Computer Vision and Natural Language Processing: Recent Approaches in Multimedia and Robotics

PERATHAM WIRIYATHAMMABHUM, University of Maryland, College Park
DOUGLAS SUMMERS-STAY, U.S. Army Research Laboratory, Adelphi
CORNELIA FERMÜLLER and YIANNIS ALOIMONOS, University of Maryland, College Park

Integrating computer vision and natural language processing is a novel interdisciplinary field that has received a lot of attention recently. In this survey, we provide a comprehensive introduction of the integration of computer vision and natural language processing in multimedia and robotics applications with more than 200 key references. The tasks that we survey include visual attributes, image captioning, video captioning, visual question answering, visual retrieval, human-robot interaction, robotic actions, and robot navigation. We also emphasize strategies to integrate computer vision and natural language processing models as a unified theme of distributional semantics. We make an analog of distributional semantics in computer vision and natural language processing as image embedding and word embedding, respectively. We also present a unified view for the field and propose possible future directions.

Categories and Subject Descriptors: I.2.0 [**Artificial Intelligence**]: General; I.2.7 [**Artificial Intelligence**]: Natural Language Processing; I.2.9 [**Artificial Intelligence**]: Robotics; I.2.10 [**Artificial Intelligence**]: Vision and Scene Understanding; I.4.9 [**Image Processing and Computer Vision**]: Applications; I.5.4 [**Pattern Recognition**]: Applications

General Terms: Computer Vision, Natural Language Processing, Robotics

Additional Key Words and Phrases: Language and vision, survey, multimedia, robotics, symbol grounding, distributional semantics, computer vision, natural language processing, visual attribute, image captioning, imitation learning, word2vec, word embedding, image embedding, semantic parsing, lexical semantics

ACM Reference Format:
Peratham Wiriyathammabhum, Douglas Summers-Stay, Cornelia Fermüller, and Yiannis Aloimonos. 2016. Computer vision and natural language processing: Recent approaches in multimedia and robotics. ACM Comput. Surv. 49, 4, Article 71 (December 2016), 44 pages.
DOI: http://dx.doi.org/10.1145/3009906

1. INTRODUCTION

We have many ways to describe the world for communication between people: texts, gestures, sign languages, and face expressions are all ways of sharing meaning. Language is unique among communication systems in that its compositionality through syntax allows a limitless number of meanings to be expressed. Such meaning ultimately must be tied to perception of the world. This is usually referred to as the *symbol*

An earlier version of this article appeared as "Computer Vision and Natural Language Processing: Recent Approaches in Multimedia and Robotics," Scholarly Paper Archive, Department of Computer Science, University of Maryland, College Park, MD, 20742.
Authors' addresses: P. Wiriyathammabhum, C. Fermüller, and Y. Aloimonos, Computer Vision Lab, University of Maryland College Park, MD 20742-3275; email: {peratham@cs.umd.edu, fer@umiacs.umd.edu, yiannis@cs.umd.edu}. D. Summers-Stay, U.S. Army Research Laboratory, Adelphi, MD 20783; email: {douglas.a.summers-stay.civ@mail.mil}.

DOI: http://dx.doi.org/10.1145/3009906

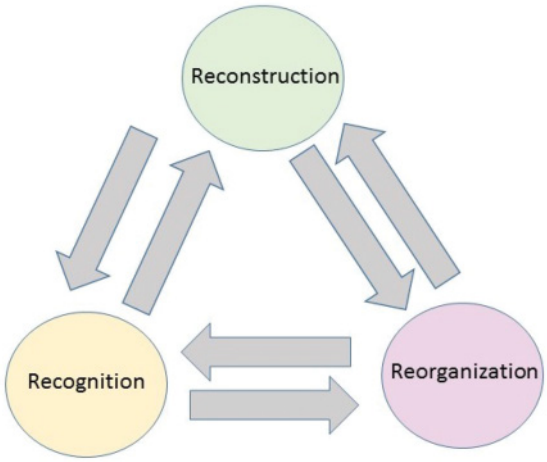

Fig. 1. The 3Rs in computer vision [Malik et al. 2016], which are reconstruction, reorganization, and recognition.

grounding problem [Harnad 1990]. Language without perception would be a pure abstraction; perception without language, on the other hand, is unable to go far beyond simple conditioned responses.

In human perception, visual information is the dominant modality for acquiring knowledge of the world—about 30% of the human brain is dedicated to visual processing. To what extent language is directly involved in the visual process is still a matter of debate. However, in the attempt to achieve artificial intelligence, making use of certain aspects of language provides interpretability and enables productive human-machine interaction. In order to bridge language and vision, let us first revisit some of the major tasks in both language and vision.

1.1. Computer Vision Tasks and Their Relationships to Natural Language Processing

Computer Vision (CV) tasks can be summarized by the concept of 3Rs [Malik et al. 2016], which are *reconstruction*, *recognition*, and *reorganization*. Reconstruction involves estimating the three-dimensional (3D) scene that gave rise to a particular visual image. It can be accomplished using a variety of processes incorporating information from multiple views, shading, texture, or direct depth sensors. Reconstruction process result in a 3D model, such as point clouds or depth images. Some examples for reconstruction tasks are Structure from Motion, scene reconstruction, and shape from shading. Recognition involves both 2D problems (like handwritten recognition, face recognition, scene recognition, or object recognition), and 3D problems (like 3D object recognition from point clouds which assists in robotics manipulation). Recognition results in assigning labels to objects in the image. *Reorganization* involves bottom-up vision: segmentation of the raw pixels into groups that represent the structure of the image. Reorganization tasks range from low-level vision like edge, contour, and corner detection, intrinsic images, and texture segmentation to high-level tasks like semantic segmentation [Tu et al. 2005; Carreira and Sminchisescu 2010; Socher et al. 2011], which has an overlapping contribution to recognition tasks. A scene can be segmented based on low-level vision [Martin et al. 2004; Teo et al. 2015] or high-level information like shadow segmentation [Horprasert et al. 1999; Ecins et al. 2014] that utilizes class information. All of these tasks can be viewed as fact finding from visual data. In the

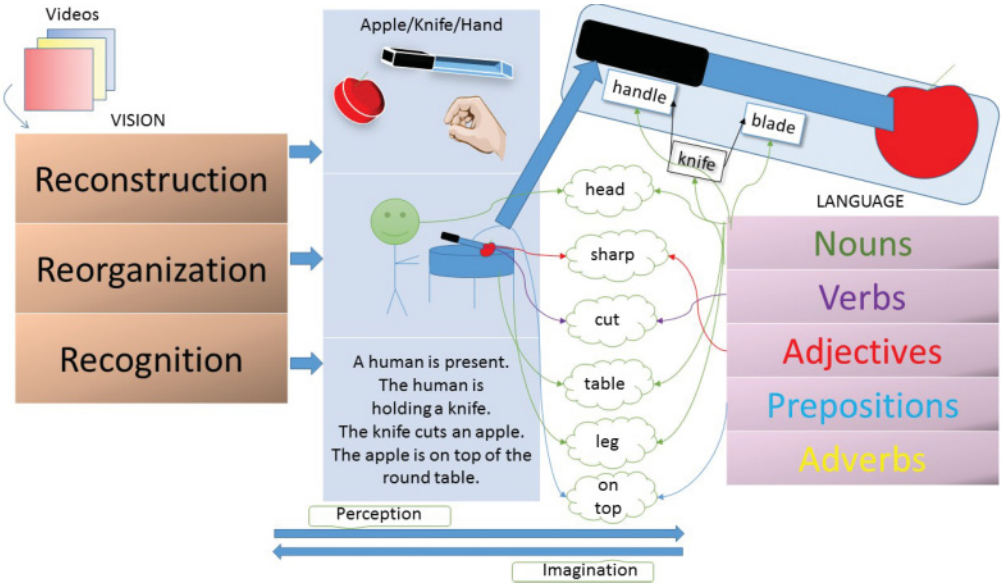

Fig. 2. Connecting the 3Rs in computer vision to language by way of semantic information. In the scene, a man cuts a red apple with a knife. There are many interpretation of the depicted image to be recognized from a cartoon level of perception by organization from reconstructed object models. Semantic representations can provide information from language to guide the interpretation of the objects, actions, events, and relations in the scene.

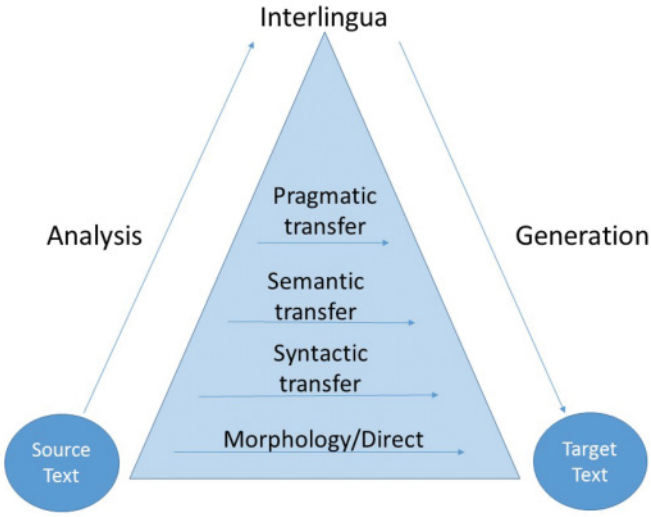

Fig. 3. The Vauquois triangle for machine translation [Vauquois 1968] with added pragmatic transfer. Pragmatics also captures meaning but it goes beyond standard semantics to consider context [Levinson 2001].

words of David Marr, "vision is the process of discovering from images what is present in the world, and where it is [Marr 1982]."

The output from one task can provide information that helps another task, so reconstruction, reorganization, and recognition can each support the other two. For example,

3D faces that are the outputs of a reconstruction task can give more information and assist face recognition [Bronstein et al. 2005]. On the other hand, recognition can give prior knowledge to create an object-specific 3D model for a reconstruction task [Barron and Malik 2015]. For reorganization and recognition, reorganization can provide contours, regions, and object candidates for object recognition [Pont-Tuset et al. 2016]. In the opposite direction, recognition can generate object proposal regions for segmentation [Hariharan et al. 2014]. This can also be viewed as recognition providing context for reorganization [Heitz and Koller 2008]. For reorganization and reconstruction, a symmetry assumption can guide reorganization and reconstruction for 3D cluttered scene segmentation [Ecins et al. 2016]. The best ways to make use of low-level features such as edges or contours to provide information to reconstruction and vice versa are links still under investigation.

Recognition tasks are most closely connected to language since the output is likely to be interpretable as words. *Semantic representations* [Gärdenfors 2014; Gupta 2009] connect language and vision in this scenario. For example, object or scene classes can be represented by concrete nouns, activities by a subclass of verbs, and object attributes by adjectives. Relations between objects or object and scene are often expressed by prepositions. Temporal relations of an object and an activity can be seen as adverbs. Reorganization tasks deal with a lower-level feature set that can be interpreted as primitive parts of shapes, textures, colors, regions, and motions. These primitive parts build up to higher-level vision: They do not typically refer to any specific object or scene that can be described in words, but they are essential for learning new words as they implicitly describe object or scene properties. Reconstruction sometimes involves 3D geometry and real-world physics, which provide richer object or scene properties than reorganization tasks. Reconstruction is currently used for real-time high-precision robotics that make use of the recovered depth dimension. This is essential for robot action manipulation [Maldonado et al. 2012].

1.2. Natural Language Processing Tasks and Their Relationships to Computer Vision

Following the *Vauquois triangle* for machine translation [Vauquois 1968], Natural Language Processing (NLP) tasks can be summarized into concepts ranging from *syntax* to *semantics* and to *pragmatics* at the top level to achieve communication. Syntax includes morphology (the study of word forms) and compositionality (the composition of smaller language units like words to larger units like phrases or sentences). Semantics is the study of meaning, including finding finding relations between words, phrases, sentences or discourse. Pragmatics studies how meaning changes in the presence of a specific context. For instance, an ironic sentence cannot be correctly interpreted without any side information that indicates the indirectness in the speaker's intention. Ambiguity in language interpretation a main obstacle for an intelligent system to overcome and achieve language understanding [Quiroga-Clare 2003]. Some complex tasks in NLP include machine translation, information extraction, dialog interface, question answering, parsing, and summarization.

