still, hashes required to store the visual similarity for 8683x8683 synset pairings amounts to 8GB alone

given word pair. Since the image search does not perform distinction of word senses e.g. a search using “apple” would return images of the Apple company logo and the fruit apple together, we have no easy way to automatically classify them. Doing so also results in additional work beyond constructing a simple and efficient baseline as we would have liked. Therefore, the set of 50 images are sense-lumped and SIFT-processed together. Apart from this, other aspects of our image-based metric remain the same. The second baseline operates on the images downloaded for the first baseline, but instead of using SIFT as our image descriptor and using the metric in Equation 13, we gauge the similarity between a given word pair as the maximum pairwise similarity between the two sets of Google images downloaded for each of the two words, where visual similarity between any two images is performed using a publicly available toolkit counting the normalized number of unchanged pixels⁷. This approach is illustrated in Table 5.7. Intuitively, the first baseline represents the advantage (if any) of a manually constructed database such as ImageNet over simply retrieving images from a major search engine like Google. In the same spirit, the second baseline represents the difference of using a sophisticated image descriptor i.e. SIFT over a simple method i.e. Absolute Error in pixels in the implementation of our image-based word relatedness metric.

Given that our coverage for word pairs has been increased using the system proposed in Figure 5.2, it is important to validate our image-based metric through its accuracy on the larger subsets of WordSim 353 data. Tables 5.8 and 5.9 show the results of using the standalone text-based metrics, our novel image-based metric and their hybrids.

First, simply augmenting synsets with gloss words and dropping up to 30% increases coverage, yet retaining competitiveness of our visual metric, which scores a Spearman correlation of 0.342. Using the hybrid metrics results in improvement in 81% (26/32) of the times when compared against the standalone text-based measures. By expanding coverage even further using Wordnet hypernyms/hyponyms and Wikipedia, we see an even greater improvement in accuracy

⁷<http://www.imagemagick.org/script/index.php>

figures. As it turns out, our standalone image-based metric scoring a correlation 0.345 even outperforms three of the text-based measures (WNE (0.335), J&C (0.328) and L&C (0.328)). Furthermore, employing the hybrid metrics now results in 97% (31/32) of the times in comparisons with corresponding text-based metrics.

Second, the above results confirm the effectiveness of a simple metric like *tfidf* in improving *both* coverage and accuracy of our image-based metric. Specifically, it can be easily applied within the expanded dataset itself, noting that our *tf* and *idf* are derived using the synsets as pseudo-documents. A plausible reason for *tfidf* being so effective is due to the redundancy of synset headwords in synergism with other content words featuring lower *tf* but high *idf*. Consider again our running synset example i.e. {dog, domestic dog, Canis familiaris}. The headword “dog” appears at least once in each of the successive expansion steps, through its direct hyponym :

“puppy (a young dog)”

and also at least once in the Wikipedia abstract :

The domestic dog (Canis lupus familiaris), is a subspecies of the gray wolf (Canis lupus), a member of the Canidae family of the mammalian order Carnivora).

Though having a low *idf*, its continued appearance in successive expansion steps leads to a compensating effect in the overall reranking. On the other hand, any other word that is reranked highly but having a low *tf* must be rare (high *idf*) to the point that its association with the synset be retained e.g. “Canis”.

Third, the benefits of using clean annotations from a manually constructed image database in ImageNet bears great significance in the implementation of our visual metric. As seen, the maximum difference between visual metrics constructed using ImageNet and Google images is close to a Spearman correlation of 0.300 (Table 5.9 : Image metric (0.345) vs Baseline (Google+SIFT) (0.079)). This difference is enormous in the semantic relatedness task where systems outperforming each other by even 0.100 or less are publishable results [34]. Our choice of using SIFT turns out to be a good choice. In the same vein, using a good visual descriptor over a simple baseline results in consistent advantage, as seen in the two baselines Google+SIFT vs Google+AE in Tables 5.8 and 5.9.

	Text-based measures								Image metric
	WNE	H&S	J&C	L&C	LIN	RES	LSA	ESA	
WS353									
STANDALONE	0.365	0.389	0.356	0.356	0.384	0.388	0.611	0.481	0.342
CombSUM	0.396	0.430	0.443	0.419	0.428	0.416	0.556	0.465	
F(0.5)	0.382	0.401	0.385	0.380	0.389	0.396	0.585	0.504	
F(1)	0.376	0.401	0.385	0.380	0.388	0.392	0.583	0.504	
F(1.5)	0.376	0.398	0.387	0.382	0.382	0.387	0.583	0.504	
<hr/>									
Baseline (Google+SIFT)					0.126				
Baseline (Google+AE)					0.088				

TABLE 5.8. Results obtained for WS353 dataset by dropping the lowest 30% (reverse sorted by *tfidf*) of expanded set of words formed by synsets/glosses. Figures in bold represent hybrid image-text metrics that are score better correlation than the individual standalone text-based metrics

	Text-based measures								Image metric
	WNE	H&S	J&C	L&C	LIN	RES	LSA	ESA	
WS353									
STANDALONE	0.335	0.361	0.328	0.328	0.358	0.364	0.612	0.514	0.345
CombSUM	0.385	0.438	0.453	0.402	0.410	0.424	0.639	0.480	
F(0.5)	0.366	0.375	0.356	0.358	0.368	0.379	0.626	0.516	
F(1)	0.371	0.375	0.356	0.363	0.367	0.381	0.626	0.515	
F(1.5)	0.372	0.375	0.356	0.368	0.368	0.380	0.628	0.515	
Baseline (Google+SIFT)					0.079				
Baseline (Google+AE)					0.056				

TABLE 5.9. Results obtained for WS353 dataset by dropping the lowest 60% (reverse sorted by *tfidf*) of expanded set of words formed by synsets/glosses/hype/hypo/Wikipedia. Figures in bold represent hybrid image-text metrics that are score better correlation than the individual standalone text-based metrics.

5.5.3. Picturability Study

In light of the effectiveness of our proposed image-based metric, it is helpful to characterize as an additional insight how and when we can employ such a metric. Specifically, we are interested to see how different data sets modeled on a criteria of choice correlates with human ratings when evaluated using the visual metric. The *picturability* of a given word [90] is defined as the ability of a human being to draw or find a good image to substitute for the word itself, given its meaning.

Conversely, a word's picturability can also be defined as the ease with which a human being is able to guess it correctly when shown an image closely associated to its meaning. Following [90], we employ a picturability logistic regression model based on raw counts of Web pages and Web images retrieved when the target word is queried using a search engine. In this model, training was performed on a manually-labeled set of 500 randomly chosen words from a large vocabulary, where 5 human annotators were requested to independently label each word as picturable ($y = 1$) or non-picturable ($y = 0$). The model is shown in Equation 15.

$$(15) \quad p(y = 1|x) = \frac{1}{1 + \exp(-(2.78x + 15.40))}$$

where

$$(16) \quad x = \log((c_1 + 10^{-9})/(c_2 + 10^{-9}))$$

Word Pair	Average Human Rating
secretary, senate	5.06
Harvard, Yale	8.13
lawyer, evidence	6.69
medium, trade	3.88
medium, gain	2.88
calculation, computation	8.44
profit, warn	3.88
investor, earn	7.13

TABLE 5.10. Table showing word pairs with their averaged human ratings of similarity from S2

achieving a score of 0.99, indicating that it is highly picturable. In another case, the word 'dividend' brought 49,300,000 Web hits and 220,000 Image hits, resulting in a score of 0.59, suggesting it is not that picturable. Here, we performed a picturability analysis of all 447 words formed by the union of the vocabulary of the three datasets, where the results are shown in Figure 5.3. Though unreadable, it can be easily inferred from the scatter plot that a vast majority of the union vocabulary are scored above the 0.9 threshold, and somewhat considered as *picturable* using the metric proposed. To observe the performance of our visual metric in the face of varying picturability of words, we divide the covered WS353 dataset into two disjoint sets, namely, S_1 , which contain 277 pairs of words, each of which has picturability score of more than or equal 0.90, and S_2 , which contains 8 pairs of words, each with a picturability score of less than 0.90. The latter is shown in Table 5.10, while Spearman correlation of the text-based and image-based methods with human judgments on the same dataset is shown in Table 5.11. Not surprisingly, the performance of our image-based metric for low picturable word pairs (S_2) is much inferior to those of other text-based methods. However, note that these word pairs only constitute a small portion of the dataset (8/337). When evaluating our image-based metric on S_1 , it is clear that such a metric is competitive with existing text-based metrics.

5.6. Related Work

Research on images and texts in cognitive science are grounded on perceptual representation and understanding of words, while most models in computer vision are tasked with discovering correlations between image features and words to enable applications such as automatic image annotation, search and retrieval. Even so, to our knowledge, very few work has applied visual representation models to the semantic relatedness task in NLP. Recently, some attention has been given to modelling synergistic relationships between the semantics of words and images [49, 10]. The research that is most closely related to ours is the work of [27], where it has been shown that it is possible to combine visual representations of word meanings into a joint bimodal representation constructed by using probabilistic generative latent topic models. Unlike our approach, however, [27] relied on a news corpus where images and words in a document are assumed to be generated by a set of latent topics, rather than a lexical resource such as ImageNet. Moreover,

in their work, no attempt has been made to evaluate the image-based models independently, or to combine image models with previously proposed knowledge-based and corpus-based measures of relatedness. While they provided a proof-of-concept that using the visual modality leads to an improvement over their purely text-based model (an increase of Spearman correlation of 0.071 on a 254-pairs subset of WordSim353 dataset), no attempt has been made to evaluate the image-based models independently, or to combine image models with previously proposed knowledge-based and corpus-based measures of relatedness.

CHAPTER 6

“MEN ARE FROM MARS AND VENUS, SO ARE WOMEN”: BUILDING A MULTIMODAL SEMANTIC SPACE USING WORDS AND IMAGES

The love between the Venusians and Martians was magical. They delighted in being together, doing things together, and sharing together. Though from different worlds, they revelled in their differences. They spent months learning about each other, exploring and appreciating their different needs, preferences, and behaviour patterns. For years they lived together in love and harmony.

John Gray

Regardless of the differences between language and vision, research performed at the joint language-vision interface has started to grow due to a number of favorable factors. The construction of a joint semantic space connecting words with images is a task that is very challenging but at the same time highly desirable for a number of applications. We posit that an image, like a word, presents meaning that can be harvested under a generic framework. The proposition is challenging due to the well-known *semantic gap* problem, first put forward by [88], which indicates that it is difficult to identify meaningful entities in an image by exclusively using its low level features like color and texture. The focus of this chapter, however, is not to predict which objects are present in an image, as in entity recognition, but rather to perform an analysis of the complete image to derive its relation to one or more semantic concepts. Specifically, we test the effectiveness of our multi-modal semantic space model by evaluating it using a cross-modal semantic relatedness framework, for which we provide the motivation in the next section.

6.1. Motivation

Traditionally, a large body of research in natural language processing has focused on formalizing word meanings. Several resources developed to date (e.g., WordNet [63]) have enabled a systematic encoding of the semantics of words and exemplify their usage in different linguistic frameworks. As a result of this formalization, computing semantic relatedness between words has

been possible and has been used in applications such as information extraction and retrieval, query reformulation, word sense disambiguation, plagiarism detection and textual entailment.

In contrast, while research has shown that the human cognitive system is sensitive to visual information and incorporating a dual linguistic-and-pictorial representation of information can actually enhance knowledge acquisition [66], the *meaning* of an image in isolation is not well-defined and it is mostly task-specific. A given image, for instance, may be simultaneously labeled by a set of words using an automatic image annotation algorithm, or classified under a different set of semantic tags in the image classification task, or simply draw its meaning from a few representative regions following image segmentation performed in an object localization framework.

Given that word meanings can be acquired and disambiguated using dictionaries, we can perhaps express the meaning of an image in terms of the words that can be suitably used to describe it. Specifically, we are interested to bridge the *semantic gap* [75] between words and images by exploring ways to harvest the information extracted from visual data in a general framework. While a large body of work has focused on measuring the semantic similarity of words,(e.g., [62]), or the similarity between images based on image content [31, 69, 70], very few researchers have considered the measure of semantic relatedness¹ between words and images.

But, how exactly is an image related to a given word? In reality, quantification of such a cross-modal semantic relation is impossible without supplying it with a proper definition. Our work seeks to address this challenge by constructing a standard evaluation framework to derive a semantic relatedness metric for arbitrary pairs of words and images. In this chapter, we explore methods to build a representation model consisting of a joint semantic space of images and words by combining techniques widely adopted in computer vision and natural language processing, and we evaluate the hypothesis that we can automatically derive a semantic relatedness score using this joint semantic space.

¹In this chapter, we are concerned with semantic *relatedness*, which is a more general concept than semantic *similarity*. Similarity is concerned with entities related by virtues of their likeness, e.g., *bank-trust company*, but dissimilar entities may also be related, e.g., *hot-cold*. A full treatment of the topic can be found in [11].

Importantly, we acknowledge that it is significantly harder to decode the semantics of an image, as its interpretation relies on a subjective and perceptual understanding of its visual components [7]. Despite this challenge, we believe this is a worthy research direction, as many important problems can benefit from the association of image content in relation to word meanings, such as automatic image annotation, image retrieval and classification (e.g., [50]) as well as tasks in the domains of text-to-image synthesis, image harvesting and augmentative and alternative communication. Arguably, this is also true for tasks that address the reverse direction of text-to-image synthesis [59, 52, 16].

The chapter² proceeds as follows. We first provide a brief overview of the semantic vectorial models used in natural language processing. In section 6.3, we provide motivations and details for implementing the cross-modal semantic relatedness evaluation framework. Next, we conduct an empirical evaluation of our hypothesis using the bag-of-visual codewords and semantic vector models to build a joint semantic space of words and images. followed by discussions in section 6.6. Finally, we experiment with augmented multimodal spaces, constructed using visual attributes, to observe if such an enriched version of semantic representation would further lead to improvements in correlation ratings. An overview of related work is provided in section 6.8.

6.2. Semantic Vector Models

The underlying idea behind semantic vector models is that concepts can be represented as points in a mathematical space, and this representation is learned from a collection of documents such that concepts related in their meanings are near to one another in that space. In the past, semantic vector models have been widely adopted by natural language processing researchers for tasks ranging from information retrieval and lexical acquisition, to word sense disambiguation and document segmentation. Several variants have been proposed, including the original vector space model [72] and the Latent Semantic Analysis [46]. Generally, vector models are attractive because they can be constructed using unsupervised methods of distributional corpus analysis and assume

²A part or whole of the work documented here has been published in [49]

little language-specific requirements as long as texts can be reliably tokenized. Furthermore, various studies [42] have shown that by using collaborative, distributive memory units to represent semantic vectors, a closer correspondence to human cognition can be achieved.

While vector-space models typically require nontrivial algebraic machinery, reducing dimensions is often key to uncover the hidden (latent) features of the terms distribution in the corpus, and to circumvent the sparseness issue. There are a number of methods that have been developed to reduce dimensions – see e.g., [85] for an overview. Here, we briefly describe one commonly used technique, namely the Latent Semantic Analysis (LSA), noted for its effectiveness in previous works for reducing dimensions.

In LSA, term co-occurrences in a corpus are captured by means of a dimensionality reduction operated by a **Singular Value Decomposition (SVD)** on the term-by-document matrix \mathbf{T} representing the corpus. SVD is a well-known operation in linear algebra, which can be applied to any rectangular matrix in order to find correlations among its rows and columns. SVD decomposes the term-by-document matrix \mathbf{T} into three matrices $\mathbf{T} = \mathbf{U}\Sigma_k\mathbf{V}^T$ where Σ_k is the diagonal $k \times k$ matrix containing the singular k values of \mathbf{T} , $\sigma_1 \geq \sigma_2 \geq \dots \geq \sigma_k$ and \mathbf{U} and \mathbf{V} are column-orthogonal matrices. When the three matrices are multiplied together the original term-by-document matrix is re-composed. Typically we can choose $k' \ll k$ obtaining the approximation $\mathbf{T} \simeq \mathbf{U}\Sigma_{k'}\mathbf{V}^T$.

6.3. Semantic Relatedness between Words and Images

Although the bag of visual codewords 3.3 has been extensively used in image classification and retrieval tasks, and vector-space models are well explored in natural language processing, there has been little connection between the two streams of research. Specifically, to our knowledge, there is no research work that combines the two techniques to model multimodal meaning relatedness. Since we are exploring new grounds, it is important to clarify what we mean by computing the semantic relatedness between a word and an image, and how the nature of this task impacts our hypothesis. The assumptions below are necessary to validate our findings:

- (1) Computing semantic relatedness between a word and an image involves comparing the concepts invoked by the word and the salient objects in the image as well as their interaction. This goes beyond simply identifying the presence or absence of specific objects indicated by a given word. For instance, we expect a degree of relatedness between an image showing a soccer ball and the word “jersey,” since both invoke concepts like {sports, soccer, teamwork} and so on.
- (2) The semantics of an image is dependent on the focus, size and position of distinct objects identified through image segmentation. During labeling, we expect this segmentation to be performed implicitly by the annotators. Although it is possible to focus one’s attention on specific objects via bounding boxes, we are interested to harvest the meaning of an image using a holistic approach.
- (3) In the case of measuring the relatedness of a word that has multiple senses with a given image, humans are naturally inclined to choose the sense that provides the highest relatedness inside the pair. For example, an image of a river bank expectedly calls upon the “river bank” sense of the word “bank” (and not “financial bank” or other alternative word senses).
- (4) A degree of semantic relatedness can exist between any arbitrary word and image, on a scale ranging from being totally unrelated to perfectly synonymous with each other. This is trivially true, as the same property holds when measuring similarity between words and texts.

Next, we evaluate our hypothesis that we can measure the relatedness between a word and an image empirically, using a parallel corpus of words and images as our dataset.

Joint Semantic Space of Words and Images		
Synsets		167
Images		230,864
Words		1144
	Nouns	783
	Verbs	140
	Adjectives	221
Image:Words ratio		202:1

TABLE 6.1. A table showing statistical information on our joint semantic space model

6.4. Dataset

As before, we turn to ImageNet 3.2. Besides the motivations for choosing this resource as listed in section 5.1, we believe we can provide an additional utility to ImageNet through this research work. Although we can search for relevant images using keywords in ImageNet, there is currently no method to query it in the reverse direction. Given a test image, we must search through millions of images in the database to find the most similar image and its corresponding synset. A joint semantic model can hopefully augment this shortcoming by allowing queries to be made in both directions.

For our experiments, we randomly select 167 synsets³ from ImageNet, covering a wide range of concepts such as plants, mammals, fish, tools, vehicles etc. We perform a simple pre-processing step using Tree Tagger [73] and extract only the nouns. Multiwords are explicitly recognized as collocations or named entities in the synset. Not considering part-of-speech distinctions, the vocabulary for synset words is 352. The vocabulary for gloss words is 777. The shared vocabulary between them is 251. Table 6.1 shows the statistical information on our joint semantic space model.

There are a total of 230,864 images associated with the 167 synsets, with an average of 1383 images per synset. We randomly select an image for each synset, thus obtaining a set of 167 test images in total. The technique explained in Section 3.3 is used to generate visual codewords

³Not all synsets in ImageNet are annotated with images. We obtain our dataset from the Spring 2010 version of ImageNet built around Wordnet 3.0.

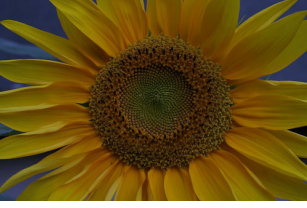


		
Synset {sunflower, helianthus}	Synset {oxygen-mask}	Synset {submarine, pigboat, sub, U-boat}
Gloss any plant of the genus Helianthus having large flower heads with dark disk florets and showy yellow rays	Gloss a breathing device that is placed over the mouth and nose; supplies oxygen from an attached storage tank	Gloss a submersible warship usually armed with torpedoes
Relatedness Scores color (5.13) dog (0.53) floret (6.53) flower (9.67) freshwater (2.40) hair (1.00) garden (6.60) head (3.80) plant (8.47) ray (3.67) sunflower (9.80) reed (2.27)	Relatedness Scores basketball (0.20) central (1.53) device (5.47) family (0.80) iron-tree (0.47) mouth (5.13) oxygen-mask (7.73) tank (4.47) storage (3.07) supply (5.20) nose (6.20) time (1.13)	Relatedness Scores africa (0.80) brass (1.73) door (1.67) good (2.40) pacific (2.40) pigboat (6.47) sub (8.20) submarine (9.67) tail (0.93) torpedo (7.60) u-boat (7.47) warship (8.73)

TABLE 6.2. A sample of test images with their synset words and glosses : The number in parenthesis represents the numerical association of the word with the image (0-10). Human annotations reveal different degree of semantic relatedness between the image and words in the synset or gloss.

for each image in this dataset.⁴ Each image is first pre-processed to have a maximum side length of 300 pixels. Next, SIFT descriptors are obtained by densely sampling the image on 20x20 overlapping patches spaced 10 pixels apart. K-means clustering is applied on a random subset of 10 million SIFT descriptors to derive a visual vocabulary of 1,000 codewords. Each descriptor is then quantized into a visual codeword by assigning it to the nearest cluster.

To create the gold-standard relatedness annotation, for each test image, six nouns are randomly selected from its associated synset and gloss words, and six other nouns are again randomly selected from the shared vocabulary words.⁵ In all, we have $167 \times 12 = 2004$ word-image pairs as our test dataset. Similar to previous word similarity evaluations [62], we ask human annotators to rate each pair on a scale of 0 to 10 to indicate their degree of semantic relatedness using the evaluation framework outlined below, with 0 being totally unrelated and 10 being perfectly synonymous

⁴For our experiments, we obtained the visual codewords computed a priori from ImageNet. Test images are not used to construct the model

⁵12 data points are generally considered sufficient for reliable correlation measures (Vania Kovic, p.c.).

with each other. To ensure quality ratings, for each word-image pair we used 15 annotators from Amazon Mechanical Turk.⁶ Finally, the average of all 15 annotations for each word-image pair is taken as its gold-standard relatedness score⁷. Note that only the pairs of images and words are provided to the annotators, and not their synsets and gloss definitions.

The set of standard criteria underlying the cross-modal similarity evaluation framework shown here is inspired by the semantic relations defined in Wordnet. These criteria were provided to the human annotators, to help them decide whether a word and an image are related to each other.

- (1) **Instance of itself:** Does the image contain an entity that is represented by the word itself (e.g. an image of “Obama” vs the word “Obama”) ?
- (2) **Member-of Relation:** Does the image contain an entity that is a member of the class suggested by the word or vice versa (e.g. an image of an “apple” vs the word “fruits”) ?
- (3) **Part-of Relation:** Does the image contain an entity that is a part of a larger entity represented by the word or vice versa (e.g. an image of a “tree” vs the word “forest”) ?
- (4) **Semantically Related:** Do both the word and the image suggest concepts that are related (e.g. an image of troops at war vs the word “peace”) ?
- (5) **Semantically Close:** Do both the word and the image suggest concepts that are not only related but also close in meaning? (e.g. an image of troops at war vs the word “gun”) ?

Criterion (1) basically tests for synonym relation. Criteria (2) and (3) are modeled after the hyponym-hypernym and meronym-holonym relations in WordNet, which are prevalent among

⁶We only allowed annotators with an approval rating of 97% or higher. Here, we expect some variance in the degree of relatedness between the candidate words and images, hence annotations marked with all 10s or 0s are discarded due to lack of distinctions in similarity relatedness

⁷Annotation guidelines and dataset can be downloaded at <http://lit.csci.unt.edu/index.php/Downloads>

nouns. Note that none of the criteria is preemptive over the others. Rather, we provide these criteria as guidelines in a *subjective* evaluation framework, similar to the word semantic similarity task in [62]. Importantly, criterion (4) models dissimilar but related concepts, or any other relation that indicates frequent association, while criterion (5) serves to provide additional distinction for pairs of words and images on a higher level of relatedness toward similarity. In Table 6.2, we show sample images from our test dataset, along with the annotations provided by the human annotators.

6.5. Experiments

Following [21], who argued that word meanings are graded over their senses, we believe that the meaning of an image is not limited to a set of “best fitting” tags, but rather it exists as a distribution over arbitrary words with varying degrees of association. Specifically, the focus of our experiments is to investigate the correlation between automatic measures of such relatedness scores with respect to human judgments.

To construct the joint semantic space of words and images, we use the SVD described in Section 6.2 to reduce the number of dimensions. To build each model, we use the 167 synsets from ImageNet and their associated images (minus the held out test data), hence accounting for 167 latent dimensions. We first represent the synsets as a collection of documents D , each document containing visual codewords used to describe their associated images as well as textual words extracted from their gloss and synset words. Thus, computing a cross-modal relatedness distance amounts to comparing the cosine similarity of vectors representing an image to the vector representing a word in the term-document vector space. Note that, unlike textual words, an image is represented by multiple visual codewords. Prior to computing the actual cosine distance, we perform a weighted addition of vectors representing each visual codeword for that image.