There is always meaning lost when translating between one language and another. When "translating" between the low-level pixels or contours of an image and a high-level description in word labels or sentences, there is a wide chasm to be crossed. *Bridging the Semantic Gap* [Zhao and Grosky 2002] means building a bridge from visual data to language data like words or phrases. To give some specific examples, labeling an image patch that contains an object with a word is object recognition. Labeling a background in an image is scene recognition. Assigning words for pixel grouping is semantic segmentation [Tu et al. 2005; Carreira and Sminchisescu 2010; Socher et al. 2011]. If we know how the words are related to each other, then it can give a clue for visual processing to better disambiguate different visual constructs. For

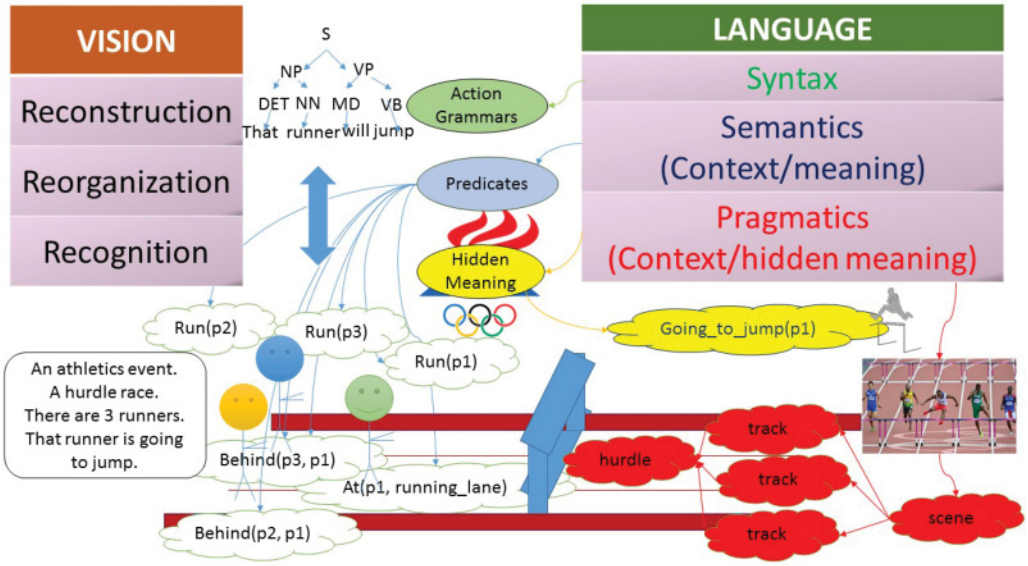

Fig. 4. From the Vanquois triangle in NLP to computer vision by commonsense reasoning. Suppose we have a video of a sporting event and we want to predict what will happen next. Using commonsense knowledge, the system can focus on the relevant actors, objects, and actions that can be thought of given a situation as its attention and awareness mechanisms that involve both language and vision. The system can predict the next action that is *that runner will jump* from language (compositional semantics—action grammar), vision (lexical semantics—recognitions), and logical inference (formal semantics—from facts).

instance, a *knife* is more likely to *cut* a *cucumber* than a *chair* in a *kitchen* because the meaning of their interaction is presented as a concept in the real world.

1.3. Categorization of the Current State in Vision and Language

The development of techniques devoted to the integration of vision and language did not happen in a top down deliberate manner, where researchers came up with a set of principles which then defined the subsequent details of the research projects. Those techniques were rather developed in a more bottom-up way, where some pioneers identified particular problems in specific domains, attempted many candidate solutions, and reported a satisfactory outcome. To impose some order on this situation, we have created a matrix where the rows represent vision and the columns represent language and reasoning (Tables I and II).

The 3Rs of vision described above divide the topic of vision well, and so each row corresponds to each of the 3Rs. The columns come from the structure of Language and Reasoning: lexical semantics, compositional semantics, formal semantics, and distributional semantics. Lexical semantics have to do with parts of speech: nouns, adjectives, verbs, prepositions, and so on. Compositional semantics are related to parsing and grammars and formal semantics to predicate and λ-calculus. Finally, distributional semantics concern the use of latent variables (word2vec, embeddings, deep learning etc.). As an example, looking at the third row and first column square, we find many works where words are associated with visual data for the purpose of recognition. Such works are of course at the high visual level, but the association could also be used in the lower levels, such as the first row or the second. Similarly, the problems of image and video captioning or visual story telling fall into the intersection of the third row with

Table I. Our Proposed Categorization for Recent Vision and Language Literatures

Vision \ Language & Reasoning	Lexical Semantics (Part-of-speech)	Compositional Semantics (Parsing)	Formal Semantics (λ-calculus)	Distributional Semantics (Latent Variables)
Reorganization Contour Extraction Segmentation	-Semantic Segmentation -Class-based Contour Section 1.1	-Scene Parsing	[Semantic Parsing on Early Vision?]	-DSMs on Colors -Deep Structured Prediction Section 4.4.2
Reconstruction 3D models	-3D Attributes -3D Semantic Segmentation	-Parsing 3D Scene -Referring Expression Section 3.3	-Semantic Parsing for Referring Expression Section 3.3	-[3D DSMs ?] -3D Deep Learning
Recognition Object Scene Action Event	-Attributes -Affordance -Spatial Relation -Force and Condition Section 2.1.1 and 3.3	-Visual Captioning -Story Telling -Attribute Grammar -Action Grammar Section 2.2 and 2.1.1	-Visual Question Answering Section 2.3	-DSMs on Descriptors -DSMs on Attributes -Deep Learning, Attention and Memory Section 4.4.2

The Rows Represents Vision and the Columns Represents Language and Reasoning. Each Cell Refers to Concepts That Involve Vision and Language as Long as the Corresponding Section Number in This Survey Article. The Interesting Questions That Might Not Have Been Asked Are in Brackets.

Table II. The Taxonomy of Representative Works According to Our Proposed Categorization for Recent Vision and Language Literatures in Table I

Vision Task	Language/Reasoning	
	Lexical Semantics	Compositional Semantics
Reorganization	[Uijlings and Ferrari 2015]	[Socher et al. 2011]
	[Myers et al. 2015]	[Tu et al. 2005]
		[Tighe and Lazebnik 2010]
Reconstruction	[Fouhey et al. 2016]	[Liu et al. 2014b]
	[Gupta et al. 2015]	[Chang et al. 2014]
	[Lai and Fox 2010]	
Recognition	[Barnard and Forsyth 2001]	[Pastra and Aloimonos 2012]
	[Farhadi et al. 2009]	[Aksoy et al. 2011]
	[Sadeghi and Farhadi 2011]	[Farhadi et al. 2010]
	[Yao et al. 2011]	[Xu et al. 2015a]
		[Park et al. 2016]
		[Yu and Siskind 2013]

Vision Task	Language/Reasoning	
	Formal Semantics	Distributional Semantics
Reorganization	[Missing work]	[Bruni et al. 2012]
		[McMahan and Stone 2015]
		[Long et al. 2015]
Reconstruction	[Fang et al. 2013]	[Wu et al. 2015a]
	[Thomason et al. 2015]	[Dosovitskiy et al. 2015]
Recognition	[Aditya et al. 2015]	[Sukhbaatar et al. 2015]
	[Kojima et al. 2002]	[Yeung et al. 2016]
	[Antol et al. 2015]	[Bruni et al. 2014]
	[Zettlemoyer and Collins 2005]	[Silberer et al. 2013]
	[Poon and Domingos 2009]	[Srivastava and Salakhutdinov 2014]
		[Andreas et al. 2016b]
		[He et al. 2016]
		[Hand and Chellappa 2016]

The Rows Represents Vision and the Columns Represents Language and Reasoning. Each Cell Refers to Works That Involve Vision and Language.

the column on compositional semantics. Finally, the table not only classifies existing research but also predicts interesting questions that might not have been asked.

This article is structured as follows. Section 2 will provide a survey of language and vision literature in the multimedia domain. Section 3 will provide another survey of language and vision works in robotics applications. Section 4 will focus on distributional semantics in language and vision. Section 5 will summarize the state-of-the-art in combining language and vision and state current limitations. Finally, Section 6 will conclude and provide suggestions for future directions.

2. LANGUAGE AND VISION FOR MULTIMEDIA

Multimedia files contain images, videos, and natural language texts. They are often harvested from the Internet. For example, a news article may contain news that was written by a journalist and a photo related to the news content. There may be a clip video that contains a reporter, and a video that depicts the snapshot of the scene where the event in the news occurred. The co-occurrence between an image and texts signals

Table III. Bloom's Taxonomy Applied to Vision and Language. The Framework Amounts to Developing Systems That Can Answer Questions About a Scene (Image/Video). These Questions Form a Hierarchy Starting from Simpler Ones Whose Answer Can Be Easily Discovered in the Scene Itself to Progressively More Difficult Questions Whose Answer Must Be Inferred, Planned, Visualized, or Reasoned. Many of the Harder Questions Have to Do With Intention, Causation, Prediction, and Utility Optimization

Bloom's Taxonomy	**Vision and Language Understanding through asking questions about the scene (images/videos)**
KNOWLEDGE	What and where questions. The system should recognize nouns, attributes, verbs, action and label events (I have seen this before)
COMPREHENSION	Why and whats next questions. The system should recognize intention, causality and should be able to predict the immediate future (I can predict what will happen).
APPLICATION	How to questions. The system should be able to perform a command given in language (like stir the coffee)
ANALYSIS	How much, how many, how similar, how different questions. The system should be able to recognize and relate the constituent components of an event.
SYNTHESIS	How to achieve a goal questions. The system should be able to produce plans by using the event space.
EVALUATION	How good, how fast, how accurate questions. The system should be able to optimize in the space of events (or plans)

that they are related. An image in the news [Berg et al. 2004], for example, is likely to have a face, and the accompanying news text will help identify the face, including a lot of side information about that person such as occupation, age or gender. This requires that the coreference between the news text and the face can be accurately discovered. Another example of finding the relationship between images and texts is entry-level categorization [Ordonez et al. 2013] by finding its natural description, the way people refer to an object in practice. A word makes sense when used to describe a specific object depending on contexts. That word can be easily used for an appropriate and unambiguous communication if it is the word people use naturally.

Language and visual data provide two sets of information that are combined into the whole story. This conforms to the theory of *semiotics* [Greenlee 1978], which is the study of the relations of signs and their meanings. Semiotics has an analogous viewpoint to the NLP concepts from the previous section [Morris 1938]. First, semiotics studies the relationship between signs and meaning—the semantics of signs. Second, the formal relation between signs is equivalent to syntax. Third, the way humans interpret signs in context is equivalent to pragmatics. If we consider purely visual signs, then this leads to the conclusion that semiotics can also be approached by computer vision, extracting interesting signs for NLP to realize the corresponding meanings.

The tasks for language and vision in multimedia mainly fall into three categories: visual properties description, visual description, and visual retrieval.

2.1. Visual Properties Description

2.1.1. Attribute-Based Vision. Associating words and pictures [Barnard and Forsyth 2001] is a form of the recognition task in CV. Object recognition traditionally tries to categorize an image to a fixed set of name tags. Farhadi [2011] argues that an image

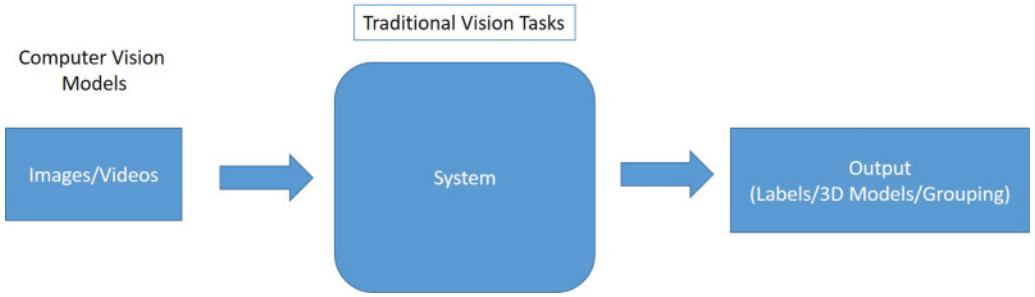

Fig. 5. The traditional computer vision framework which passively outputs language.

has more information than just a set of name tags, and categorization should change to description. Attribute-based recognition [Farhadi et al. 2009; Lampert et al. 2009; Kumar et al. 2009] describes and summarizes object *properties* in words in which an unusual property of that object can be detected and recognizing novel objects can be done with a few or zero training examples from category textual descriptions. The attributes may be binary values for some easily recognizable properties like *four-legged animal* or *walking left*. However, some properties may not be easily recognizable, like smiling. For these types of properties, the relative attributes [Parikh and Grauman 2011] help describe the strength of the property in each image by using a learning-to-rank framework [Liu 2009].

The key is that the attributes will provide a set of contexts as a knowledge source for recognizing a specific object by its properties. The attributes can be discovered using a classifier that learns a mapping from an image to each property [Farhadi et al. 2009; Lampert et al. 2009]. The attribute words become an intermediate representation that helps bridge the semantic gap between the visual space and the label space. In other words, the attributes are textual abstractions of an object. This introduces another dimension for feature engineering in which there are common features that can be shared across tasks [Farhadi et al. 2010], such as object parts, and some features that will be task-specific and unique for each task, such as the hardness of a diamond or the softness of gold when exposed to heat. Most attributes will be shared for objects in the same category. Hence, this enables a paradigm of zero-shot learning in which we can predict the class label of an unknown object given just the attributes. Zero-shot learning is very similar to text classification in NLP if attributes are thought of as words in a document, and we are trying to label that document.

From this viewpoint, we can incorporate the feature learning framework into attribute-based recognition. That is, we need to determine the possible set of attributes, whether they are words, phrases, or sentences. Also, which attributes should be used, to recognize what kind of object, becomes a feature selection problem that will impact the recognition performance explicitly. In addition, it can be further inferred that a specific set of words will correspond to a specific set of objects. Fu et al. [2015] proposed learned visual attributes with data in an unsupervised setting with feature selection but with visual words rather than textual words. Predicting visual attributes fall into the category of traditional computer vision framework (Figure 5). These traditional systems consider either images or videos as an input and create meaningful output attributes. For zero-shot learning, the system takes attributes as the input and output class labels.

Attributes can be incorporated into a more advanced feature learning framework. For parsing, Park et al. [2016] proposed an attribute and-or grammar (A-AOG) model that simultaneously predicts human body pose and human attributes in a parse graph

in which attributes are augmented to nodes in the parse tree. For multitask learning, Hand and Chellappa [2016] depicted an improved concept of attributes, where features in different convolutional neural networks are shared based on word relation.

Attribute-based vision was found useful in many specific applications where property contexts are crucial and informative, including animal recognition [Berg et al. 2006; Lampert et al. 2009], face recognition [Berg et al. 2004; Kumar et al. 2009], finding iconic images [Berg and Berg 2009], unsupervised attribute recovery on web data [Berg et al. 2010], and object identification for robotics [Sun et al. 2013]. Recognizing an image should result in a rich meaning that informatively describes what is going on in the image. Beyond words in isolation, phrases [Sadeghi and Farhadi 2011] and sentences [Farhadi et al. 2010; Yang et al. 2011] can expand more dimensions from an image.

2.1.2. Human Object Interaction (HOI) Prediction. Another aspect of predicting object properties is to *understand actions*. Any human action can be described by many different verbs, depending on the point of view of the speaker; conversely, each verb potentially describes a class of actions. In this way, verbs can behave for actions like attributes do for objects. Action verbs often involve human subjects and physical objects as their predicate arguments. (This differs from recognizing human activity using attributes [Liu et al. 2011], which is similar to object properties prediction.) To recognize actions, verbs and objects need to be recognized and human poses need to be estimated [Yao et al. 2011]. The recognition relies mostly on discriminative classifiers like SVMs. However, this is a special setting where class-imbalance and multilabel correlation should be considered. Verbs can encompass a variety of visually distinct actions: *cutting a tree* and *cutting a cucumber* involve different tools, poses, and tools. Some representative data sets for HOI prediction are Stanford action 40 [Yao et al. 2011], Trento Universal Human Object Interaction Dataset (TUHOI) [Le et al. 2014], and Humans Interacting with Common Objects (HICO) [Chao et al. 2015]. The datasets used before TUHOI and HICO were typically relatively small.

Gupta and Malik [2015] believe that just predicting HOI without semantic role understanding is not adequate, because such prediction lacks fine-grained information in the output. They introduced the V-COCO dataset in which people are shown performing multiple actions at once and assigns roles for each object that is involved with each action. In contrast to traditional HOI prediction, V-COCO encourages object detectors instead of discriminative classifiers. This serves as an introduction of semantic role labeling (which is a well-known task in NLP) to CV. Language resources like FrameNet [Baker et al. 1998] and the VerbNet [Schuler 2005] ontology are used as a commonsense knowledge source for NLP. Along the same line, Yatskar et al. [2016] introduces the imSitu dataset, and a structured prediction baseline: a Conditional Random Field (CRF) on the top of a Convolutional Neural Networks (CNN). The baseline shows its superiority to traditional discriminative classifiers like SVMs. The ability to reliably assign semantic roles to objects in images associated with actions will make this a useful type of tool on which to build intelligent responsive agents.

2.2. Visual Description

2.2.1. Image Captioning. The grammatical structure of a sentence provides a more informative description of an image than a bag of unordered words. Images often come with accompanying text that tells a story. For example, an image from a sport headline may depict the decisive moment of the game and the corresponding text will describe the details. To be assigned a concrete meaning, terms should have interpretations that are visually discoverable [Dodge et al. 2012]. Sentence generation systems may discard nouns or adjectives that are non-visual from their visual recognition results to reduce

bias errors. Scene information can also reduce bias errors in object recognition since only a specific set of objects will naturally occurs in a given scene [Yu et al. 2011].

Collecting captions from visually similar images can generate good descriptions. Ordonez et al. [2011] finds the best caption from the most visually similar image based on content matching, where the distance measurement consists of the object, people, stuf,f and scene detectors. Kuznetsova et al. [2012] goes further by summarizing the captions from the candidate similar images. The motivation for borrowing captions from similar images is that measuring similarity between visual features is easier than measuring in both visual and text features. Nearest-neighbor methods are shown to work well for image captioning by the remarkable automatic evaluation scores given a good embedding space by Kernel Canonical Correlation Analysis (KCCA) [Hodosh et al. 2013] or Neural Networks [Socher et al. 2014].

To *generate a sentence* describing an image, a certain amount of low-level visual information needs to be extracted. ⟨Objects, Actions, Scenes⟩ triplets can be used to represent meaning as a Markov Random Field of potential edges [Farhadi et al. 2010]. Then, the parameters are learned using human annotated examples. Considering part-of-speech, the quadruplets of ⟨Nouns, Verbs, Scenes, Prepositions⟩ can represent meaning extracted from visual detectors [Yang et al. 2011]. Visual modules extract objects that are either a subject or an object in the sentence. Then a Hidden Markov Model (HMM) is used to decode the most probable sentence from a finite set of quadruplets along with some corpus-guided priors for verb and scene (preposition) predictions. Li et al. [2011] represents meaning using objects (nouns), visual attributes (adjectives), and spatial relationships (prepositions) as ⟨⟨adj1, obj1⟩, prep, ⟨adj2, obj2⟩⟩. Then the sentence is generated by phrase fusion using web-scale n-grams for determining probabilities. Babytalk [Kulkarni et al. 2013] goes further by using CRF for predicting the best ⟨⟨adj1, obj1⟩, prep, ⟨ adj2, obj2⟩⟩ triplet. Then the output is decoded using a language model and generated as a template-based sentence.

Midge [Mitchell et al. 2012] makes an additional improvement by tying syntactic models to visual detections so the template is more relaxed and the sentence looks more natural. Visual Dependency Grammar [Elliott and Keller 2013] proposes dependency constraints, such as spatial relations of pixels, so the visual detection stage will have a structured output to be fed into a template-based generation system. This step reduces noise from object detectors and provides more stability given gold standard region annotations. Recent methods use a CNN to detect visual features and using Recurrent Neural Networks (RNNs) [Karpathy and Fei-Fei 2015a] or Long-Short Term Memory (LSTM) [Vinyals et al. 2015] to generate the sentence description. Both methods are implemented in the NeuralTalk2 system.[1] Xu et al. [2015a] incorporates the attention mechanism to learn to describe the salient object in an image. Another approach [Aditya et al. 2015; Wu et al. 2015b; Zhu et al. 2015b] uses commonsense knowledge to generate more humanlike sentences based on pragmatic knowledge.

Some standard datasets for image captioning are Flickr8k [Hodosh et al. 2013], Flickr30k [Young et al. 2014], and MS COCO [Chen et al. 2015a]. Other interesting datasets include the Amazon product data [McAuley et al. 2015a, 2015b], the Comprehensive Cars (CompCars) dataset [Yang et al. 2015c], and Visual Genome [Krishna et al. 2016]. For a recent trend of image captioning datasets, Ferraro et al. [2015] provides a detailed explanation along with an empirical evaluation across standard datasets.

Automatic evaluation metrics like BLEU [Papineni et al. 2002], ROUGE [Lin 2004], METEOR [Elliott and Keller 2013], or CIDEr [Vedantam et al. 2015], developed for evaluating natural language processing tasks like machine translation or text summarization, can be used for evaluating generated image captions. Elliott and Keller

[1]https://github.com/karpathy/neuraltalk2.

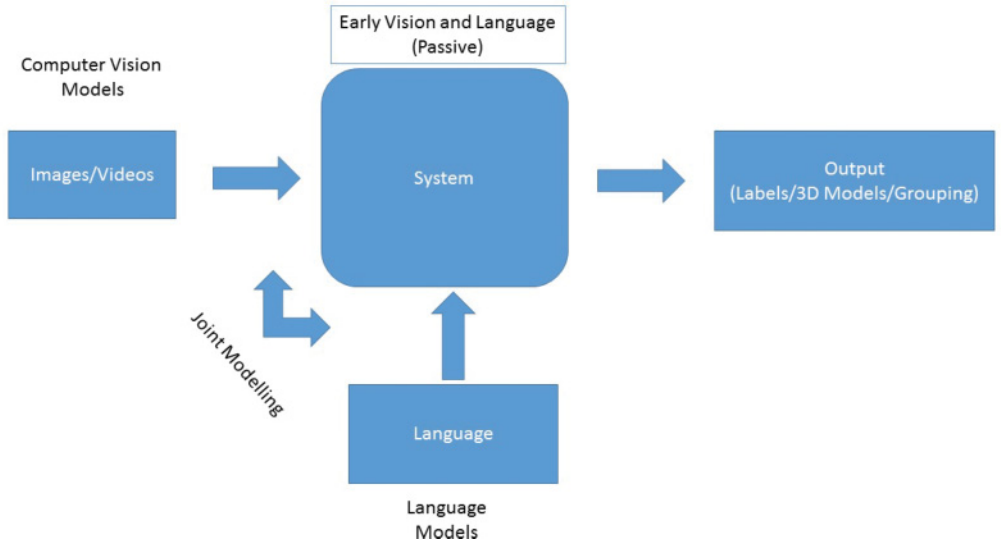

Fig. 6. The early vision and language framework which passively utilizes high-level language information as an additional information context.

[2014] provide a comparison among BLEU, ROUGE, and METEOR that reports a weak correlation with human judgments for BLEU. METEOR has the strongest correlation but is still far from real human judgments. Another evaluation approach is human evaluation using Amazon Mechanical Turks or CrowdFlower, which employs metrics like relevance and thoroughness [Aditya et al. 2015] or readability, correctness, and humanlikeness [Yang et al. 2015a]. Bernardi et al. [2016] provides another comparison on image captioning systems.

As a general framework (Figure 6), most methods in image captioning are trying to either model language information as another layer on top or jointly model language and vision simultaneously by a carefully designed loss function or algorithm. These systems consider structural multimodal input and create structural output in contrast to the traditional system.

2.2.2. Video Captioning. A video can be described by a sentence or a discourse that is a structured set of sentences that tells a specific story. A sentence prior can be learned from web-scale corpora to bias the model and penalize unlikely combinations of actors/actions/objects [Guadarrama et al. 2013]. Motion tracking for improved activity recognition, and event semantics (which describe the motion, force dynamics, action or activity that happens in a video under a specific situation) can both improve the results on activity understanding [Siskind 2001].

The sentence tracker [Barbu et al. 2012a, 2012b; Barrett et al. 2016; Yu and Siskind 2013; Siddharth et al. 2014; Yu et al. 2015b] is the first such system that has been proposed. It links the compositional structure of languages and the compositional structure of video events using natural language semantics and three essential computer vision tasks, which are tracking, object detection, and event recognition. These three tasks are done simultaneously using a single cost function including the attention mechanism to focus on the most salient event to produce the best sentence description for activity recognition.

From a natural language processing perspective, the sentence tracker utilizes lexical semantics and contains the information of "who did what to whom, and where and how

they did it" [Barbu et al. 2012a]. An object is described as a noun phrase; the observed action is described as a verb; object properties are described as adjectives; spatial relations between objects are described by prepositions; and the event characteristics are represented as prepositions and adverbs. The system has a predefined vocabulary and a sentence is composed using a set of predefined grammars.