To illustrate, consider a single document d_i , representing the synset “snail,” which consists of $\{cw0, cw555, cw23, cw124, cw876, snail, freshwater, mollusk, spiral, shell\}$, where cwX represents a particular visual codeword indexed from 0-999⁸, and the textual words are nouns extracted from the associated synset and gloss. Given a test image I , it can be expressed as a bag of visual codewords $\{cw_1, \dots, cw_k\}$. We first represent each visual codeword in I as a

⁸For simplicity, we only show the top 5 visual codewords

vector of length $|D|$ using term-frequency inverse-document-frequency (*tfidf*) weighting, e.g., $cw_k = \langle 0.4 * d_1, 0.2 * d_2, \dots, 0.9 * d_m \rangle$, where $m=167$, and perform an addition of k such vectors to form a final vector v_i . To measure the semantic relatedness between image I and a word w , e.g., “snail,” we simply compute the cosine similarity between v_i and v_w , where v_w is also a vector of length $|D|$ calculated using *tfidf*.

This work seeks answers to the following questions. First, what is the relation between the discriminability of the visual codewords and their ability to capture semantic relatedness between a word and an image, as compared to the gold-standard annotation by humans? Second, given the unbalanced dataset of images and words, can we use a relatively small number of visual codewords to derive such semantic relatedness measures reliably? Third, what is the efficiency of an unsupervised vector semantic model in measuring such relatedness, and is it applicable to large datasets?

Analogous to text-retrieval methods, we measure the discriminability of the visual codewords using two weighting factors. The first is *term-frequency (tf)*, which measures the number of times a codeword appears in all images for a particular synset, while the second, *image-term-frequency (itf)*, captures the number of images using the codeword in a synset. For the two weighting schemes, we apply normalization by using the total number of codewords for a synset (for *tf* weighting) and the total number of images in a synset (for *itf* weighting).

We are interested to quantify the relatedness for pairs of words and images under two scenarios. By ranking the 12 words associated with an image in reverse order of their relatedness to the image, we can determine the ability of our models to identify the most related words for a given image (**image-centered**). In the second scenario, we measure the relatedness of words and images regardless of the synset they belong to, thus evaluating the ability of our methods to capture the relatedness between any word and any image. This allows us to capture the correlation in an (**arbitrary-image**) scenario. For the evaluations, we use the Spearman’s Rank correlation.

To place our results in perspective, we implemented two baselines and an upper bound for each of the two scenarios above. The *Random* baseline randomly assigns ratings to each word-image pair on the same 0 to 10 scale, and then measures the correlation to the human gold-standard.

	Spearman's Rank Coefficient (image-centered)									
Top K codewords	100	200	300	400	500	600	700	800	900	1000
<i>LSA tf</i>	0.228	0.325	0.273	0.242	0.185	<u>0.181</u>	0.107	0.043	-0.018	0.000
<i>LSA tf (norm)</i>	0.233	0.339	<u>0.293</u>	<u>0.254</u>	0.202	0.180	<u>0.124</u>	<u>0.047</u>	<u>-0.012</u>	0.000
<i>LSA tf*itf</i>	<u>0.268</u>	0.317	0.256	0.248	<u>0.219</u>	0.166	0.081	-0.004	-0.037	0.000
<i>LSA tf*itf (norm)</i>	0.252	0.327	0.257	0.246	0.211	0.153	0.097	0.002	-0.042	0.000
<i>VB tf</i>	0.243	0.168	0.101	0.055	-0.021	-0.084	-0.157	-0.210	-0.236	-0.332
<i>VB tf (norm)</i>	0.240	0.181	0.110	0.062	-0.010	-0.082	-0.152	-0.204	-0.235	-0.332
<i>VB tf*itf</i>	0.262	0.181	0.107	0.065	-0.019	-0.081	-0.156	-0.211	-0.241	-0.332
<i>VB tf*itf (norm)</i>	0.257	0.180	0.116	0.068	-0.014	-0.079	-0.150	-0.250	-0.237	-0.332
Random	0.001	0.018	0.016	-0.008	0.008	0.005	-0.001	0.014	-0.035	0.012
IHA	0.687									
	Spearman's Rank Coefficient (arbitrary-image)									
Top K codewords	100	200	300	400	500	600	700	800	900	1000
<i>LSA tf</i>	0.236	0.341	0.291	0.249	0.208	0.183	0.106	<u>0.033</u>	-0.039	0.000
<i>LSA tf (norm)</i>	0.230	0.353	<u>0.301</u>	<u>0.271</u>	0.220	<u>0.186</u>	<u>0.115</u>	0.032	<u>-0.029</u>	0.000
<i>LSA tf*itf</i>	<u>0.291</u>	0.332	0.289	0.262	<u>0.235</u>	0.172	0.092	0.008	-0.041	0.000
<i>LSA tf*itf (norm)</i>	0.277	0.345	0.292	0.269	0.234	0.164	0.098	0.015	-0.046	0.000
<i>VB tf</i>	0.272	0.195	0.119	0.059	-0.012	-0.088	-0.164	-0.218	-0.240	-0.339
<i>VB tf (norm)</i>	0.277	0.207	0.130	0.069	-0.003	-0.083	-0.160	-0.215	-0.242	-0.339
<i>VB tf*itf</i>	0.287	0.206	0.127	0.062	-0.008	-0.085	-0.161	-0.214	-0.241	-0.339
<i>VB tf*itf (norm)</i>	0.286	0.212	0.132	0.071	-0.005	-0.081	-0.158	-0.214	-0.241	-0.339
Random	-0.024	-0.014	0.015	-0.015	-0.004	-0.014	0.024	-0.009	-0.007	0.007
IHA	0.764									

TABLE 6.3. Correlation of automatically generated scores with human annotations on cross-modal semantic relatedness, as performed on the ImageNet test dataset of 2004 pairs of word and image. Correlation figures scoring the highest within a weighting scheme are marked in bold, while those scoring the highest across weighting schemes and within a visual vocabulary size are underlined.

The *Vector-Based (VB)* method is a stronger baseline aimed to study the correlation performance in the absence of dimensionality reduction. As an upper bound, the *Inter-Human-Agreement (IHA)* measures the correlation of the rating by each annotator against the average of the ratings of the rest of the annotators, averaged over the 167 synsets (for the image-centered scenario) and over the 2004 word-image pairs (for the arbitrary-image scenario).

6.6. Discussion

Our experimental results are shown in Table 6.3. A somewhat surprising observation is the consistency of correlation figures between the two scenarios. In both scenarios, a representative set of 200 visual codewords is sufficient to consistently score the highest correlation ratings across the 8

weighting schemes. Intuitively, based on the experimental results, automatically choosing the top 10% or 20% of the visual codewords seems to suffice and gives optimal correlation figures, but requires further justification. Conversely, the relatively simple weighting scheme using tf (*normalized*) produces the highest correlation in six visual codeword sizes ($K=200,300,400,700,800,900$) for the image-centered scenario, as well as in another six visual codeword sizes ($K=200, 300, 400, 600, 700, 900$) for the arbitrary-image scenario. Unlike stopwords in text retrieval accounting for most of the highest tf scores, visual codewords weighted by the same scheme tf and a similar tf (*normalized*) scheme seem to be the most discriminative. Note that the higher the discriminative power of a set of visual codewords for a synset⁹, the more representative they are able to visually describe the synset. As the discriminability of the codewords decreases, the composite vector representation of an image deviates from the same direction as the vector representing a word in a word-image pair judged to be similar. This translates directly to a lower cosine similarity score, and consequently, an inverse monotonic relationship with human judgments. For this reason, we observe a phenomenon of decreasing correlation scores into the negative region as the visual codeword size increases for the vector based model. The correlation for including the entire visual vocabulary set (1000) produces identical results for all vector-based and LSA weighting schemes, as images across synsets are now encoded by the same set of visual codewords without discrimination between them.

Dimensionality reduction using SVD gains an advantage over the vector-based method for both scenarios, with the highest correlation rating in LSA (200 visual codeword, $tf(norm)$) achieving 0.077 points better than the corresponding highest correlation in Vector-based (100 visual codeword, $tf*itf$) for the image-centered scenario, representing a 29.3% improvement. Similarly, in the arbitrary-image scenario, the increase in correlation from 0.287 (VB $tf*itf$ at 100 visual codeword) to 0.353 (LSA $tf(norm)$ at 200 visual codeword) underlines a gain of approximately 23.0%. Overall, the arbitrary-image scenario also scores consistently higher than the image-centered scenario under similar experimental conditions. For instance, for the top 200 visual words, the same weighting schemes produce consistently lower correlation figures for the image-centered scenario.

⁹It is important to realize that for any top K visual codeword size, the *actual* codewords used for different synsets may be very different, i.e., a specific subset of $^{1000}C_{200}$ codeword combinations, where $K=200$

This is also true for the Inter-Human-Agreement score, which is higher in the arbitrary-image scenario (0.764) compared to the image-centered scenario (0.687). Note that for all the experiments, the semantic relatedness scores generated from the semantic vector space are significantly more correlated with the human gold-standard than the random baselines.

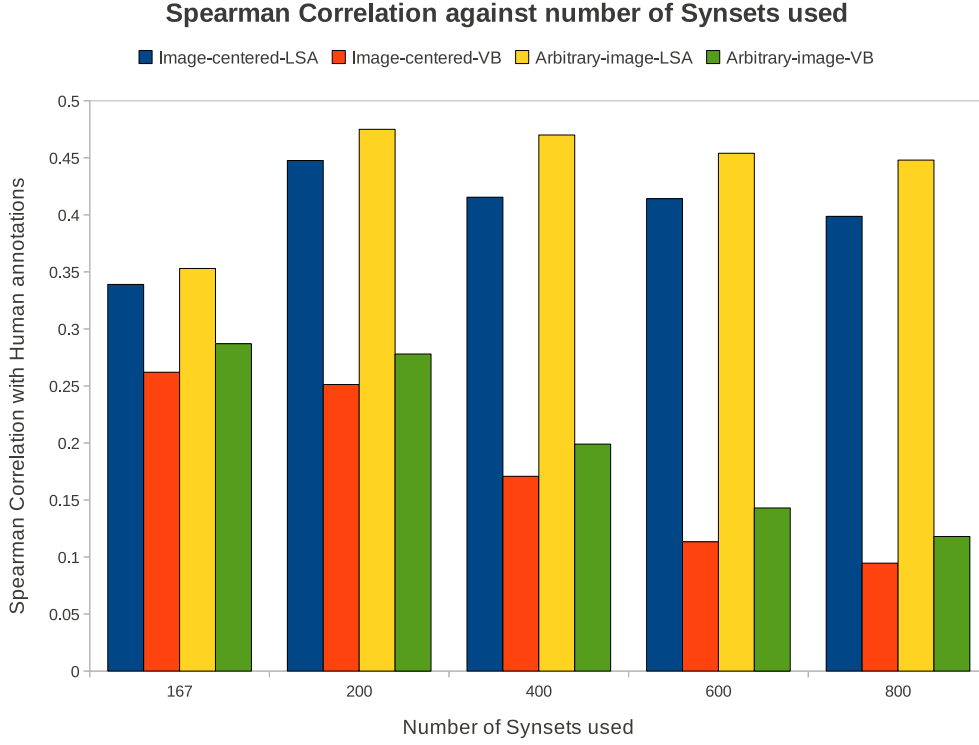


FIGURE 6.1. Spearman correlation performance with human annotations against number of synsets used

6.6.1. Scalability Effectiveness

To investigate the effectiveness of the model when scaling up to large datasets, we employ the best combination of weighting scheme and vocabulary size shown in Table 6.3, i.e., a visual vocabulary size of 200 and *tf (normalized)* weighting for LSA, and vocabulary size of 100 and *tf*idf* weighting for the vector-based model, and incrementally construct models ranging from 167 synsets to 800 synsets (all randomly selected from ImageNet). We then measure the correlation of relatedness scores generated using the same test dataset with respect to human annotations. The

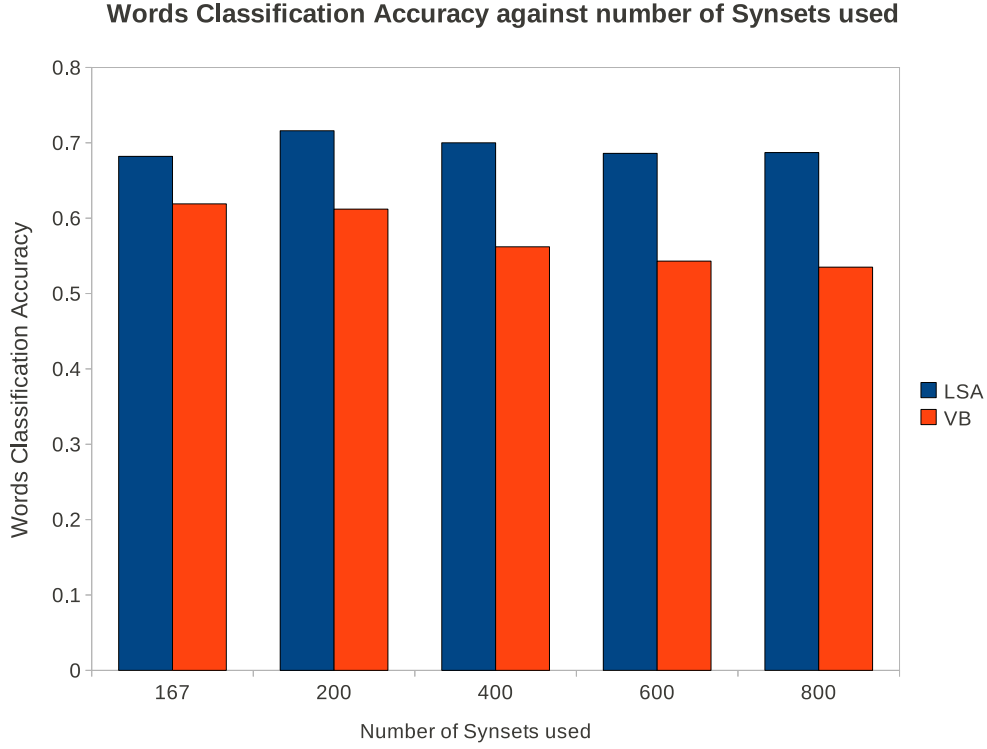


FIGURE 6.2. Classification accuracy, as more data is added to construct the semantic space model.

dataset was randomly selected to increase by approximately five times, from a total of 230,864 images with 878 words to a total of 1,014,528 images with 3887 words. Furthermore, for each unseen test image taken from Synset S_i and the associated 12 candidate words, we evaluate the ability of the model to identify which of the candidate words actually appear in the gloss or the synset of S_i , in a task we term as word classification. Here, the top six words are predictably classified as those appearing in S_i while the last six are classified as outside of S_i , after all 12 words are ranked in reverse order of their relatedness to the test image. We measure the accuracy of the word classification task using $\frac{TP+TN}{2004}$, where TP is the number of words correctly classified as synset or gloss words, and TN is the number of words correctly classified as outside of synset or gloss, both summed over the 2004 pairs of words and images.

As shown in Figure 6.1, when a small number of synsets (33) was added to the original semantic space, correlation with human ratings increased steeply to around 0.45 and higher for LSA in both scenarios, while the vector-based method suffers a slight decrease in correlation ratings from 0.262 to 0.251 (image-centered) and from 0.287 to 0.278 (arbitrary-image). As more images and words are added, correlation for the vector-based model continues to decrease markedly. Comparatively, LSA is less sensitive to data scaling, as correlation figures for both scenarios decreases slightly but stays within a 0.40 to 0.45 range. Additionally, we infer that LSA is consistently more effective than the vector-based model in the words classification task (as seen in Figure 6.2). Even with more data added to the semantic space, word classification accuracy stays consistently at 0.7 for LSA, while it drops to 0.535 for the vector-based model at a synset size of 800.

For both LSA and vector-based baseline models, as the size of the semantic space increases (via adding more synsets), there is an initial and significant increase in the performance on both the semantic relatedness and word classification tasks. This is perhaps due to the introduction of additional and ‘informative’ dimensions into the semantic space. Note that each successive synset that is randomly selected for constructing the semantic space is not necessarily closely related to previously selected synsets in the WordNet concept hierarchy. Rather, it could be any one of a diverse set of synsets covering different topics, as long as each of them can be illustrated with images. However, as more synsets are added, the probability of a given textual word being in more than one synset increases. On the other hand, the same set of visual codewords describing an image, which is initially associated with the textual word, may not always be present in each additional synset containing the latter. This decreasing dissociation between the image-word pair probably causes the slow but steady decrease in performance of both tasks, as more and more synsets are added.

6.7. Extended Study : Using Image Attributes for Measuring Cross-Modal Semantic Relatedness

While advances in computer research have seen a maturity in the detection of object categories (such as cars, trees, persons etc) and even object parts (e.g. “leg” or “wheel”), recent interests have shifted to the modeling of *visual attributes* that arise from semantic description of objects [45, 23, 22]. For instance, besides training an object classifier to recognize the presence or absence



FIGURE 6.3. Examples of object attributes taken directly from ImageNet website, which are present in our dataset

of a “dog” in an image, there is a growing effort to extract object attributes, e.g., automatically describing the dog as “black” or “furry”. Figure 6.3 illustrates some objects with their extracted attributes. Attribute learning like this bears important implications. It not only enriches the set of semantic tags for the object, but it also eases the construction of large-scale image databases and image ontologies. As the number of object categories increases to thousands and millions, training of individual classifiers for each object is impractical. Instead, the acquisition of new images for each category or the recognition of image categories can be performed by employing sufficiently large set of attributes via transfer learning [71]. Indeed, it has been shown that combining verbal descriptions of attributes with just a few training images, or even the exclusive use of verbal descriptions, leads to more efficient and effective image categorization [23].

In this section, we present an extended study on solving the cross-modal semantic relatedness problem by combining visual codeword and visual object attributes, using unsupervised corpus-based approaches. As mentioned, visual object attributes learning or modeling is a relatively new research area in computer vision, where efforts have focused on modeling different types of attributes that are appearance-based [45, 23, 22], object parts-based [45, 22] or attributes exhibiting similarity between objects [45]. Of particular interest to us is the detection of appearance-based attributes that are semantic in nature, i.e., the attribute is visual and can be described sufficiently using natural language.

This section addresses the following research questions. First, is it possible to augment the bag-of-visual codewords approach with learned visual attributes in a corpus-based representation? Second, given such an augmented multimodal semantic representation, can we derive a semantic relatedness measure between arbitrary pairs of words and images that achieves an improved correlation rating with respect to human judgments on the same evaluation dataset?

6.7.1. Dataset

To our knowledge, semantic, appearance-based visual attributes have not been employed in any previous work related to measuring semantic relatedness between words and images. An intuitive approach to acquire a suitable dataset for evaluation is to perform data mining on any available image database consisting of paired word and images, according to a list of desirable attributes. However, datasets constructed in this way often tend to be overly sparse and unbalanced in the attribute space [71]. Furthermore, due to its infancy, the extraction of semantic, appearance based attributes suffers from low accuracy [71], and is not suitable for evaluating our hypothesis due to the percolation of noise. Since we have shown previously that deployment of low-level descriptors such as SIFT is promising for solving the cross-modal semantic relatedness problem, our goal here is to look forward in other orthogonal directions for exploitation of image-based information cues. For this reason, acquiring a dataset of clean annotations of visual attributes is important to establish our hypothesis about their applicability to solve our task at hand. For our dataset, we use the ground-truth annotated 384 synsets from ImageNet collected for attribute learning in large-scale datasets in [71]. The dataset contains 9600 images covering 384 synsets, each

annotated by 3 annotators from Amazon Mechanical Turk. For each of the 25 selected attributes, each annotator indicated on each image the presence/absence of the attribute. Overall, the image is marked as (+1) if all three annotators agree about its presence, (-1) if all agree on its absence, or otherwise ambiguous (0) if there is a lack of consensus. Table 6.4 shows the categorized attributes found in our dataset¹⁰

CATEGORY	ATTRIBUTES
Color	black, blue, brown, gray, green, orange, pink, red, violet, white, yellow
Pattern	spotted, striped
Shape	long, round, rectangular, square
Texture	furry, smooth, rough, shiny, metallic, vegetation, wooden, wet

TABLE 6.4. Table showing high-level categories and the corresponding attributes

As before, we employ the bag-of-visual codewords extraction procedure outlined in 3.3 to represent each image in our dataset, using a subset of the visual codeword vocabulary. For each synset, we processed its associated images to derive a representative set of visual codewords, which is appended to all its synset and gloss words. In order to incorporate our newly acquired attributes in this bag-of-words representation, mapping from the former to the latter is necessary. Formally, let $I_s = \{i_1, \dots, i_k\}$ be the set of images associated with synset s , $A = \langle a_j | j = 1, \dots, N \rangle$ be the N (25) semantic, appearance-based attributes from above. For each image $i \in I_s$, $L_i = \langle l_i^k | k = 1, \dots, N \rangle$, where $\text{Label}(a_j) = l_i^k$, and $l_i^k \in \{0, 1, -1\}$. We seek to find function $g : L_i \rightarrow \{0, 1\}^{|L_i|}$, where 0 indicates the absence of the attribute, and 1 indicates its presence. In other words, we are trying to find a mapping function from a ternary-valued attribute label to a binary-valued visual attribute codeword. Trivially, it may seem that we can discard the ambiguous label where $l_i^k = 0$, however, note that the same set of visual attributes is extracted from all images in all 384 synsets. In some cases, a clear establishment of an attribute in some synsets may be unclear, as positive instances are rare for discrimination purposes. Indeed, a close examination of the data confirms this intuition. Instead, g is determined for each synset based on a majority vote

¹⁰<http://www.image-net.org/download-attributes>

from labels from its member images. An attribute label from L_i is mapped only if it is demonstrated in more than half of the synset images in the process of constructing the semantic space. In case of a tie, i.e., the attribute is established in half of the synset images, then we are unsure of its representativeness of the synset entity. In this case, we perform mapping to attribute codewords both binary co-dominant attribute values. For instance, if only half of the training images for synset {dog, domestic dog, Canis familiaris} consists of dogs in black, while the color of the dog in the other half of the images could be all ambiguous e.g. gray, or all non-black e.g. brown, then we append to the synset both attribute codeword BLACK and NOT_BLACK, since it is uncertain that the color black is a representative attribute of dog.

Consequently, the final composition of our appended synsets can be summarized as follows. Suppose TC , VC and AC are the vocabulary of textual words, visual codewords and attribute codewords respectively, then any synset $s = \{t_i, v_j, a_k | t_i \in TC, v_j \in VC, a_k \in AC\}$. Specifically, we set $j=1$ to 200 for the visual codeword size based on optimal performance using the LSA model from the work performed earlier, while $k = 1$ to 25.

For evaluation, we randomly select 150 synsets from the 384 synsets annotated with visual attributes. A randomly selected image from each of these 150 synsets is set aside for testing, while the rest of images are used to build the multimodal semantic space. As before, for each test image, we randomly selected 6 words from the synset and gloss, and another 6 words elsewhere. For each pair of image and word, we invited 15 annotators with at least 97% approval rating from AMT to rate their association from 0 to 10 using the cross-modal evaluation framework explained earlier. Altogether, we obtained $150 \times 12 = 1800$ annotated pairs of word and image. For each pair, we computed the average score from the 15 annotations. Table 6.5 shows examples of test image and word pairs.

Following earlier work, we again compute the upper bound *Inter-Human-Agreement (IHA)*, which measures the correlation of the rating by each annotator against the average of the ratings of the rest of the annotators, averaged over the 150 synsets (for the image-centered scenario) and over the 1800 word-image pairs (for the arbitrary-image scenario). Since the focus here is on improvement in correlation, not error reduction, baselines are not necessary.