The sentence tracker can be divided into two subsystems [Barbu et al. 2012a]. The first subsystem consists of three steps. First, object detection is performed with a high-recall setting. Second, tracking with forward projection to increase precision is performed. Third, the optimal set of detections is chosen using dynamic programming with the Viterbi algorithm, outputting a result consistent with the optical flow. For the second subsystem, events are recognized using HMMs, also computed using the Viterbi algorithm. The unified objective function from the final step of the first subsystem and the second subsystem can be merged, since both are based on HMMs. The sentence tracker has been tested in three scenarios: sentence-guided focus of attention [Barbu et al. 2012b; Siddharth et al. 2014; Yu et al. 2015b], sentence generation [Barbu et al. 2012a; Yu and Siskind 2013; Siddharth et al. 2014; Yu et al. 2015b], and video retrieval [Barrett et al. 2016; Siddharth et al. 2014; Yu et al. 2015b].

Video captioning can also use meaning representations such as action concept hierarchy [Kojima et al. 2002], ⟨Subject, Verb, Object⟩ triplets with text-mined knowledge [Krishnamoorthy et al. 2013], and semantic hierarchies [Guadarrama et al. 2013] or ⟨Subject, Verb, Object, Place⟩ quadruplets that use a factor graph for inference [Thomason et al. 2014]. These works try to emphasize the requirement of the ideal system that can handle videos in the wild with a large vocabulary and are not restricted to a limited domain, such as "cooking."

Recent approaches deploy powerful deep-learning frameworks to model both image and word sequences. These approaches can support a larger vocabulary than other methods that have a small set of predefined vocabulary [Venugopalan et al. 2015b]. Similar to image captioning, Venugopalan et al. [2015b] combines a sequence of CNN and another two sequences of LSTM to generate sentence description from video. AlexNet is deployed as a pretrained CNN model, and the output features are mean-pooled before feeding to the LSTM sentence decoding layer. This work is inspired by Donahue et al. [2015], which uses CRF to extract image features for the intermediate representation for an LSTM. Venugopalan et al. [2015a] makes an improvement to Venugopalan et al. [2015b] that discards the temporal information and models the image frames as a bag-of-images. It is able to add temporal information between image frames by feeding optical flow features to a CNN and incorporates temporal information with a modified LSTM. Venugopalan et al. [2015a] claims their S2VT system[2] as the first sequence-to-sequence model since the system can jointly handle a variable number of input frames, learn temporal structure of the video, and learn the language model. Xu et al. [2015b] makes a further improvement by using multi-scale image frames that use CNN on the whole image frame and fully connected convolutional neural network (FCN) for simulating receptive fields in small regions. Then, a Multiple Instance Learning framework is used to combine multiple CNN and FCN models. A similar idea is deployed in Johnson et al. [2016] that exhibits better object localization for video captioning. Yao et al. [2015] also makes an improvement to the S2VT system [Venugopalan et al. 2015a] by using 3D spatial-temporal CNN to capture both local and global temporal structures in a video with the temporal attention mechanism using the soft alignment method [Bahdanau et al. 2014] borrowed from neural machine translation. Yu et al. [2015c] refines the CNN-LSTM system [Venugopalan et al. 2015a] by introducing hierarchical

[2]https://vsubhashini.github.io/s2vt.html.

recurrent neural networks (hierarchical RNN) to enable the system to describe a long video with multiple sentences or a paragraph, since the sentence decoder with the hierarchical RNN will be able to capture long-term dependencies between sentences.

Another interesting direction in video captioning attempts to perform movie captioning by aligning movie frames with sentences and paragraphs [Zhu et al. 2015a] or book chapters [Tapaswi et al. 2015] from the novelization of the movie. Zhu et al. [2015a] also relies on deep learning (specifically context-aware CNN and LSTM), but this system goes beyond words by having a sentence embedding representation from skip-thought vectors [Kiros et al. 2015]. Tapaswi et al. [2015] aligns book chapters with movies, which is simpler and can be done with finding shortest path on graphs, similar to gene alignments.

Some datasets for video captioning are the Microsoft Research Video Description Corpus (MSVD) [Chen and Dolan 2011], the TACoS Multi-Level corpus [Rohrbach et al. 2014], YouCook [Das et al. 2013], Montreal Video Annotation Dataset (M-VAD) [Torabi et al. 2015], and the MPII Movie Description dataset [Rohrbach et al. 2015]. Ferraro et al. [2015] provides some explanations and experiments for recent video captioning datasets.

2.3. Visual Question Answering

Visual Question Answering (vQA) [Antol et al. 2015; Yu et al. 2015a] is a rich new task in which senses and knowledge from the textual question must be considered in the visual extraction process to generate the correct answer. vQA is similar to image captioning in that the output is textual descriptions of images. However, it is a much more difficult task to answer queries well based on given visual information, because it is impossible to simply use generic descriptions to avoid true, deep image understanding. First, the classical "to know what is where by looking" [Marr 1982] is still applied in the sense that the answer must correctly recognize objects and background, along with the spatial relations between them. Second, the contextual information needs to be able to answer 'when, for what, and how?" [Parikh 2009] to make sure that this information is relevant. For example, a scene can visually differ from summer to winter, and in this this context, a shovel should be used to clean out snow piles in front of the house. The main challenge in vQA lies in the design of meaning representation and system mechanisms that extract the information and perform a reasoning process on them.

The *meaning representation* depends on the nature of the data. For a static single image, object and scene relations will often suffice, and the system needs to generate sentence descriptions [Antol et al. 2015]. But if the data are an image sequence or video, the system needs to model the temporal dimension of what is going on in the image sequence, whether it is an activity, causal event, or any other thing that changes over time. Tapaswi et al. [2016] proposes a question-answering dataset similar to vQA [Antol et al. 2015]. The proposed system uses a memory network [Sukhbaatar et al. 2015] to model three-way relationships of story, question, and answer for the question answering task.

Another thing to be considered is the *system mechanism* that will extract relevant information. Visual madlibs [Yu et al. 2015a] is another visual Question Answering dataset that requires a system to answer fill-in-the-blank/multiple-choice questions. Visual madlibs has a nice categorization of visual questions. This helps the system select a mechanism, whether it will be saliency detection to answer the image's interesting questions (type 3) or attribute prediction to answer object's attribute questions (type 6). As suggested by the dataset creators, a simple baseline should be joint-embeddings or deep-learning methods. Visual7w [Zhu et al. 2016] is also another visual Question Answering dataset that has a categorization of visual questions. Visual7w categorizes the questions into what, where, when, who, why, how many, and which (7w) in a

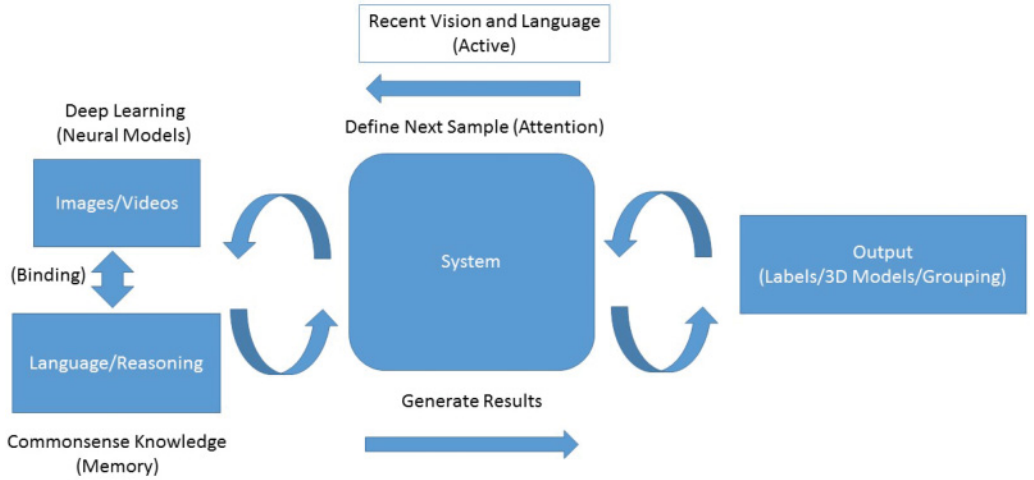

Fig. 7. The modern vision and language framework which actively defines the next sample and exhibits cognitive phenomena like attention and memory.

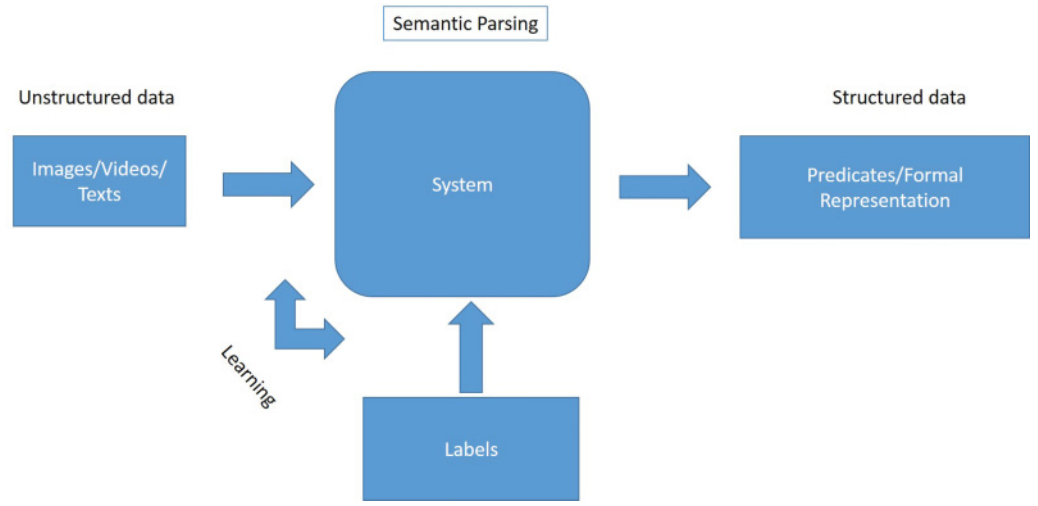

Fig. 8. The semantic parsing framework pipeline. Taking both unstructured data and labels as its input, the system learns the mapping to structured formal representation (predicates).

multiple-choice setting. Visual7w is a subset of Visual Genome, which is a larger amount of data that tries to map images to WordNet with a very rich context setting including captioning, question answering, attributes, relations, region descriptions and object instances in the scale of millions for each of the contexts.

Recent methods (Figure 7) are able to begin to tackle vQA because they decompose the structures of both visual and language modalities in a cognitive-inspired way. The systems are designed in an *active vision* framework where the system actively defines the next sample to be processed. This attention mechanism defines how the output can provide feedback reaching low-level features through the system. Reinforcement learning techniques like REINFORCE [Williams 1988] can make this happen naturally by using an exploration-exploitation framework as the attention mechanism [Yeung et al.

2016]. The agent architecture has memory and can be trained with the standard back-propagation algorithm. In addition, the system needs to deal with the joint modeling of vision and language as a symbol-signal binding problem. In other words, the system needs to perform an unambiguous mapping between those two modalities.

2.4. Visual Retrieval

Content-based Image Retrieval (CBIR) is another field in multimedia that widely utilizes language in the form of query strings or concepts. ACM CSUR already has a decent survey on this topic [Li et al. 2016] so we will merely briefly visit this topic for the completeness in our article.

Traditionally, images are indexed by their low-level vision features like color, shape, and texture [Zhang et al. 2012]. Recent CBIR systems try to annotate an image region with a word, similarly to semantic segmentation, so the keyword tags are close to human interpretation. CBIR systems use keywords to describe an image for image retrieval but visual attributes describe an image for image understanding. Nevertheless, visual attributes provide a suitable middle layer for CBIR with an adaptation to the target domain. This difference stems from the fact that CBIR is less noisy compared to other application domains like robotics.

Some CBIR approaches use co-occurrence models on words in image grids [Mori et al. 1999], machine translation on image blobs to words [Duygulu et al. 2002], probabilistic models on blobs to words [Barnard et al. 2003], topic models on blobs and words [Blei and Jordan 2003], the Cross-media Relevance Model on the joint distribution of image blobs and words [Jeon et al. 2003], and the Continuous-space Relevance Model, which further models the semantics rather than color, texture, or shape features [Lavrenko et al. 2003]. Since image class labels and annotated tags are likely to have some relations to each other, Wang et al. [2009] proposes an extension of supervised Latent Dirichlet Allocation [Mcauliffe and Blei 2008] to jointly model the latent spaces of image classification and image annotation. Gong et al. [2014] uses Canonical Correlation Analysis (CCA) and its kernel extension to perform mapping on data from three different spaces of visual features, tags, and visual semantics to the joint subspace and perform multimodal retrieval. Deep learning and word embedding can help provide representations for the images and their tags as an intermediate feature transformation to CCA and can improve the accuracy [Murthy et al. 2015].