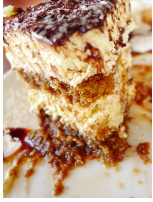
		
Synset {silverback}	Synset {cross}	Synset {tiramisu}
Gloss an adult male gorilla with grey hairs across the back	Gloss a wooden structure consisting of an upright post with a transverse piece	Gloss an Italian dessert consisting of layers of sponge cake soaked with coffee and brandy or liqueur layered with mascarpone cheese and topped with grated chocolate
Relatedness Scores adult (8.07) deep (2.27) gorilla (9.93) grey (5.33) hair (5.33) male (7.13) oil (1.00) herbivorous (6.47) muzzle (2.93) neck (4.13) horse (0.20) silverback (6.87)	Relatedness Scores consist (0.87) cross (10.00) fleet (0.53) habit (0.33) hoof (1.13) hungarian (0.47) hunting (0.80) post (5.20) structure (6.80) upright (5.93) wild (0.87) wooden (7.60)	Relatedness Scores cake (9.27) chocolate (7.67) dessert (8.27) grate (2.60) hinny (0.27) italian (4.93) low (1.00) mixture (3.93) ridge (0.80) rum (2.53) soak (1.93) toe (0.27)
Visual Attributes black (color) furry (texture)	Visual Attributes black (color) long (shape) rectangular (shape) smooth (texture)	Visual Attributes black (color) brown (color)

TABLE 6.5. A sample of test images with their synset words and glosses, human annotations : The number in parenthesis represents the numerical association of the word with the image (0-10). All content words are considered. The visual attributes agreeable by all annotators are listed for each image

6.7.2. Experiments

Our extended study involving semantic, appearance-based attributes is to observe if an augmented multimodal semantic space that includes visual attribute codewords can lead to improvements in measuring cross-modal semantic relatedness over the previously employed best-performing system i.e. vector space model using LSA (**VSM-LSA**) to reduce dimensionality. Upon this basic multimodal semantic space constructed using the 150 synsets, we further made three enhancements. First, in order to investigate the effect of adding each individual attribute to the basic model, we constructed 25 augmented semantic spaces, each using VSM-LSA and

mapped codewords belonging to one and only one of the 25 visual attributes. Second, we investigate addition of categorized attributes to the basic model, by constructing another four augmented semantic spaces, each integrating mapped codewords belonging to all related attributes within a high-level attribute category i.e. color (**CLR-VSM-LSA**), pattern (**PTT-VSM-LSA**), shape (**SHP-VSM-LSA**), and texture (**TXT-VSM-LSA**). Finally, we want to observe the performance by incorporating all 25 attributes in a composite semantic space, **COMB-ATT-VSM-LSA**, by augmenting VSM-LSA using all codewords belonging to all attributes. For each of the semantic spaces constructed above, we compute the Spearman correlation of the automatically generated semantic relatedness scores between the 1800 pairs of word and image under both image-centered and arbitrary-image scenarios. Note that tf-normalized weighting ($tf(norm)$) is used for selecting the top K visual codewords for constructing all the semantic spaces, where $K = 200$, as this was the the visual codeword size resulting in the highest performance for VSM-LSA in Table 6.3.

	Arbitrary	Image-Centered
CLR-VSM-LSA	0.348	0.361
PTT-VSM-LSA	0.342	0.363
SHP-VSM-LSA	0.348	0.366
TXT-VSM-LSA	0.351	0.366
COMB-ATT-VSM-LSA	0.369	0.378
VSM-LSA	0.344	0.364
Upperbound (Humans)	0.613	0.523

TABLE 6.6. Spearman correlation performance of augmented VSM-LSAs using visual attributes and their integration within high level categories

6.7.3. Discussion

The results for the semantic spaces enhanced with individual visual attributes are shown in Figure 6.4. Across both arbitrary-image and image-centered scenarios, attributes scoring consistent improvements over the basic VSM-LSA are green (color), white (color), yellow (color), rectangular (shape), round (shape), smooth (texture), shiny (texture), vegetation (texture) and wooden

(texture). For the rest of the attributes, findings are largely inconclusive. Furthermore, even the highest improvement scored, i.e., white (color) for image-centered scenario, is within a margin of 0.01. These marginal improvements can be attributed to the inherent nature of our dataset. A simple analysis reveals that out of 9600 annotated images, approximately 8% of the attributes are labeled as positives, while 80% are negatives, and 12% are ambiguously labeled. The unbalanced labels constituted a challenging scenario for effective discrimination between object classes based on appearance-based attributes, and therefore, translates indirectly to loose association between visual words in one synset and textual words in another synset, resulting in little or no improvement over the basic VSM-LSA. Given that the labels are already human-labeled ground truths, it is worth trying to explore other directions, such as weighing in on attributes that are better represented across diverse object categories.

The results for semantic spaces constructed by using integrated codewords within high-level categories are shown in Table 6.6. Overall, compared to our basic VSM-LSA semantic space, all enhanced versions score improvements in either arbitrary-image or image-centered scenarios, or both, with the exception of PTT-VSM-LSA. The highest improvement from the four semantic spaces is represented by TXT-VSM-LSA (0.351 for arbitrary-image, and 0.366 for image-centered). A closer examination reveals perhaps a correlation between the number of attributes contained in the high-level category and its performance. This is especially true for PTT-VSM-LSA where there exists only two pattern attributes. However, combining all 25 visual attributes together into a single composite semantic space yields the highest performance of all systems we experimented with, scoring a Spearman correlation of 0.369 for the arbitrary-image and 0.378 for the image-centered scenarios respectively. Note that the correlation upperbounds for human on this dataset are lower than the ones computed for earlier work, most likely due to the inclusion of content words of all part-of-speech i.e. nouns, adjectives and adverbs. This presents significant difficulties in quantifying relation between an image and a non-noun in its synset, e.g., the word-image pair of ‘consist’ and the image depicting ‘cross’ in Table 6.5, because the former is

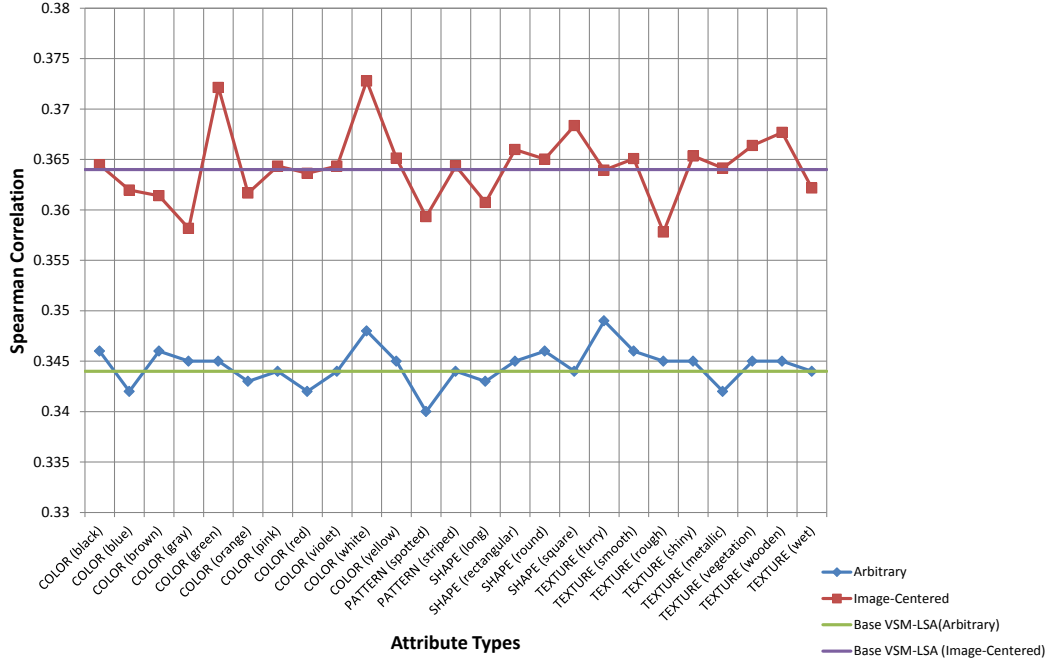


FIGURE 6.4. Spearman correlation performance of VSMs augmented with individual attributes

arguably much more general than concrete nouns in its semantics, but represents our first-cut approach at building a multimodal semantic space covering all words and images, instead of focusing exclusively on nouns.

6.8. Related Work

Despite the large amount of work in computing semantic relatedness between words or similarity between images, there are only a few studies in the literature that associate the meaning of words and pictures in a joint semantic space. The work most similar to ours was done by [84], who employed LSA to combine textual words with simple visual features extracted from news images using colors and textures. Although it was concluded that such a joint textual-visual representation model was promising for image retrieval, no intensive evaluation was performed on datasets on a large scale, or datasets other than the news domain. Similarly, [33] compared different methods

such as LSA and probabilistic LSA to construct joint semantic spaces in order to study their effects on automatic image annotation and semantic image retrieval, but their evaluation was restricted exclusively to the Corel dataset, which is somewhat idealistic and not reflective of the challenges presented by real-world, noisy images.

Another related line of work by [6] used a generative hierarchical model to learn the associative semantics of words and images for improving information retrieval tasks. Their approach was supervised and evaluated again only on the Corel dataset.

More recently, [27] showed that it is possible to combine visual representations of word meanings into a joint bimodal representation constructed by using latent topics. While their work focused on unifying meanings from visual and textual data via supervised techniques, no effort was made to compare the semantic relatedness between arbitrary pairs of word and image.

CHAPTER 7

CONCLUSION

Language and vision endure together, or perish alone. The goal of this dissertation is to provide empirical evidence in further support of the synergism between language and vision. Specifically, we are interested to provide computational approaches to explore and exploit the synergistic relationships between the visual and textual modalities, and to further stretch the boundaries of their applicability in generating novel solutions to existing problems.

7.1. Research Questions Revisited

1. Is it possible to decode information in one modality to help tasks existing in the other modality ?

With regards to interoperability of textual features for tasks in vision, we introduced several text-based extractive approaches for the selected task of automatic image annotation and showed that they compare favorably with the state-of-the-art model using both text and image features. We believe our work has practical applications in mining and annotating images over the Web, where texts are naturally associated with images, and scalability is important. Our next direction seeks to derive robust annotation models using additional ontological knowledge-bases. We would also like to advance the the state-of-the-art by augmenting current textual models with image features.

Research in the other direction also shows a lot of promise, as we are able to exploit the visual information presented by images in a parallel corpora containing words and images, and devise a new image-based metric for measuring word relatedness. Using pair-wise combination of the outputs from several text-based metric and this new image-based metric, we are able to obtain the state-of-the-art for the relatedness, as measured on three standard evaluation datasets. For improving coverage, we bridged the gap between WordNet and a corpus-based resource, Wikipedia,

for further validating the promise of our approach, and outline circumstances when such an image-based metric should be used.

2. Considering the supplementary and complementary advantages of each modality over the other, can we integrate image and word features into a unified framework for the construction of a richer semantic space ?

The meaning of words are represented by discrete lexical units, each of which is defined according to a dictionary. On the contrary, the meaning of an image is not determined a priori, but rather depends on the vision task required at hand to justify for its semantics. The first step in constructing a multimodal semantic space requires quantization of an image into discrete units, termed as visual codewords. By using simple statistical relationships between words and images mined from a large parallel corpora, we are able to employ semantic vectorial models used in natural language processing to construct a joint semantic space.

The construction of such a joint multimodal semantic space bears huge promises, such as its ability to measure relatedness between pairs of words, a word and an image, and pairs of images directly using a concept-based distributional semantics framework.

3. Can we formalize the meaning of images by using words in languages?

We provided a proof of concept in quantifying the semantic relatedness between words and images through the use of visual codewords and textual words in constructing a joint semantic vector space. Our experiments showed that the relatedness scores have a positive correlation to human gold-standards, as measured using a novel evaluation framework devised to quantify cross-modal semantic relatedness, which is inspired by the semantic relations in WordNet.

We believe many aspects of this work can be explored further. For instance, other visual codeword attributes, such as pixel coordinates, can be employed in a structured vector space along with the existing model for improving vector similarity measures.