Some standard data sets for CBIR are Corel5k [Feng et al. 2004], NUS-WIDE [Chua et al. 2009], MIRFlickr08 [Huiskes and Lew 2008], PASCAL-VOC [Everingham et al. 2010], ESP Game [Von Ahn and Dabbish 2004], UIUC [Li and Fei-Fei 2007], Labelme [Russell et al. 2008], and IAPR TC-12 [Makadia et al. 2008]. For recent trends in CBIR, Zhang et al. [2012], Tousch et al. [2012], and Wang et al. [2012] provide a summary and comparison of recent methods and future challenges. Another direction in CBIR deals with video retrieval by events [Han et al. 2015; Singh et al. 2015]. The task is closely related to activity recognition, because of the temporal dimension in a video.

3. LANGUAGE AND VISION FOR ROBOTICS

3.1. Symbol Grounding

Robotics is a research field involving both *perception* and *manipulation* of the world [Thrun et al. 2005]. There are many modalities of perception: vision, sound, smell, taste, taction or balance, and so on. By mean of hardware sensors, each of these modalities provides sensory information to be processed. To manipulate the environment, a robot controls its body parts and applies physical forces, embodied from its mechanics, to perform actions that may result in its own movements or a change to its environment.

In this section, we will consider research works in language and vision where the resulting applications are in robotics.

For robotics, language becomes symbols, and vision dominates perception. Bridging language and vision is equivalent to the *symbol grounding problem* [Harnad 1990] in the lens of robotics that tries to foster an autonomous agent to reason and react in the real world. The symbol grounding problem is about grounding the meaning of the symbols to the perception of the real world. Grounding the meaning only in symbols can be done for some cases (such as Compositional semantic or machine reading [Etzioni et al. 2006]) but may result in closed reference loops without some connection to sensing as a context to be grounded in. For example, words like *yellow*, *board*, *fast*, *prawn*, or *walk* need real-world perceptions in order to understand their meaning. Another example [Mooney 2013] in Wordnet [Miller 1995] is the word pair *sleep* and *asleep*, which have pointers to each other as a loop.

The symbol grounding problem can be categorized into five subcategories [Coradeschi et al. 2013]. First, *physical symbol grounding* deals with grounding symbols into perception. This conforms to the original definition in Harnad [1990]. Second, *perceptual anchoring* [Coradeschi and Saffiotti 2000] connects the sensor data from an object to a higher-order symbol that refers to that object while also maintaining the connection in time. Third, *grounding words in action* [Roy 2005] maintains a hierarchical representation of concepts for abstract words: higher-order concepts are grounded into basic concepts and sensorimotor grounded actions. Fourth, *social symbol grounding* [Cangelosi 2006] tries to share the connection after anchoring for one agent to many agents. This can involve pragmatics as well as lexical semantics. Fifth, *grounding symbols in the semantic web* [Johnston et al. 2008] tries to ground the symbols into a large ontological knowledge base from the internet. This grounding in text is the least restricted setting of the symbol grounding problem. The symbol grounding problem represents a gap between symbols and perception.

3.2. Robotics Vision

Visual data can enable perception in a cognitive system that fulfills the grounding process. Current computer vision techniques, however, are limited when only low-level information, like pixels, are used [Aloimonos and Fermüller 2015]. Humans process perceptual inputs by using their knowledge about things they perceive in all modalities in the form of words, phrases, and sentences [Mooney 2008]. The language may be the knowledge about objects, scenes, actions, or events in the real world in which these perceptions can be given by Computer Vision systems. This knowledge needs relations to make sense of and understand meaning.

This gives rises to a *top-down active visual process* [Aloimonos and Fermüller 2015] where language requests some actions from a sensorimotor system that will instantiate a new data collection from perceptual sensors. For example, if an object is far away, a language executive may request an action that ends in a clearer viewpoint that has a better pose and scale of that object. The action may be decomposable into subtasks and may need a planning system for a robot to perform the action. The visual information from the new viewpoint may be collectible only from direct observation, not by inference. This scheme is interactive between the conventional bottom-up and this novel top-down vision system. This is like active learning in education [Meyers and Jones 1993] where a student can actively ask a teacher for a clearer understanding of the subject but a student also needs to be smart so the answer from his question will provide the desired information or lead to an interesting discourse between the two.

Robotics Vision (RV) [Darrell 2010] differs from the traditional CV. Nowadays, a large part of CV relies on machine learning where the performance relies on the volume of the data. Recognition in CV focuses on category-level recognition that aims to be general

and can be acquired from the internet to create a big-data paradigm. In contrast, RV uses reliable hardware sensors like depth camera [Lai et al. 2011] or motion camera [Barranco et al. 2014] so the performance has typically relied on the number and quality of sensors instead (although this field, too, is beginning to incorporate more deep-learning, big-data approaches: in robotic vehicles, for example). Also, recognition in RV focuses on situated-level recognition where the context environment is limited from the embodiment of the sensor hardware. Robotics Vision tasks relate to how a robot can perform sequences of actions on affordable objects to manipulate the real-world environment. Such tasks need some information involving detecting and recognizing objects, object motion tracking, human activity recognition, and so on. This is to give a robot both static and dynamic information about its surrounding contexts.

Interfacing CV and RV needs domain adaptation techniques [Daumé III 2007] since online images and real-world objects differ. Lai and Fox [2010] tries to incorporate data available from the Internet, namely Google's 3D Warehouse, to solve 3D point cloud object detection in the real world in which the problem describes the need for domain adaptation and the methods involve domain adaptation formulations.

3.3. Situated Language in Robotics

For robotics, languages are used to describe the physical world for a robot to understand its environment. This problem is another form of the symbol grounding problem known as *grounded language acquisition* or embodied language problem. A robot should be able to perceive and transform the information from its contextual perception into language using semantic structures. By this, a language can be grounded into another predefined grammar that can represent meaning.

The most well-known approach to represent meaning is *Semantic Parsing (SP)* [Zelle and Mooney 1996], which transforms words into logic predicates. SP tries to map a natural language sentence to a corresponding meaning representation that can be a logical form like λ-calculus using Combinatorial Categorical Grammar (CCG) [Steedman 1996] as rules to compositionally construct a parse tree. Like many other parsing problems, there can be many possible parse trees, in which some of them may be incorrect. They may contain an incorrect lambda form in the tree, or the meaning may be incorrect. Therefore, a prediction model can be employed to solve this problem as another structured prediction problem [Daumé III et al. 2009; Taskar et al. 2005; Bakir 2007]. SP transforms texts to logical forms and enables a dialog agent to perform symbol grounding to a knowledge base [Thomason et al. 2015] in which the system can make a query about the environment via logical forms.

Another alternative to SP is Abstract Meaning Representation (AMR) [Banarescu et al. 2012]. AMR tries to map a sentence to a graph representation that encodes the meaning of that sentence about "who is doing what to whom." By this, AMR makes morphology and syntax abstract into predicate-argument structures. A sentence becomes a rooted directed acyclic graph and is labeled on edges for relations and on leaves on concepts.

Not all words can describe the geometric properties of the world. For example, words like *special* or *perhaps* provide emotional information but not about the real-world context. For robotics, most words should involve either *actions* or contexts. Actions can be words like *move*, *put*, *push*, or *pick up* that need an argument that can be the robot itself or other things in the vicinity for execution. The contexts can be words or phrases like *a table on the right*, *a green button on the wall*, or *the red car under the roof* that show *affordance* [Chemero 2003] of an object so it can be specified for the robot to perform actions with attention. Prepositions from these phrases, like *on* in *on the wall* or *under* in *under the roof*, encode *spatial relations* that are essential contexts to specify where to perform actions for a robot [Zampogiannis et al. 2015]. Affordance

is a property of an object related to tools and actions, for example, a cucumber can be cut with a knife or a green button can be pushed with a hand. This defines an affordance of *able to cut* to a knife and *can be cut* to a cucumber. The property that defines affordance can be shape, color, or texture, which can be perceived as low-level vision features [Myers et al. 2015]. Affordance helps reasoning by giving the relational information about objects, tools, and actions along with their pre-conditions and post-conditions after the action is applied. Adverbs depict the details of an action that can be *force* or *speed*. A robot should know how to perform a successful action by adjusting its own control parameters to be precise with tools and object affordances. For example, cutting a *cucumber* and a *log* need different tools, like a *knife* and an *axe*, in which a robot should handle them differently and apply a specific amount of forces for each tool. Moreover, to perform a *rescue* task and a *finding object* task need different speed where *rescue* should be executed with full speed while *find* can be done with a normal pace. In addition to vision and language, recent advanced tactile sensors [Yogeswaran et al. 2015] should help in perceiving and adjusting forces by sensing forces and frictions directly.

3.4. Recent Works in Language and Vision for Robotics

Robotics has various interesting applications, and we will describe only a salient set of them. Robotics tasks involving language and vision can be categorized into three main tasks: a robot talking to human, a robot that learns from human actions, and a robot that performs navigation.

First, a robot can *interact with human via language* that is situated and grounded in perception. This needs both language understanding and generation as well as some representation that will integrate perception into language. For situated language generation tasks, Chen and Mooney [2008] apply semantic parsing to ground simulated Robocup soccer events into language for commentary. This work goes beyond a manual template system to learning to perform. Automated sport game models (ASPOGAMO) [Beetz et al. 2007] try to track sportsmen and ball positions via detection and tracking systems on broadcasted football games. ASPOGAMO can handle changing lighting conditions, fast camera motions, and distant players. Unifying both systems for tracking and sportcasting is another promising direction. For situated language understanding tasks, Matuszek et al. [2013] parses user natural language instructions into a formal representation that commands robots with Probabilistic CCG for semantic parsing [Zettlemoyer and Collins 2005]. A robot will follow the parsed instructions and execute its control for routing. Matuszek* et al. [2012] further incorporates visual attributes for grounding words describing objects based on perception. The model incorporates both semantic parsing for language and visual attribute classification for vision and is trained via EM algorithm that jointly learns language and attribute relations.

To unify generation and understanding, Grounded Situation Model (GSM) [Mavridis and Roy 2006] is a situated conversation agent that bridges perceptions, language, and actions with semantic representation based on parsing. Its belief is updated with a mixture of visual, language, and proprioceptive data. GSM can answer questions and perform basic actions via verbal interaction. Tellex et al. [2014] categorizes robot language tasks into following instructions, asking questions, and requesting help. A robot will try to find uncertain parts in the command and ask a targeted question for clarification, and then it will perform better actions based on the obtained information. To achieve this, the G^3 framework [Tellex et al. 2011], which is a probabilistic model, is used to model the inverse semantics from the uncertain part of the world to a word in a sentence. This is not an obvious mapping since an expression can refer to many things in a scene. This problem is called the "situated referential grounding" problem because there is a mismatch between the perceptions of a human and a robot. One

way to overcome the problem is to have a robot generate a collaborative dialogue [Liu et al. 2014a; Liu and Chai 2015; Fang et al. 2013] so mismatched perceptions will be dissolved when more information, such as object properties and spatial information, can be obtained in a subjective way.

Walter et al. [2015] unifies language and vision for robotics again by bridging visual, language, speech, and control data for a forklift robot. Their robot can recognize objects based on one example using one-shot visual memory. Its natural language interface works by speech processing or pen gestures. It is equipped with reliable sensors and an anytime motion planner that enables its local actions without global information. It has annunciation and visualization interfaces. The robot also has a safety mechanism for nearby workers using pedestrian detection and shout detection. For further information in this topic, Mavridis [2015] provides a survey for verbal and nonverbal human-robot interaction.

Second, a robot can *learn to perform actions* by imitating or observing human actions. This setting is sometimes denoted as robot learning from demonstration (LfD), imitation learning [Piaget 2013] or observational learning [Bandura 1974]. Instead of manual hard coding, a robot can learn from either a human teacher or other robots. LfD helps program robotic controls to perform actions. The motivation is from the *mirror-neuron system* [Rizzolatti and Craighero 2004], which is a neuron that will fire when an animal both performs and observes a certain action. This phenomenon enables humans to see and learn other people's actions, including understanding intentions and emotions attached in those actions.

LfD [Argall et al. 2009] tries to learn a mapping of state and action pairs from a teacher's demonstration (s_t, a_t) as a supervised learning setting so the learned policy from state S to action A, which is $\pi : S \rightarrow A$, will have some performance guarantees. The mapping between actions is defined by $T(s'|s, a) : S \times A \times S \rightarrow [0, 1]$. Moreover, the states may not be fully observable, so the observed state Z is from another mapping $S \rightarrow Z$, and a policy will be $\pi : Z \rightarrow A$ instead. For more information, Argall et al. [2009] provides a unified framework for LfD.

Julian Jaynes's bicameral mind [Jaynes 2000] theorizes an existence of language intervention in human consciousness and motivates an incorporation of language and vision to LfD. The "bicameral mind" refers to a mind with two chambers in which one is speaking as an executive in language and the other just obeys and perform actions. Jaynes theorizes, rather speculatively, that such a mental organization existed in humans that lived in earlier cultures but not in modern people. The theory is probably unprovable in humans, but it seems useful as an organizational principle for a robot with language that can see, talk, and perform actions. Language and vision can provide data and models for this scenario. A robot would recognize actions by watching a demonstration from either a real-world teacher or a video. Then it would ground its perceptions to language and learn to plan its own actions. Finally, the planned imitated actions would be carried out by its motor actuators. A real robotics system implementation of this concept can be found in Marge et al. [2016] using a Wizard-of-Oz (WOz) method. WOz describes a method where a human acts in place of the robot's language executive, and the robot just executes actions. It is used for gathering a corpus of data about human-robot interaction to help develop future systems that will be able to understand natural language in a particular situated context.

Action grammar [Guerra-Filho and Aloimonos 2007; Karapurkar 2008; Pastra and Aloimonos 2012] is the most common interface for language and vision in LfD. Action grammar is a Probabilistic Context Free Grammar that models human actions in parallel with natural language descriptions. Human pose is aligned with phonemes as the lowest level description. Human movement that makes changes to human pose to create actions is analogous with morphemes. Human activity, composed from some specific

human actions, is paralleled with syntax or grammar. Karapurkar [2008] provides an excellent overview of the action grammar framework.

An action grammar can model the hand-object-tool relations and incrementally construct an activity tree [Summers-Stay et al. 2012]. An interesting feature is that interleaving activities can be recognized using an action grammar. This provides a strength from the nature of the sequential data in which there can be many motif sequences to be recognized. Teo et al. [2012] introduces a prior knowledge of action-tool relations mined from their cooccurence in the Gigawords corpus. This results in a learned prior knowledge that reflects the real world and helps improve activity recognition using textual knowledge. Yang et al. [2013] further models the hand-object-tool, object-scene, and object-scene-attribute relations as a multi-label classification task with a prior from Gigawords corpus as in Teo et al. [2012]. The results suggest that the language prior will rule out some relations that will never occur because they do not make sense such as using a cup of water to fold a t-shirt. Furthermore, many actions can be easily observed from real-world descriptions, so having the meaning from texts helps a robot learn to connect words to actions.

Nowadays, there are many websites where demonstration videos intended to teach humans are available. A robot can potentially watch these videos and imitate actions. A robot may need to be given all language, speech, and visual data for a complete understanding [Malmaud et al. 2015]. However, using only language and vision can efficiently teach a robot to perform a task successfully as in cooking [Yang et al. 2015b] or t-shirt folding [Shukla et al. 2015]. Semantic Parsing is an enhancement to the action grammar in which a post-condition can be inferred using the semantics of the Combinatorial Categorical Grammar itself [Yang et al. 2015]. For example, cutting a cucumber will result in divided cucumbers. This is also called "manipulating action consequences" [Yang et al. 2013], which represents object-centric consequences as a knowledge graph. The results can be considered as the desired goals of the action performers. The corresponding semantics is $\lambda.x\lambda.y\, cut(x, y) \rightarrow divided(y)$, where x is a cutting tool, such as a knife, and y is a cuttable object, such as a cucumber. Shukla et al. [2015] goes further and introduces the knowledge transfer scheme, where a robot will learn from a human demonstration and then transfer its knowledge by teaching to another human.

Third, a robot can perform *planning for navigation*. With an understanding of the surrounding world, a robot can perform reasoning and make a plan to achieve its goals. However, real-world navigation requires map making with some techniques like Simultaneous Localization And Mapping (SLAM) [Durrant-Whyte and Bailey 2006]. For robotics research, a simulated world is created to test the software so a navigation is on a visual world instead. The map can be either egocentric or top view (aerial). The spatial relations are frequently used to refer to a subjective relation between a robot and a map. Therefore, the language used will involve more pragmatics, since many meanings are hidden as a presupposition or an implicature (such as the left-right-straight-backward directions). There are also many scenarios where navigation plans can distinctively differ. The scenario can range from an event or a place where an action should be immediately carried, like an emergency event with a dense human crowd, to a mysterious and dangerous place but with a few people, like a mine, to a household kitchen, which is a safe place, but the objects are cluttered.

The planning for navigation problem can be cast as a situated language generation [Garoufi 2014]. The task is to generate instructions in virtual environments that guide both directions and actions of an agent [Byron et al. 2007]. The system cares about the agent embodiment so the generated language becomes listener-centric. For example, the system will generate "push the second button on the wall" to help the agent accomplish the final task, which is taking the trophy instead of just giving a high

level instruction like "take the trophy." The contexts in a virtual environment are converted into natural language using a Tree-Adjoining Grammar, which is then further converted into a planning problem [Koller and Stone 2007; Garoufi and Koller 2010]. Some visual cues, such as gaze direction [Garoufi et al. 2016] can help the system in generating a more meaningful discourse because it will have some feedback information that helps in inferring the mental state of the listener while giving instructions.

4. DISTRIBUTIONAL SEMANTICS IN LANGUAGE AND VISION

4.1. Distributional Semantics

Distributional Semantics [Harris 1954] relies on the hypothesis that words that occur in similar contexts are similar in meaning. This hypothesis can recover word meaning from co-occurrence statistics between words and contexts in which they appear. Distributional Semantic Models (DSMs) use the vector space and its properties to model meaning. A semantic vector space represents a word as a data point and encodes the similarity and relatedness between words in term of measurements between those data points. To obtain such a semantic vector space, various methods have been proposed. Those methods mostly belong to the class of *latent variable models* where the vector space is spanned by the latent variable vectors.

Word similarity can be measured by Euclidean distance or cosine similarity of the angle between word vectors. Similar words will have similar vector representations and will be near to each other in the vector space. Word relatedness can be observed from the displacements or offsets between the vectors that represent relations. For instance, the word *king* can be mapped to another vector that is very similar to the word *queen* by subtracting the word vector for *man* and adding the word vector for *woman*.

Modeling meaning by word co-occurence as the only source of information limits its connection to the real world. One may argue that meaning can be described in language and provide understanding without any perception as a prerequisite for communication. So, word co-occurence can be derived and provide meaning because an appropriate corpus can be created to serve any specific meaning. However, one can also argue that a lot of knowledge cannot be understood without a grounded perception.

A coconut can occur with other words that describe some related aspects of a *coconut* like *fruits* or *tropical countries*. Nevertheless, one cannot understand its unique shape, color, and texture without perceiving a real coconut. If there are any similar fruits whose shape, color or texture are the same as a coconut, then one can effectively describe a coconut by each aspect, but there will be still a question of "What does it really look like?" Therefore, perception is essential in order to answer these questions and provide more information sources for modeling meaning in addition to only words-to-words relations.

An object's context can be the spatial location where this specific object appears. This encapsulates the natural information about how scenes and objects can relate to each other [Choi et al. 2012]. The context also includes the meaning of the real world that similar objects in the same category will be likely to appear in the same context, from which it can be further inferred that those objects co-occur with each other in some specific patterns as well. For example, a *mountain lion* and a *deer* are likely to be in a *forest* or a *zoo*. If we were to observe them both in a *forest*, then the *mountain lion* is likely to be chasing the *deer* for its meal. In addition, the context can also be the intrinsic properties of an object. In this case, DSMs can model low-level visual features such as colors, shapes, or textures and describe an object using those contexts. For example, a *tennis ball* tends to be yellow, round, and furry. Understanding these meanings will help the reasoning process about the real world.

4.2. Distributed Word Representation

There are two types of connectionist representations, local and distributed [Plate 1997; Hinton 1984]. Local representations have a small number of features for each item while distributed representations have a large number of features from nearly all features in the feature pool. Local representations are sparse and capture salient signals from particular features. In contrast, distributed representations are dense and continuous and depict patterns over many signals. Local representations have lower representation power than distributed representations that are dense and compact.

4.2.1. Latent Semantic Analysis and Topic Models. Latent Semantic Analysis (LSA) [Dumais 2007] or Latent Semantic Indexing (LSI) is the most well-known instance of distributed word representation that tries to recover word relations from a corpus. LSA uses Singular Value Decomposition (SVD) on the term-document matrix "*C*" which outputs a low-rank approximation that can be used as a weighting scheme. Probabilistic Latent Semantic Indexing (pLSI) [Hofmann 1999] is the probabilistic version of LSA that models each word in a document as a sample from a mixture model of conditionally independent multinomial distributions. Each document consists of topics, and each topic consists of words. pLSI has an improvement over LSA in terms of the interpretability of word-topic and topic-document relations. Latent Dirichlet Allocation (LDA) [Blei et al. 2003] was proposed to overcome the overfitting problem in pLSI by introducing the Dirichlet prior over the multinomial topic distribution.

4.2.2. word2vec. While LSA takes a geometric approach to extracting meaning from co-occurance statistics, and LDA uses a probabilistic generative approach to the problem [Bruni et al. 2014], the recently proposed word2vec [Mikolov et al. 2013] uses a neural network to extract the semantic vectors. word2vec produces a word embedding, in other words, a distributed representation [Turian et al. 2010] that is equivalent to factorizing the term-term matrix [Levy and Goldberg 2014b]. Word embeddings are typically induced by neural language models [Bengio et al. 2003; Morin and Bengio 2005] that predict the context given an input word.

Because it can be trained on such large corpora, word2vec can provide a significant improvement over LSA and LDA from its quality of the output word embedding. The resulting representation is encoded with word meaning so similar words will have similar vector representations, for example, the *vector('car')* will be similar to the *vector('driver')*. Moreover, the relationship between words is also preserved in term of displacement between points such that basic vector operations on these points will be meaningful, for example, *vector('Paris') - vector('France') + vector('Italy')* will result in a vector very similar to the *vector('Rome')*. Also, the displacement can capture syntactic relations, for instance, *vector('sweet') - vector('sweetest') + vector('fastest')* will result in a vector very similar to the *vector('fast')*.

4.2.3. State-of-the-art word2vec Models. There are a lot of interesting follow-up works that try to improve word2vec in various ways. For example, GloVe [Pennington et al. 2014] incorporates global word co-occurence information. Dependency-Based Word Embeddings [Levy and Goldberg 2014a] include syntactic dependency information. Marginal Contrast Embedding [Chen et al. 2015b] uses a pairwise ranking loss function in the learning-to-rank framework to model antonyms on an antonym word list. Symmetric Pattern Based Word Embeddings [Schwartz et al. 2015] creates its word embedding based on symmetric patterns extracted from corpora such as "X such as Y" or "X is a Y" or "X and Y" that also enables antonym modeling. Modeling lexical contrast [Mohammad et al. 2013] is an important contribution because it solves the fundamental problem of co-occurrence between the target word and its synonym or antonym, which previously was a blind spot of modeling distributional semantics by word co-occurrence.

Other methods improve word embeddings with respect to interpretability, since typical models output dense vectors that cannot be easily interpreted by humans. Faruqui et al. [2015] and Yogatama et al. [2014] add interpretability into a word embedding using sparse coding methods. AutoExtend [Rothe and Schütze 2015] can extend existing word embedding from synsets and lexemes to make a better embedding. Retrofitting [Faruqui et al. 2014] is a graph-based method that also extends an existing word embedding. word2vec triggered a lot of interest with its high-quality word representation output, so it has created a renaissance of word embedding research.

Still other works generalize word embedding to logic for better reasoning. Traditionally, this was done by the method of binding roles and fillers by using operations such as tensor product [Clark and Pulman 2007]. Roles are logical predicates and fillers are logical atoms that are both represented as vectors. Recently, there are efforts that try to map Boolean structures to distributional semantics for recognizing textual entailment, which decides the entailment between two sentences. The proposed approaches are Markov Logic Networks [Beltagy et al. 2015; Garrette et al. 2014] and learning a mapping function with Boolean Distributional Semantic Model [Kruszewski et al. 2015].

4.3. Bag-of-Visual-Words and Image Embedding

4.3.1. Bag-of-Visual Words. In CV, the idea of bag-of-words representation (BoW) has long been borrowed from the NLP community in solving recognition tasks under the name of bag-of-visual-words representation (BoVW). BoW representation discards spatial and grammatical relations between words and creates a representation of a document based on only word frequencies that outputs the term-document matrix. Similarly, BoVW discards location and shape information from an image, saving only a large number of local clues whose statistical properties can be analyzed.