APPENDIX A
EVALUATION DATASETS

MC30 Dataset

WORD	WORD	SCORE
asylum	madhouse	3.610
bird	cock	3.050
bird	crane	2.970
boy	lad	3.760
brother	monk	2.820
car	automobile	3.920
cemetery	woodland	0.950
chord	smile	0.130
coast	forest	0.420
coast	hill	0.870
coast	shore	3.700
crane	implement	1.680
food	fruit	3.080
food	rooster	0.890
forest	graveyard	0.840
furnace	stove	3.110
gem	jewel	3.840
glass	magician	0.110
journey	car	1.160
journey	voyage	3.840
lad	brother	1.660
lad	wizard	0.420
magician	wizard	3.500
midday	noon	3.420
monk	oracle	1.100
monk	slave	0.550
noon	string	0.080
rooster	voyage	0.080
shore	woodland	0.630
tool	implement	2.950

RG65 Dataset

WORD	WORD	SCORE
gem	jewel	3.940
midday	noon	3.940
automobile	car	3.920
cemetery	graveyard	3.880
cushion	pillow	3.840
boy	lad	3.820
cock	rooster	3.680
implement	tool	3.660
forest	woodland	3.650
coast	shore	3.600
autograph	signature	3.590
journey	voyage	3.580
serf	slave	3.460
grin	smile	3.460
glass	tumbler	3.450
cord	string	3.410
hill	mound	3.290
magician	wizard	3.210
furnace	stove	3.110
asylum	madhouse	3.040
brother	monk	2.740
food	fruit	2.690
bird	cock	2.630
bird	crane	2.630
oracle	sage	2.610
sage	wizard	2.460
brother	lad	2.410
crane	implement	2.370
magician	oracle	1.820
glass	jewel	1.780
cemetery	mound	1.690
car	journey	1.550
hill	woodland	1.480
crane	rooster	1.410
furnace	implement	1.370
coast	hill	1.260
bird	woodland	1.240
shore	voyage	1.220
cemetery	woodland	1.180
food	rooster	1.090
forest	graveyard	1.000
lad	wizard	0.990
mound	shore	0.970
Continue on the next page		

WORD	WORD	SCORE
automobile	cushion	0.970
boy	sage	0.960
monk	oracle	0.910
shore	woodland	0.900
grin	lad	0.880
coast	forest	0.850
asylum	cemetery	0.790
monk	slave	0.570
cushion	jewel	0.450
boy	rooster	0.440
glass	magician	0.440
graveyard	madhouse	0.420
asylum	monk	0.390
asylum	fruit	0.190
grin	implement	0.180
mound	stove	0.140
automobile	wizard	0.110
autograph	shore	0.060
fruit	furnace	0.050
noon	string	0.040
rooster	voyage	0.040
cord	smile	0.020

WS353 Dataset

WORD	WORD	SCORE
admission	ticket	7.69
alcohol	chemistry	5.54
aluminum	metal	7.83
announcement	effort	2.75
announcement	news	7.56
announcement	production	3.38
announcement	warning	6.00
arafat	jackson	2.50
arafat	peace	6.73
arafat	terror	7.65
architecture	century	3.78
arrangement	accommodation	5.41
arrival	hotel	6.00
asylum	madhouse	8.87
atmosphere	landscape	3.69
attempt	peace	4.25
baby	mother	7.85
bank	money	8.12
baseball	season	5.97
bed	closet	6.72
benchmark	index	4.25
bird	cock	7.10
bird	crane	7.38
bishop	rabbi	6.69
board	recommendation	4.47
book	library	7.46
book	paper	7.46
boxing	round	7.61
boy	lad	8.83
bread	butter	6.19
brother	monk	6.27
calculation	computation	8.44
canyon	landscape	7.53
car	automobile	8.94
car	flight	4.94
cell	phone	7.81
cemetery	woodland	2.08
century	nation	3.16
century	year	7.59
championship	tournament	8.36
chance	credibility	3.88
change	attitude	5.44
Continue on the next page		

WORD	WORD	SCORE
chord	smile	0.54
closet	clothes	8.00
coast	forest	3.15
coast	hill	4.38
coast	shore	9.10
company	stock	7.08
competition	price	6.44
computer	internet	7.58
computer	keyboard	7.62
computer	laboratory	6.78
computer	news	4.47
computer	software	8.50
concert	virtuoso	6.81
consumer	confidence	4.13
consumer	energy	4.75
country	citizen	7.31
crane	implement	2.69
credit	card	8.06
credit	information	5.31
cucumber	potato	5.92
cup	article	2.40
cup	artifact	2.92
cup	coffee	6.58
cup	drink	7.25
cup	entity	2.15
cup	food	5.00
cup	liquid	5.90
cup	object	3.69
cup	substance	1.92
cup	tableware	6.85
currency	market	7.50
day	dawn	7.53
day	summer	3.94
death	inmate	5.03
death	row	5.25
decoration	valor	5.63
delay	news	3.31
delay	racism	1.19
deployment	departure	4.25
deployment	withdrawal	5.88
development	issue	3.97
direction	combination	2.25
Continue on the next page		

WORD	WORD	SCORE
disability	death	5.47
disaster	area	6.25
discovery	space	6.34
dividend	calculation	6.48
dividend	payment	7.63
doctor	liability	5.19
doctor	nurse	7.00
doctor	personnel	5.00
dollar	buck	9.22
dollar	loss	6.09
dollar	profit	7.38
dollar	yen	7.78
drink	car	3.04
drink	ear	1.31
drink	eat	6.87
drink	mother	2.65
drink	mouth	5.96
drug	abuse	6.85
energy	crisis	5.94
energy	laboratory	5.09
energy	secretary	1.81
environment	ecology	8.81
equipment	maker	5.91
exhibit	memorabilia	5.31
experience	music	3.47
family	planning	6.25
fbi	fingerprint	6.94
fbi	investigation	8.31
fertility	egg	6.69
fighting	defeating	7.41
five	month	3.38
focus	life	4.06
food	fruit	7.52
food	preparation	6.22
food	rooster	4.42
football	basketball	6.81
football	soccer	9.03
football	tennis	6.63
forest	graveyard	1.85
fuck	sex	9.44
furnace	stove	8.79
game	defeat	6.97
Continue on the next page		

WORD	WORD	SCORE
game	round	5.97
game	series	6.19
game	team	7.69
game	victory	7.03
gem	jewel	8.96
gender	equality	6.41
glass	magician	2.08
glass	metal	5.56
government	crisis	6.56
governor	interview	3.25
governor	office	6.34
grocery	money	5.94
harvard	yale	8.13
holy	sex	1.62
hospital	infrastructure	4.63
hotel	reservation	8.03
hundred	percent	7.38
image	surface	4.56
impartiality	interest	5.16
investigation	effort	4.59
investor	earning	7.13
jaguar	car	7.27
jaguar	cat	7.42
japanese	american	6.50
jerusalem	israel	8.46
jerusalem	palestinian	7.65
journal	association	4.97
journey	car	5.85
journey	voyage	9.29
king	cabbage	0.23
king	queen	8.58
king	rook	5.92
lad	brother	4.46
lad	wizard	0.92
law	lawyer	8.38
lawyer	evidence	6.69
liability	insurance	7.03
life	death	7.88
life	lesson	5.94
life	term	4.50
line	insurance	2.69
liquid	water	7.89
Continue on the next page		

WORD	WORD	SCORE
listing	category	6.38
listing	proximity	2.56
lobster	food	7.81
lobster	wine	5.70
love	sex	6.77
lover	quarrel	6.19
luxury	car	6.47
magician	wizard	9.02
man	governor	5.25
man	woman	8.30
maradona	football	8.62
marathon	sprint	7.47
mars	scientist	5.63
mars	water	2.94
media	gain	2.88
media	radio	7.42
media	trading	3.88
mexico	brazil	7.44
midday	noon	9.29
mile	kilometer	8.66
minister	party	6.63
ministry	culture	4.69
minority	peace	3.69
money	bank	8.50
money	cash	9.15
money	currency	9.04
money	deposit	7.73
money	dollar	8.42
money	laundering	5.65
money	operation	3.31
money	possession	7.29
money	property	7.57
money	wealth	8.27
money	withdrawal	6.88
monk	oracle	5.00
monk	slave	0.92
month	hotel	1.81
morality	importance	3.31
morality	marriage	3.69
movie	critic	6.73
movie	popcorn	6.19
movie	star	7.38
Continue on the next page		

WORD	WORD	SCORE
movie	theater	7.92
murder	manslaughter	8.53
museum	theater	7.19
music	project	3.63
nature	environment	8.31
nature	man	6.25
network	hardware	8.31
news	report	8.16
noon	string	0.54
observation	architecture	4.38
oil	stock	6.34
opec	country	5.63
opec	oil	8.59
opera	industry	2.63
opera	performance	6.88
peace	atmosphere	3.69
peace	insurance	2.94
peace	plan	4.75
phone	equipment	7.13
physics	chemistry	7.35
physics	proton	8.12
plane	car	5.77
planet	astronomer	7.94
planet	constellation	8.06
planet	galaxy	8.11
planet	moon	8.08
planet	people	5.75
planet	space	7.92
planet	star	8.45
planet	sun	8.02
population	development	3.75
possibility	girl	1.94
practice	institution	3.19
precedent	antecedent	6.04
precedent	cognition	2.81
precedent	collection	2.50
precedent	example	5.85
precedent	group	1.77
precedent	information	3.85
precedent	law	6.65
prejudice	recognition	3.00
preservation	world	6.19
Continue on the next page		

WORD	WORD	SCORE
president	medal	3.00
problem	airport	2.38
problem	challenge	6.75
production	crew	6.25
production	hike	1.75
professor	cucumber	0.31
professor	doctor	6.62
profit	loss	7.63
profit	warning	3.88
psychology	anxiety	7.00
psychology	clinic	6.58
psychology	cognition	7.48
psychology	depression	7.42
psychology	discipline	5.58
psychology	doctor	6.42
psychology	fear	6.85
psychology	freud	8.21
psychology	health	7.23
psychology	mind	7.69
psychology	psychiatry	8.08
psychology	science	6.71
reason	criterion	5.91
reason	hypertension	2.31
record	number	6.31
registration	arrangement	6.00
report	gain	3.63
rock	jazz	7.59
rooster	voyage	0.62
school	center	3.44
seafood	food	8.34
seafood	lobster	8.70
seafood	sea	7.47
secretary	senate	5.06
seven	series	3.56
shore	woodland	3.08
shower	flood	6.03
shower	thunderstorm	6.31
sign	recess	2.38
situation	conclusion	4.81
situation	isolation	3.88
size	prominence	5.31
skin	eye	6.22
Continue on the next page		

WORD	WORD	SCORE
smart	student	4.62
smart	stupid	5.81
soap	opera	7.94
space	chemistry	4.88
space	world	6.53
start	match	4.47
start	year	4.06
stock	cd	1.31
stock	egg	1.81
stock	jaguar	0.92
stock	life	0.92
stock	live	3.73
stock	market	8.08
stock	phone	1.62
street	avenue	8.88
street	block	6.88
street	children	4.94
street	place	6.44
stroke	hospital	7.03
student	professor	6.81
sugar	approach	0.88
summer	drought	7.16
summer	nature	5.63
telephone	communication	7.50
television	film	7.72
television	radio	6.77
tennis	racket	7.56
territory	kilometer	5.28
territory	surface	5.34
theater	history	3.91
tiger	animal	7.00
tiger	carnivore	7.08
tiger	cat	7.35
tiger	fauna	5.62
tiger	feline	8.00
tiger	jaguar	8.00
tiger	mammal	6.85
tiger	organism	4.77
tiger	tiger	10.00
tiger	zoo	5.87
tool	implement	6.46
train	car	6.31
Continue on the next page		

WORD	WORD	SCORE
travel	activity	5.00
treatment	recovery	7.91
type	kind	8.97
victim	emergency	6.47
video	archive	6.34
viewer	serial	2.97
vodka	brandy	8.13
vodka	gin	8.46
volunteer	motto	2.56
war	troops	8.13
water	seepage	6.56
weapon	secret	6.06
weather	forecast	8.34
wednesday	news	2.22
wood	forest	7.73
word	similarity	4.75

APPENDIX B
PICTURABILITY SCORES

Word Picturablity Scores (Vocabulary words of MC30, RG65 and WS353)

WORD	SCORE
American	0.99898
Arafat	0.97671
Brazil	0.99932
CD	0.99950
FBI	0.92509
Freud	0.95649
Harvard	0.89777
Israel	0.99333
Jackson	0.99666
Japanese	0.99885
Jerusalem	0.99405
Maradona	0.99981
Mars	0.99893
Mexico	0.99985
OPEC	0.50377
Palestinian	0.95800
Wednesday	0.99958
Yale	0.82643
abuse	0.99521
accommodation	0.98674
activity	0.99605
admission	0.93071
airport	0.99817
alcohol	0.99753
aluminum	0.99698
animal	0.99996
announcement	0.99138
antecedent	0.98406
anxiety	0.82794
approach	0.94356
architecture	0.99997
archive	0.99994
area	0.99864
arrangement	0.99779
arrival	0.96780
article	0.99973
artifact	0.99710
association	0.97094
astronomer	0.99864
Continue on the next page	

WORD	SCORE
asylum	0.99383
atmosphere	0.98869
attempt	0.93329
attitude	0.98268
autograph	0.99699
automobile	0.99986
avenue	0.98445
baby	0.99976
bank	0.98603
baseball	0.99896
basketball	0.99987
bed	0.99404
benchmark	0.97072
bird	0.99961
bishop	0.83676
block	0.99057
board	0.99826
book	0.99901
boxing	0.99961
boy	0.99904
brandy	0.98045
bread	0.98793
brother	0.99220
buck	0.86854
butter	0.99193
cabbage	0.98011
calculation	0.75510
canyon	0.99708
car	1.00000
card	0.99846
carnivore	0.99852
cash	0.91636
cat	0.99971
category	0.99991
cell	0.99735
cemetery	0.97861
center	1.00000
century	0.99779
challenge	0.98451
Continue on the next page	