BoVW representation is a descriptor-image matrix where descriptors are local features in the visual codebook. Theses descriptors are salient keypoints of an image extracted using techniques such as Scale-invariant feature transform (SIFT) [Lowe 2004] or speeded up robust features (SURF) [Bay et al. 2006] that can reliably find descriptors across images under difficulties like rotation, translation, or illumination changes. Then, the descriptors are clustered together by a clustering algorithm such as k-means [MacQueen et al. 1967] to create a visual codebook. There are varying numbers of visual descriptors in each image unlike text documents whose words come off the shelf, so the codebook step is needed for visual data. Thus, the clustering step is needed in order to make the frequency histogram comparable across images by fixing the codewords. This BoVW model does not go beyond point descriptors to edges or surfaces. The local descriptors tend to find image patches with similar appearance, but these may not be correlated with the object-level parts in an image [Grauman and Leibe 2010].

Some landmark works incorporate location and shape information into the BoVW model and achieve a test-of-time popularity like Spatial Pyramid Matching [Lazebnik et al. 2006] or the Constellation model [Fei-Fei et al. 2007].

4.3.2. Image Embedding. Natural images lie in a low-dimensional manifold in the space of all possible images. The efforts to model that manifold result in image embedding techniques [Pless and Souvenir 2009]. Image embedding is similar to the word vector embedding representation because it also results in a dense low-dimensional feature vector. Besides this, images that are close to each other in the embedding space are similar and each dimension captures factors of variations in the images such as pose, illumination, or translations.

The image embedding is an output or intermediate output from a representation learning methods such as dimensionality reduction methods [van der Maaten et al. 2009] including deep-learning techniques [Bengio et al. 2013]. One of the most

dominant examples for image embedding is face recognition [Zhao et al. 2003]. Eigenfaces [Turk and Pentland 1991] uses Principal Component Analysis (PCA) to project to a low-dimensional image manifold, which represents faces along with common variations. (PCA is related to SVD used in LSA.) Fisherfaces [Belhumeur et al. 1997] also uses a dimensionality reduction method, namely Fisher's Linear Discriminant Analysis [Fisher 1936], to compute a discriminative projection that is better for classification. Other well-known dimensionality reduction techniques based on SVD, for instance, ISOMAP [Tenenbaum et al. 2000], Locally Linear Embedding [Roweis and Saul 2000], or Laplacian Eigenmaps [Belkin and Niyogi 2003], can capture nonlinear variations in images such as pose, age, or facial expressions along some manifold directions. This looks similar to the analogy property in word2vec that can perform reasoning based on vector space displacements. These image embedding techniques fall into the category of Spectral and Tensor Methods [Anandkumar et al. 2012a, 2012b, 2014]. They all follow a common three-step procedure: Build a statistical matrix or tensor, recover latent variable vectors via eigendecomposition or tensor decomposition, and perform dimensionality reduction with the learned projection vectors to obtain the final manifold vector space. This area is still an active research area for machine learning.

4.3.3. Deep Learning: State-of-the-Art Image Embedding Models. Both BoVW and image embedding are used as a feature set for classification mainly for recognition tasks but are not limited to them. For example, RNNs was applied to another recognition task of semantic segmentation in the context of scene parsing [Socher et al. 2011]. Recently, image embedding from deep CNNs [LeCun et al. 1998] that exhibits similar characteristics to word2vec [Garcia-Gasulla et al. 2015] is applied in various tasks in reorganization (like Optical Flow [Dosovitskiy et al. 2015], segmentation [Long et al. 2015]), and Recognition (like visual analogy using Siamese CNNs [Sadeghi et al. 2015]).

There are many attempts in many benchmarks in open competitions to design a better architecture of CNNs. We are not going to describe all of them in depth since this deserves a survey of its own. Some notable architectures are AlexNet [Krizhevsky et al. 2012], GoogLeNet [Szegedy et al. 2015], VGG net [Simonyan and Zisserman 2014], and ResNet [He et al. 2016]. The main insight from these models is that deeper models are better for classification. Another insight is that a convolutional layer behaves like a receptive field while fully connected layers give better discriminative power. Based on these models for recognition, more models are proposed for other computer vision tasks. For example, R-CNN [Girshick et al. 2014] or Fast R-CNN [Girshick 2015] have been proposed for object detection. Another widely used architecture is FCN [Long et al. 2015] for semantic segmentation. It is a fully convolutional neural networks that can perform pixelwise labeling. The idea delves further into a deeper problem of structured prediction when recurrent neural networks can be seen as a generic sequence model like CRFs [Zheng et al. 2015].

In short, one can conclude that performing representation learning on image data may result in an image embedding with similar properties to word2vec. However, image embedding is likely to be more domain specific and has more data set bias [Torralba et al. 2011]. Even though it is trained on a data set of millions or billions images like AlexNet [Krizhevsky et al. 2012] and provides a breakthrough in recognition on the ImageNet LSVRC-2010 challenge [Berg et al. 2010], the coverage of real-world objects is just around 1,000 categories and is still far from learning from text alone like training word2vec on Google's Billion Words corpus, which has 793,471 vocabulary words [Chelba et al. 2014]. For face recognition, the recently proposed DeepFace [Taigman et al. 2014] model was created to recognize around 4,000 identities on the Labelled Face in the Wild data set [Huang et al. 2007], which is very remarkable but still far from a system that can recognize anybody's face on the fly.

Solving this problem by learning on a larger-scale image collection without a bounded set of categories, such as learning from the Internet [Chen et al. 2013], to provide a general image embedding that models infrequent data like word2vec is a promising future direction. It is a major challenge to make such a system ready for in-the-wild autonomous systems, however.

4.4. Multimodal Distributional Semantics Models

4.4.1. Early Multimodal Distributional Semantics Models. DSMs have been applied to jointly model semantics based on both visual features like colors, shape or texture and textual features like words. The common pipeline is to map visual data to words and apply distributional semantics models like LSA or topic models on top of them. Visual attributes can approximate the linguistic features for a distributional semantics model. Silberer et al. [2013] uses various DSMs, such as CCA or the Attribute-topic Model [Andrews et al. 2009], to model visual attributes annotated based on Nelson's association norms [Nelson et al. 2004] and McRae's feature norms [McRae et al. 2005]. Visual attributes are predicted from images using SVMs following the traditional attribute prediction pipeline. BoVW can become input features for DSMs. Bruni et al. [2012] empirically compare SIFT and LAB visual features (BoVW), DSMs based on word co-occurance counts, and a joint feature of both. Roller and Im Walde [2013] create BoVW based on SURF [Bay et al. 2008] and GIST [Oliva and Torralba 2001] descriptors. They propose an extended version of the attribute-topic model to incorporate word co-occurance features, cognitive features, and visual features. Finally, DSMs can be applied directly to interesting channels of visual signals. McMahan and Stone [2015] released a Lexicon of Uncertain Color Standards that uses topic models to find an association between 829 English word meanings and colors in HSV space. The word meaning is a probability distribution over the color spaces. Joulin et al. [2015] trains a convolutional networks on 100 million Flickr photos and captions in a weakly supervised setting and found that the weights in the multi-label loss layer can be used as word embedding similar to word2vec.

The outcome from the aforementioned works emphasizes that visual space and textual space represent different meaning spaces which complement and are at least partially orthogonal to each other. This conclusion is supported by the brain activity evidence from fMRI images [Anderson et al. 2013] and automatic evaluation metrics [Leong and Mihalcea 2011]. In a technical perspective, this insight says that a better joint semantic subspace can be created and yield a better performance for some tasks [Bruni et al. 2012]. Bruni et al. [2014] and Baroni [2016] are recent excellent review articles about multimodal DSMs that describe these methods.

4.4.2. Recent Neural Multimodal Distributional Semantics Models. Recent waves of deep-learning works have received a lot of attention based on the undeniable power and performance of deep neural networks. Neural models have surpassed many traditional methods in both vision and language by learning better distributed representation from the data. Srivastava and Salakhutdinov [2014] show that Multimodal Deep Boltzmann Machines can model joint visual and textual features better than topic models. In addition, neural models can model some cognitively plausible phenomena such as attention and memory.

The cognitive science community has established many well-studied theories of *visual attention and memory*, from which neural networks and computer vision often get their inspiration. Attention is a mechanism that selects a subset of available information channels for enhanced processing. Carrasco [2011] categorizes visual attention into three main types: spatial attention, feature-based attention, and object-based attention. Spatial attention corresponds to eye movement. Feature-based

attention corresponds to a specific set of properties of an object such as colors, orientations, or motions regardless of the object location in a scene. Object-based attention is guided by object structure. Memory is the record of experiences presented in the brain [Eichenbaum 2008]. Memory can be characterized to long-term and short-term memory [Cowan 2008]. A short-term memory can hold a small amount of temporal information in a very accessible state. Working memory is not separate from short-term memory. Working memory is used for planning, decision making, and behavior. Short-term memory and working memory are highly related to attention process [Fougnie 2008]. Long-term memory can hold a large amount of records about facts, knowledge, and prior events. Procedural memory is a long-term memory that holds for tasks without conscious awareness like daily routines such as walking without thinking [Cohen and Bacdayan 1994].

Several recent works have shown promising results of how visual attention and memory can be efficiently simulated as a computer process. For attention, an image can initially give an image embedding representation using CNNs and RNNs. An LSTM network is placed on top and acts like a state machine that simultaneously generates outputs, such as image captions [Xu et al. 2015a], and attentively looks at relevant regions of interest in an image one at a time [Karpathy and Fei-Fei 2015b]. For memory, there are many recent works that try to incorporate commonsense knowledge into visual question answering [Andreas et al. 2016a, 2016b; Kumar et al. 2016; Xiong et al. 2016], which has shown to outperform multimodal embedding methods like those from Gao et al. [2015], Ren et al. [2015], and Malinowski et al. [2015]. A multi-step attention mechanism is needed to utilize the process of getting relevant information based on the structure of the input question (semantic structure prediction). The building blocks for memory networks are memory units like LSTMs or Gate Recurrent Units [Cho et al. 2014] to contain commonsense knowledge in the form of vectors. These memory networks have many different function modules that will act as a step in attention mechanisms. Those functions in the state-of-the-art at the time of writing this article [Andreas et al. 2016a, 2016b] are "find," "transform," "combine," "describe," and "measure," which will have inputs and outputs of "image," "attention" (simply speaking, a feature map), and "label."

5. DISCUSSION: SURVEY WRAP-UP, THE CURRENT STATE OF THE FIELD AND LIMITATIONS OF THE STATE-OF-THE-ART

In this survey, we discuss two artificial intelligence tasks, computer vision, and natural language processing, and applications based on their integration in multimedia and robotics. In multimedia, natural language processing can provide high-level meaning to assist low-level computer vision processes in large corpora as observed in visual attributes, image captioning, video captioning, and image annotation. In robotics, natural language processing can help a robot to perform a more accurate reasoning and control given the situated data stream from low-level computer vision processes. Next we emphasize the unified theme of distributional semantics as the concept that is able to perform model integration for computer vision and natural language processing. More specifically, we reviewed both traditional approaches such as bag-of-words models and topic models as well as recent approaches like word2vec and deep learning. Finally, we summarize all approaches into the framework of distributional semantics and provide a lookahead into a cognitive-inspired recent framework.

The current state of the field exhibits a rapid progress that can be summarized in three words: *accuracy*, *scalability*, and *creativity*. The accuracy is achieved from the widely recognized success of deep learning. The scalability is provided by the advancements in GPUs acceleration from high-performance computing. The creativity can be observed from a lot of novel applications such as captioning, question answering,

and dialog systems. There has been substantial progress in previous research problems, and the field has moved on to the next step, continuously pushing the limitations.

Compared to traditional approaches in computer vision, deep learning with CNNs provide the best result in accuracy, and the system can be trained end-to-end, from input to output. Still, there remain challenges at the corners of the system capabilities. For example, the *number of categories* that a system can categorize is still far from achieving human-level vision. The system may need an out-of-sample extension to project the knowledge and decide that it should either make a decision (know) or request the data to learn (do not know). Moreover, to recognize an *accurate structured knowledge* in the setting of fine-grained classification is still far from commonsense knowledge. For example, to recognize a bird, a human has a lot of contextual knowledge that comes up when he sees a bird. Saying that a bird can fly, he needs to also verify that the bird is flying or has a potential to fly (recognizing that the bird does not have a broken wing or is still alive are crucial). A bird is a stimulus to perception and his mind thinks about what it has seen in the context of prior knowledge. The current framework cannot efficiently do that. At least a new model design is needed in order to perform a *scalable structured prediction*. Compared to traditional approaches in natural language processing, deep learning with RNNs, word embedding models like word2vec, and memory models like LSTM provide the best result in accuracy for sequence to sequence learning. However, sometimes LSTM fails when the structure of the input data is not known beforehand. In such cases, an extension like TreeLSTM that handles tree-structured data [Tai et al. 2015] or even a feed-forward neural networks with random weights [Hochreiter and Schmidhuber 1997] may have a better result. This raises a new point to be taken care of, especially for a system with attention mechanism.