WORD	SCORE
championship	0.99639
chance	0.98205
change	0.99822
chemistry	0.97836
children	0.99891
chord	0.99712
citizen	0.93576
clinic	0.95070
closet	0.99694
clothes	0.99742
coast	0.99566
cock	0.99944
coffee	0.99572
cognition	0.87171
collection	0.99985
combination	0.99389
communication	0.98093
company	0.99998
competition	0.99610
computation	0.78135
computer	0.99976
concert	0.99894
conclusion	0.91104
confidence	0.92908
constellation	0.99866
consumer	0.95259
cord	0.94010
country	0.99844
crane	0.97496
credibility	0.66468
credit	0.99897
crew	0.99794
crisis	0.98470
criterion	0.85832
critic	0.98792
cucumber	0.82088
culture	0.99901
cup	0.99834
currency	0.99407
Continue on the next page	

WORD	SCORE
cushion	0.98276
dawn	0.99721
day	0.99968
death	0.99783
decoration	0.99994
defeat	0.92686
defeating	0.91241
delay	0.93634
departure	0.94028
deployment	0.96347
deposit	0.97087
depression	0.97478
development	0.99602
direction	0.99714
disability	0.71583
disaster	0.98568
discipline	0.75835
discovery	0.98575
dividend	0.58771
doctor	0.98643
dollar	0.97757
drink	0.99614
drought	0.98384
drug	0.94053
ear	0.99981
earning	0.88197
eat	0.99437
ecology	0.98970
effort	0.95990
egg	0.99478
emergency	0.96078
energy	0.99272
entity	0.87839
environment	0.99332
equality	0.84827
equipment	0.98823
evidence	0.89248
example	0.99721
exhibit	0.99325
Continue on the next page	

WORD	SCORE
experience	0.98867
eye	0.99817
family	0.99867
fauna	0.99977
fear	0.99752
feline	0.99490
fertility	0.91725
fighting	0.99863
film	0.99986
fingerprint	0.99270
five	0.99382
flight	0.99651
flood	0.95960
focus	0.99464
food	0.99949
football	0.99989
forecast	0.98582
forest	0.99900
fruit	0.99855
fuck	0.99954
furnace	0.90124
gain	0.89790
galaxy	0.99968
game	0.99982
gem	0.98541
gender	0.98073
gin	0.95624
girl	0.99992
glass	0.99853
government	0.98418
governor	0.91298
graveyard	0.99193
grin	0.98833
grocery	0.71730
group	0.99909
hardware	0.99236
health	0.99701
hike	0.98783
hill	0.98690
Continue on the next page	

WORD	SCORE
history	0.99935
holy	0.99604
hospital	0.98257
hotel	0.99939
hundred	0.96095
hypertension	0.95103
image	1.00000
impartiality	0.23983
implement	0.88426
importance	0.95494
index	0.99940
industry	0.99301
information	0.99893
infrastructure	0.94562
inmate	0.84381
institution	0.82395
insurance	0.98674
interest	0.97194
internet	0.99941
interview	0.99697
investigation	0.84636
investor	0.36210
isolation	0.96559
issue	0.99466
jaguar	0.99999
jazz	0.99158
jewel	0.98481
journal	0.99765
journey	0.99412
keyboard	0.99955
kilometer	0.98447
kind	0.99468
king	0.99786
laboratory	0.96097
lad	0.98680
landscape	0.99999
laundering	0.75040
law	0.98622
lawyer	0.86294
Continue on the next page	

WORD	SCORE
lesson	0.98881
liability	0.89161
library	0.99622
life	0.99969
line	0.99828
liquid	0.97938
listing	0.99175
live	0.99918
lobster	0.97687
loss	0.98938
love	0.99990
lover	0.99530
luxury	0.99962
madhouse	0.99591
magician	0.98060
maker	0.99410
mammal	0.99994
man	0.99940
manslaughter	0.94173
marathon	0.97773
market	0.99525
marriage	0.99806
match	0.99749
medal	0.99449
media	0.99996
memorabilia	0.99782
metal	0.99859
midday	0.90586
mile	0.97707
mind	0.99736
minister	0.97823
ministry	0.98891
minority	0.79544
money	0.99338
monk	0.97661
month	0.99829
moon	0.99897
morality	0.81823
mother	0.99824
Continue on the next page	

WORD	SCORE
motto	0.94844
mound	0.86758
mouth	0.97435
movie	0.99998
murder	0.95727
museum	0.99949
music	0.99995
nation	0.98680
nature	0.99997
network	0.99847
news	0.99992
noon	0.96682
number	1.00000
nurse	0.95429
object	0.99499
observation	0.96131
office	0.99473
oil	0.99758
opera	0.99667
operation	0.97030
oracle	0.90012
organism	0.98436
paper	0.99981
party	0.99904
payment	0.98169
peace	0.99105
people	0.99967
percent	0.97931
performance	0.99749
personnel	0.96458
phone	0.99999
physics	0.98076
pillow	0.98624
place	0.99909
plan	0.99924
plane	0.99848
planet	0.99856
planning	0.99441
popcorn	0.99623
Continue on the next page	

WORD	SCORE
population	0.99302
possession	0.85838
possibility	0.85824
potato	0.98657
practice	0.97334
precedent	0.99995
prejudice	0.87006
preparation	0.96405
preservation	0.93071
president	0.99082
price	0.99947
problem	0.99109
production	0.99781
professor	0.99256
profit	0.85738
project	0.99840
prominence	0.94706
property	0.99865
proton	0.99989
proximity	0.91276
psychiatry	0.82496
psychology	0.94845
quarrel	0.68294
queen	0.99676
rabbi	0.95747
racism	0.95325
racket	0.97329
radio	0.99878
reason	0.98841
recess	0.90531
recognition	0.95441
recommendation	0.94597
record	0.99274
recovery	0.94993
registration	0.99787
report	0.99945
reservation	0.97645
rock	0.99876
rook	0.88379
Continue on the next page	

WORD	SCORE
rooster	0.99141
round	0.99585
row	0.98902
sage	0.96477
school	1.00000
science	0.99845
scientist	0.96872
sea	0.99971
seafood	0.96556
season	0.99927
secret	0.99638
secretary	0.54860
seepage	0.94921
senate	0.76639
serf	0.88594
serial	0.99710
series	0.99978
seven	0.99571
sex	0.99984
shore	0.98310
shower	0.99916
sign	0.99979
signature	0.99448
similarity	0.97024
situation	0.93133
size	0.99999
skin	0.99760
slave	0.89951
smart	0.99921
smile	0.99898
soap	0.87436
soccer	0.99988
software	0.99854
space	0.99990
sprint	0.99155
star	0.99979
start	0.99715
stock	0.99975
stove	0.93449
Continue on the next page	

WORD	SCORE
street	0.99828
string	0.98284
stroke	0.99016
student	0.99274
stupid	0.98347
substance	0.85060
sugar	0.99351
summer	0.99948
sun	0.99797
surface	0.99540
tableware	0.99959
team	0.99811
telephone	0.98833
television	0.99934
tennis	0.99907
term	0.99482
territory	0.98484
terror	0.99970
theater	0.98177
thunderstorm	0.99762
ticket	0.98273
tiger	0.99953
tool	0.99064
tournament	0.99283
trading	0.97838
train	0.99683
travel	0.99990
treatment	0.98177
troops	0.99077
tumbler	0.98820
type	0.99962
valor	0.99971
victim	0.92915
victory	0.99028
video	1.00000
viewer	0.99966
virtuoso	0.94609
vodka	0.99684
volunteer	0.97888
Continue on the next page	

WORD	SCORE
voyage	0.99958
war	0.99925
warning	0.98795
water	0.99915
wealth	0.89861
weapon	0.99765
weather	0.99475
wine	0.98861
withdrawal	0.64496
wizard	0.99734
woman	0.99961
wood	0.99900
woodland	0.97030
word	0.99996
world	0.99987
year	0.99955
yen	0.97351
zoo	0.99802

BIBLIOGRAPHY

- [1] *Gcap: Graph-based automatic image captioning*, Proceedings of the 4th International Workshop on Multimedia Data and Document Engineering (MDDE) (Washington, DC), 2004.
- [2] Barton A., Sevcik R., and Ronski M., *Exploring visual-graphic symbol acquisition by pre-school age children with developmental and language delays*, Augmentative and Alternative Communication 22 (2006), 10–20.
- [3] Michael Argyle, *Bodily communication*, International Universities Press, 1988.
- [4] K. Barnard and D.A. Forsyth, *Learning the semantics of words and pictures*, Proceedings of the IEEE International Conference on Computer Vision, 2001.
- [5] K. Barnard, M. Johnson, and D. Forsyth, *Word sense disambiguation with pictures*, Proceedings of the HLT-NAACL 2003 Workshop on Learning Word Meaning from Non-Linguistic Data (Edmonton, Canada), 2003.
- [6] Kobus Barnard and David Forsyth, *Learning the semantics of words and pictures*, Proceedings of International Conference on Computer Vision, 2001.
- [7] Irving Biederman, *Recognition-by-components: A theory of human image understanding*, Psychological Review, vol. 94, 1987, pp. 115–147.
- [8] David Blei and John Lafferty, *A correlated topic model of science*, Annals of Applied Statistics, vol. 1, 2007, pp. 17–35.
- [9] A. Borman, R. Mihalcea, and P. Tarau, *Picnet: Augmenting semantic resources with pictorial representations*, Proceedings of the AAAI Spring Symposium on Knowledge Collection from Volunteer Contributors (Stanford, CA), 2005.
- [10] Elia Bruni, Giang Binh Tran, and Marco Baroni, *Distributional semantics from text and images*, Proceedings of the EMNLP Geometrical Models for Natural Language Semantics Workshop (2011).
- [11] Alexander Budanitsky and Graeme Hirst, *Evaluating wordnet-based measures of lexical semantic relatedness*, Computational Linguistics, vol. 32, 2005.
- [12] Gustavo Carneiro, Antoni Chan, Pedro Moreno, and Nuno Vasconcelos, *Supervised learning of semantic classes for image annotation and retrieval*, IEEE Trans. on Pattern Analysis and Machine Intelligence 29 (2006), no. 3.
- [13] Chih-Chung Chang and Chih-Jen Lin, *LIBSVM: a library for support vector machines*, 2001.
- [14] S. Chang and G. Polese, *A methodology and interactive environment for iconic language design*, Proceedings of the IEEE workshop on visual languages, 1992.

- [15] S. Clay and J. Wilhelms, *Put: Language-based interactive manipulation of objects*, IEEE Computer Graphics and Applications 16 (1996), no. 2, 31–39.
- [16] Brendan Collins, Jia Deng, Kai Li, and Li Fei-Fei, *Towards scalable dataset construction: An active learning approach*, Proceedings of European Conference on Computer Vision, 2008.
- [17] B. Coyne and R. Sproat, *Wordseye: An automatic text-to-scene conversion system*, Proceedings of the ACM Conference on Computer Graphics and Interactive Techniques (Los Angeles, CA), 2001, pp. 487–496.
- [18] Jia Deng, Wei Dong, Richard Socher, Lia-Ji Li, Kai Li, and Li Fei-Fei, *ImageNet: A Large-Scale Hierarchical Image Database*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
- [19] Koen Deschacht and Marie-Francine Moens, *Text analysis for automatic image annotation*, Proceedings of the Association for Computational Linguistics, 2007.
- [20] Pinar Duygulu, Kobus Barnard, Nando de Freitas, and David Forsyth, *Object recognition as machine translation: learning a lexicon for a fixed image vocabulary*, Proceedings of the 7th European Conference on Computer Vision, 2002.
- [21] Katrin Erk and Diana McCarthy, *Graded word sense assignment*, Proceedings of Empirical Methods in Natural Language Processing, 2009.
- [22] Ali Farhadi, Ian Endres, and Derek Hoiem, *Attribute-centric recognition for cross-category generalization*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2010.
- [23] Ali Farhadi, Ian Endres, Derek Hoiem, and David Forsyth, *Describing objects by their attributes*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
- [24] Li Fei-Fei and Pietro Perona, *A bayesian hierarchical model for learning natural scene categories*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2005.
- [25] Shaolei Feng, R Manmatha, and Victor Lavrenko, *Multiple bernoulli relevance models for image and video annotation*, International Conference on Computer Vision and Pattern Recognition, 2004.
- [26] Yansong Feng and Mirella Lapata, *Automatic image annotation using auxiliary text information*, Proceedings of the Association for Computational Linguistics, 2008.
- [27] ———, *Visual information in semantic representation*, Proceedings of the Annual Conference of the North American Chapter of the ACL, 2010.
- [28] Rob Fergus, Pietro Perona, and Andrew Zisserman, *Object class recognition by unsupervised scale-invariant learning*, Proceedings of the International Conference on Computer Vision and Pattern Recognition, 2003.
- [29] Edward A. Fox and Joseph A. Shaw, *Combination of multiple searches*, Proceedings of the 2nd Text REtrieval Conference (TREC-2), 1994.
- [30] Evgeniy Gabrilovich and Shaul Markovitch, *Computing semantic relatedness using wikipedia-based explicit semantic analysis*, International Joint Conferences on Artificial Intelligence, 2007.

- [31] Jacob Goldberger, Shiri Gordon, and Hayit Greenspan, *An efficient image similarity measure based on approximations of kl-divergence between two gaussian mixtures*, Proceedings of IEEE International Conference on Computer Vision, 2003.
- [32] Michael Grubinger, Clough Paul, Mller Henning, and Deselaers Thomas, *The iapr benchmark: A new evaluation resource for visual information systems*, International Conference on Language Resources and Evaluation, 2006.
- [33] Jonathon S. Hare, Sina Samangooei, Paul H. Lewis, and Mark S. Nixon, *Investigating the performance of auto-annotation and semantic retrieval using semantic spaces*, Proceedings of the international conference on content-based image and video retrieval, 2008.
- [34] Samer Hassan and Rada Mihalcea, *Semantic relatedness using salient semantic analysis*, Proceedings of AAAI Conference on Artificial Intelligence, 2011.
- [35] Graeme Hirst and David St-Onge, *Lexical chains as representations of context for the detection and correction of malapropisms*, WordNet: An Electronic Lexical Database, MIT Press, 1998, pp. 305–332.
- [36] Aminul Islam and Diana Inkpen, *Semantic Similarity of Short Texts*, Recent Advances in Natural Language Processing V (Nicolas Nicolov, Galia Angelova, and Ruslan Mitkov, eds.), Current Issues in Linguistic Theory, vol. 309, John Benjamins, Amsterdam & Philadelphia, 2009, pp. 227–236.
- [37] Mario Jarmasz, *Rogets thesaurus as a lexical resource for natural language processing*, Ph.D. Dissertation (Ottawa-Carleton Institute for Computer Science, School of Information Technology and Engineering, University of Ottawa), 2003.
- [38] Jiwoon Jeon, Victor Lavrenko, and R Manmatha, *Automatic image annotation and retrieval using cross-media relevance models*, Proceedings of the ACM SIGIR Conference on Research and Development in Information Retrieval, 2003.
- [39] Jay J. Jiang and David A. Conrath, *Semantic similarity based on corpus statistics and lexical taxonom*, Proceedings of International Conference Research on Computational Linguistics (ROCLING X), 1997.
- [40] Yohan Jin, Latifur Khan, Lei Wang, and Mamoun Awad, *Image annotations by combining multiple evidence & wordnet*, Proceedings of Annual ACM Multimedia, 2005.
- [41] R. Johansson, A. Berglund, M. Danielsson, and P. Nugues, *Automatic text-to-scene conversion in the traffic accident domain*, Proceedings of the Nineteenth International Joint Conference on Artificial Intelligence (Edinburgh, Scotland), 2005, pp. 1073–1078.
- [42] Pentti Kanerva, *Sparse distributed memory*, MIT Press, 1998.
- [43] Anita Komlodia, Weimin Houb, Jenny Preeceb, Allison Druinb, Evan Golubb, Jade Alburob, Sabrina Liaob, Aaron Elkissb, and Philip Resnikb, *Evaluating a cross-cultural childrens online book community: Lessons learned for sociability, usability, and cultural exchange*, Interacting with Computers, vol. 19, 2007, pp. 494–511.

- [44] Miroslav Kubat and Stan Matwin, *Addressing the curse of imbalanced training sets: one-sided selection*, Proceedings of International Conference on Machine Learning, 1997.
- [45] Christoph H. Lampert, Hannes Nickisch, and Stefan Harmeling, *Learning to detect unseen object classes by between-class attribute transfer*, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, 2009.
- [46] Thomas Landauer and Susan Dumais, *A solution to platos problem: The latent semantic analysis theory of acquisition*, Psychological Review, vol. 104, 1997, pp. 211–240.
- [47] Claudia Leacock and Martin Chodorow, *Combining local context and wordnet similarity for word sense identification*, The MIT Press, 1998, pp. 265–283.
- [48] Chee Wee Leong and Rada Mihalcea, *Going beyond text: A hybrid image-text approach for measuring word relatedness*, Proceedings of International Joint Conference on Natural Language Processing, 2011.
- [49] ———, *Measuring the semantic relatedness between words and images*, Proceedings of International Conference on Computational Semantics, 2011.
- [50] Chee Wee Leong, Rada Mihalcea, and Samer Hassan, *Text mining for automatic image tagging*, Proceedings of the International Conference on Computational Linguistics, 2010.
- [51] Jia Li and James Wang, *Real-time computerized annotation of pictures*, Proceedings of International Conference on Computer Vision, 2008.
- [52] Li-Jia Li and Li Fei-Fei, *Optimol: automatic online picture collection via incremental model learning*, International Journal of Computer Vision, 2008.
- [53] Wei Li and Andrew McCallum, *Pachinko allocation: Dag-structured mixture models of topic correlations*, Proceedings of the International Conference on Machine learning, 2006.
- [54] Dekang Lin, *An information-theoretic definition of similarity*, Proceedings of the Fifteenth International Conference on Machine Learning, 1998.
- [55] David Lowe, *Distinctive image features from scale-invariant keypoints*, International Journal of Computer Vision, 2004.
- [56] Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar, *A new baseline for image annotation*, Proceedings of European Conference on Computer Vision, 2008.
- [57] Diana McCarthy and Roberto Navigli, *The semeval English lexical substitution task*, Proceedings of the ACL Semeval workshop, 2007.
- [58] Rada Mihalcea, Courtney Corley, and Carlo Strapparava, *Corpus-based and knowledge-based measures of text semantic similarity*, Proceedings of Association for the Advancement of Artificial Intelligence, 2006, pp. 775–780.

- [59] Rada Mihalcea and Chee Wee Leong, *Towards communicating simple sentences using pictorial representations*, Machine Translation, vol. 22, 2009, pp. 153–173.
- [60] Rada Mihalcea and Paul Tarau, *Textrank: Bringing order into texts*, Proceedings of Empirical Methods in Natural Language Processing, 2004.
- [61] G. Miller, *Wordnet: A lexical database*, Communication of the ACM 38 (1995), no. 11, 39–41.
- [62] G. Miller and W. Charles, *Contextual correlates of semantic similarity*, Language and Cognitive Processes 6 (1998), no. 1.
- [63] George Miller, *Wordnet: A lexical database for english*, Communications of the ACM, vol. 38, 1995, pp. 39–41.
- [64] Ted Pedersen, Siddharth Patwardhan, and Jason Michelizzi, *Extended gloss overlaps as a measure of semantic relatedness*, Proceedings of International Joint Conference on Artificial Intelligence, 2003.
- [65] ———, *Wordnet::similarity - measuring the relatedness of concepts*, Proceedings of the Nineteenth National Conference on Artificial Intelligence, 2004, pp. 1024–1025.
- [66] Mary C. Potter and Babara A. Faulconer, *Time to understand pictures and words*, Nature, vol. 253, 1975, pp. 437–438.
- [67] M.C. Potter, J.F. Kroll, B. Yachzel, E. Carpenter, and J. Sherman, *Pictures in sentences: understanding without words*, Journal of Experimental Psychology 115 (1986), no. 3, 281–294.
- [68] Philip Resnik, *Using information content to evaluate semantic similarity in a taxonomy*, Proceedings of the International Joint Conference on Artificial Intelligence, 1995.
- [69] Bernice E. Rogowitz, Thomas Fresez, John R. Smithy, Charles A. Bouman, and Edward Kaliny, *Perceptual image similarity experiments*, Human Vision and Electronic Imaging III, 1998.
- [70] Daniel B. Russakoff, Carlo Tomasi, Torsten Rohlfing, and Calvin R. Maurer Jr., *Image similarity using mutual information of regions*, European Conference on Computer Vision, 2004.
- [71] Olga Russakovsky and Li Fei-Fei, *Attribute learning in large-scale datasets*, European Conference of Computer Vision (ECCV), International Workshop on Parts and Attributes (Crete, Greece), September 2010.
- [72] G. Salton, A. Wong, and C.S. Yang, *A vector space model for automatic indexing*, Readings in Information Retrieval, Morgan Kaufmann Publishers, San Francisco, CA, 1997, pp. 273–280.
- [73] Helmut Schmid, *Probabilistic part-of-speech tagging using decision trees*, Proceedings of the International Conference on New Methods in Language Processing, 1994.
- [74] Jitendra Sharma, Alessandra Angelucci, and Mriganka Sur, *Induction of visual orientation modules in auditory cortex*, Nature, vol. 404, 2000, pp. 841–847.
- [75] Arnold W.M. Smeulders, Marcel Worring, Simone Santini, Amarnath Gupta, and Ramesh Jain, *Content-based image retrieval at the end of the early years*, IEEE Transactions on Pattern Analysis and Machine Intelligence, vol. 22, 2000, pp. 1349–1380.

- [76] Rion Snow, Brendan O'Connor, Daniel Jurafsky, and Andrew Ng, *Cheap and fast - but is it good? evaluating non-expert annotations for natural language tasks*, Proceedings of Empirical Methods in Natural Language Processing, 2008.
- [77] Srihari and Burhans, *Visual semantics: Extracting visual information from text accompanying pictures*, Proceedings of the American Association for Artificial Intelligence, 1994.
- [78] Munirathnam Srikanth, Joshua Varner, Mitchell Bowden, and Dan Moldovan, *Exploiting ontologies for automatic image annotation*, Proceedings of the ACM Special Interest Group on Research and Development in Information Retrieval, 2005.
- [79] Peter D. Turney and Patrick Pantel, *From frequency to meaning: Vector space models of semantics*, Journal of Artificial Intelligence Research, vol. 37, 2010, pp. 141–188.
- [80] 2006, http://en.wikipedia.org/wiki/Visual_language.
- [81] Luis von Ahn and Laura Dabbish, *Labeling images with a computer game*, Proceedings of the ACM Special Interest Group on Computer Human Interaction, 2004.
- [82] Laurie von Melchner, Sarah L. Pallas, and Mriganka Sur, *Visual behaviour mediated by retinal projections directed to the auditory pathway*, Nature, vol. 404, 2000, pp. 871–876.
- [83] Chong Wang, David Blei, and Li Fei-Fei, *Simultaneous image classification and annotation*, Proceedings of IEEE Conference on Computer Vision and Pattern Recognition, 2009.
- [84] Thijs Westerveld, *Image retrieval: Context versus context*, Content-Based Multimedia Information Access, 2000.
- [85] Dominic Widdows and Kathleen Ferraro, *Semantic vectors: a scalable open source package and online technology management application*, Proceedings of the Sixth International Language Resources and Evaluation (LREC'08), 2008.
- [86] A. Yamada, T. Yamamoto, H. Ikeda, T. Nishida, and S. Doshita, *Reconstructing spatial image from natural language texts*, Proceedings of the 14th International Conference on Computational linguistics (Nantes, France), 1992, pp. 1279–1283.
- [87] Jun Yang, Yu-Gang Jiang, Alexander G. Hauptmann, and Chong-Wah Ngo, *Evaluating bag-of-visual-words representations in scene classification*, ACM Multimedia Information Retrieval Workshop, 2007.
- [88] Rong Zhao and William Grosky, IEEE Transactions on Multimedia, 2002.
- [89] X. Zhu, A. B Goldberg, M. Eldawy, C. R. Dyer, and B. Strock, *A text-to-picture synthesis system for augmenting communication*, Proceedings of the American Association for Artificial Intelligence: Integrated Intelligence Track (Vancouver, Canada), 2007, pp. 1590–1595.
- [90] Xiaojin Zhu, Andrew Goldberg, Mohamed Eldawy, Charles Dyer, and Bradley Strock, *A text-to-picture synthesis system for augmenting communication*, Integrated Intelligence Track of the Twenty-Second AAAI Conference on Artificial Intelligence, 2007.