From a robotics perspective, robots are programmed for a fixed and situated task like cutting or pouring. The knowledge in those fixed tasks are well defined in syntax and semantics. However, there is still a limitation in *modeling human users*. To know what and when and where an action is "relevent" requires a deep understanding for a clear-cut decision. For instance, a survey robot that receives commands needs adaptation and self-inference. The commander may have no knowledge about the scene, and a robot needs to reason on its own about how to proceed and make the commander happy. A robot should be able to recall events (episodic memory) from its training session. This is a critical issue to achieve satisfactory human-robot interaction. We believe that borrowing some frameworks from the database and data-mining community such as *recommendation systems* is a promising direction. Situated understanding at the level of a robot that can successfully perform a task is also needed for multimedia analysis. For example, visual retrieval will output more relevant images or videos given a more fine-grained knowledge about user requirements. From a multimedia perspective, the system can index a gigantic amount of data with a few or no training samples. If a robot need to perform a new task, then it is more favorable to *learn quickly from a short demonstration* (a few minutes) rather than hundreds of videos (hours). Learning from demonstration currently look at the whole video. Recent active vision systems with attention mechanism can generate key viewpoints from a single short demonstration video. The systems also need embodied semantic memory, episodic memory and accurate binding to complete the attention mechanism.

Human bodies and robot bodies differ. We want to be able to teach a robot the best control for its body by allowing it to observe human poses, but at the same time we hope that the robot will learn a more efficient technique or a new routine to perform that action based on its own form and capabilities. A robot should have some prior knowledge of itself in its own memory. This should define a different attention mechanism and a binding from human poses. A robot needs the ability to represent one thing in many

ways so it can approach problems creatively. When a robot learns, its memory should grow and can learn new things easily. This zero-shot learning is also needed for a general artificial intelligence robots that can perform various tasks on the fly with a versatile working memory. The next dream is that a robot should have emotional states [Minsky 2006]. It seems impossible to understand humans without an ability to feel in the same way. People can think of different things in different emotional states. They may have cognitive biases related to emotional states, and a robot should know the nature of the emotional transition. We wish a robot to know this and understand humans better.

No system can perform well without a solid knowledge in methodology (a great set of tools). *Tensor decomposition* is an interesting future direction for modeling multimodal structured data. If we want interesting statistics, then we can decompose a tensor into a set of orthogonal bases in the similar spirit to SVD. This may be a good starting point for the next generation of meaning representation for multimodal distributional semantics. A tensor can encode many different things. We can make a tensor of word co-occurrences for topic models. We may create a tensor of linked data to perform graph segmentation. On top of those, we may combine multiple tensors and create a new tensor with more dimensions to model word co-occurrences segmented in graphs as a simultaneous community detection and topic tracking system. For deep learning, a tensor decomposition can also be used to train deep networks [Denton et al. 2014] or even compress them in a compress sensing setting [Novikov et al. 2015]. A theoretical analysis on deep learning and tensor decomposition is described in Cohen et al. [2016].

It is undeniable that a long-lived dream of every artificial intelligence researcher is to simulate a mind with a computer. Despite progress in neural networks, many human mental behaviors and capabilities are far beyond what any such system has been able to demonstrate. At the extreme this means incremental learning and big visual and language data for a "baby AI" [Bengio et al. 2007] that learns like a child and will grow up and behave like a human. A baby AI needs to learn on larger-scale image collection with an open set of categories, such as learning from the Internet [Chen et al. 2013], in order to provide a general visual representation that models infrequent data like word2vec. Also, a baby AI would need to have a capability of machine reading where the system can learn new concepts, including new words, new syntaxes, new meanings, and alternative interpretations in a never-ending setting [Carlson et al. 2010]. From both viewpoints, a simple conclusion is to solve the "out-of-sample/vocab" problem in the umbrella of the "generalization" problem in machine learning. An inspiration from theories of human memory is an interesting starting point.

6. LOOKAHEAD: POSSIBILITIES FOR THE FUTURE

The future of the field 'language and vision" looks very promising, since governments and industry are intensifying funding in this arena as they look for solutions to particular problems. Thus, the potential for increased activity is very strong. What form is this research going to take? Certainly, research will continue along similar lines with the state of the art today addressing the variety of problems described in this article. However, this kind of research is not only geared towards applications but also has a very basic scientific component, as it is related to the problem of the grounding of meaning in language. Thus we feel that the field will soon regroup to re-examine basic questions and techniques. Towards that goal we offer two insights.

First, since the integration of vision and language is a fundamentally cognitive problem, we would like to take inspiration from the cognitive sciences. In this field, there has been recent ongoing strong activity on the problem of "event understanding." This is of course for a good reason, because events are how people experience what happens to them, and a lot of our actions and behaviors are guided by our understanding of

events. We perceive events, we participate in events with our actions, we simulate events we hear or read about, and we solve problems using our knowledge of events [Radvansky and Zacks 2014]. This year the Annual Meeting of the Cognitive Science Society (CogSci 2016)[3] is devoted to the problem of event understanding. Events correspond to stories, but events consist of actions, and actions correspond to sentences. In turn, actions consist of actors (nouns, agents, hands), objects and tools (nouns and adjectives), movements and actions and their manner (verbs, adverbs) and goals and changes. Thus, there is good reason to expect that a compositional theory of event understanding is possible. Aspects of that theory will determine the specific research questions to be asked. It is worth noting at this point that the most prominent psychological theory of event representations (TEC-Theory of Event Coding) [Hommel et al. 2001] suggests that perceptual contents and action plans are coded in a common representational medium by feature codes with distal reference. Perceived events (perceptions) and to-be-produced events (actions) are equally represented by integrated, task-tuned networks of feature codes—cognitive structures that they call event codes. Thus events are used both in perception and planning.

The second insight comes from the field of education. What does it really mean for a system to successfully integrate vision and language? What does it really mean for a system to have an understanding of vision and language? Or, to turn the question around, what does it mean for a human to understand, or how do we know that a human has understanding of something? The answer to this question is visible in our educational system: To assess whether someone has an understanding of something, we give them a test. We ask them questions. All these tests are based on the idea that comprehension can be tested by asking the subject to perform certain tasks. If the subject performs those tasks successfully, then comprehension is said to have happened. Thus, comprehension is a set of testable skills. One popular educational theory devoted to this is Bloom's Taxonomy [Bloom et al. 1956; Summers-Stay 2013]. The taxonomy was an effort to categorize skills that could be taught in schools and to make more clear the meaning of text (language) comprehension. At the lowest level of this hierarchy was *knowledge*, because it was a necessary precursor to any other kind of cognition. It amounts to recognizing and retrieving relevant knowledge from memory. Next in the hierarchy comes *comprehension*, that is, to construct meaning through summarizing, inferring, comparing, and explaining. Next comes *application*, that is, the carrying out of a procedure. It is followed by *analysis* (breaking material into parts, determining how the parts relate to one another) by *synthesis* (planning, producing) and *evaluation* (judgments and comparisons). Clearly, by following this viewpoint, educators can design questions at progressive levels of difficulty, as far as language comprehension is concerned.

We suggest that the same framework can be applied to the problem of understanding vision and language. Of course children do not need to be taught image understanding because by the time they go to school they "know" how to see. But we could follow this framework to teach computers how to integrate vision and language. This is shown in Figure 1, which also provides our view for future research on the topic. To assess the understanding of vision and language by a system, we need to ask the system a set of questions about images and videos and evaluate the answers. Indeed, at the *knowledge* level, the system should be able to answer questions related to *what* is *where*. The system will need capabilities to recognize objects and tools, human and their parts, and specific actions and label events. At the next level, of *comprehension*, the system should be able to answer questions related to *why* and *what's next*. In other words, the system will need capabilities to recognize intentions of the event participants as

[3]http://cognitivesciencesociety.org/conference2016/index.html.

well as to predict what comes next in the sequence. Next comes application and here the questions have to do with *how*. Given knowledge of an event, can an agent (robot) actually perform the same task (learning from imitation). In the *analysis*, we ask questions about the components of events and how they relate to each other, while in the *synthesis* we ask how to compose new events from our knowledge of existing events. Finally, in evaluation, we ask questions about the utility of events. We believe this is a fruitful framework in charting future research.

ACKNOWLEDGMENTS

The authors thank the anonymous reviewers and editors for their insightful comments and suggestions. We thank Dr. Hal Daumé III and Dr. Cynthia Matuszek for very useful advice and discussions. We thank Dr. Marco Baroni for providing a useful knowledge source. We also thank Aleksandrs Ecins, Austin Myers, Gregory Kramida, Bharat Singh, Xintong Han, Ang Li, Yi Zhang, Yezhou Yang, Ching Lik Teo, and other members of UMD Computer Vision Lab (CVL) and UMD Computational Linguistics and Information Processing Lab (CLIP) for useful insights and support. We also thank Dr. Desmond Elliott, Dr. Kevin Lai, and Dr. Joseph Tighe for a useful informal discussion.

REFERENCES

Somak Aditya, Yezhou Yang, Chitta Baral, Cornelia Fermuller, and Yiannis Aloimonos. 2015. From images to sentences through scene description graphs using commonsense reasoning and knowledge. *arXiv preprint arXiv:1511.03292* (2015).

Eren Erdal Aksoy, Alexey Abramov, Johannes Dörr, Kejun Ning, Babette Dellen, and Florentin Wörgötter. 2011. Learning the semantics of object–action relations by observation. *Int. J. Robot. Res.* (2011), 0278364911410459.

Yiannis Aloimonos and Cornelia Fermüller. 2015. The cognitive dialogue: A new model for vision implementing common sense reasoning. *Image Vis. Comput.* 34 (2015), 42–44.

Animashree Anandkumar, Rong Ge, Daniel Hsu, Sham M. Kakade, and Matus Telgarsky. 2014. Tensor decompositions for learning latent variable models. *J. Mach. Learn. Res.* 15, 1 (2014), 2773–2832.

Animashree Anandkumar, Daniel Hsu, and Sham M. Kakade. 2012a. A method of moments for mixture models and hidden Markov models. In *COLT*, Vol. 1. 4.

Anima Anandkumar, Yi-kai Liu, Daniel J. Hsu, Dean P. Foster, and Sham M. Kakade. 2012b. A spectral algorithm for latent Dirichlet allocation. In *Advances in Neural Information Processing Systems*. 917–925.

Andrew J. Anderson, Elia Bruni, Ulisse Bordignon, Massimo Poesio, and Marco Baroni. 2013. Of words, eyes and brains: Correlating image-based distributional semantic models with neural representations of concepts. In *EMNLP*. 1960–1970.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016a. Learning to compose neural networks for question answering. In *NAACL*.

Jacob Andreas, Marcus Rohrbach, Trevor Darrell, and Dan Klein. 2016b. Neural module networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 39–48.

Mark Andrews, Gabriella Vigliocco, and David Vinson. 2009. Integrating experiential and distributional data to learn semantic representations. *Psychol. Rev.* 116, 3 (2009), 463.

Stanislaw Antol, Aishwarya Agrawal, Jiasen Lu, Margaret Mitchell, Dhruv Batra, C. Lawrence Zitnick, and Devi Parikh. 2015. Vqa: Visual question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2425–2433.

Brenna D. Argall, Sonia Chernova, Manuela Veloso, and Brett Browning. 2009. A survey of robot learning from demonstration. *Robot. Auton. Syst.* 57, 5 (2009), 469–483.

Dzmitry Bahdanau, Kyunghyun Cho, and Yoshua Bengio. 2014. Neural machine translation by jointly learning to align and translate. In *ICLR 2015*.

Collin F. Baker, Charles J. Fillmore, and John B. Lowe. 1998. The Berkeley framenet project. In *Proceedings of the 36th Annual Meeting of the Association for Computational Linguistics and 17th International Conference on Computational Linguistics–Volume 1*. Association for Computational Linguistics, 86–90.

Gökhan Bakir. 2007. Predicting structured data. MIT press, 2007.

Laura Banarescu, Claire Bonial, Shu Cai, Madalina Georgescu, Kira Griffitt, Ulf Hermjakob, Kevin Knight, Philipp Koehn, Martha Palmer, and Nathan Schneider. 2012. Abstract meaning representation (AMR)

1.0 specification. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. ACL. 1533–1544.

Albert Bandura. 1974. *Psychological Modeling: Conflicting Theories*. Transaction Publishers.

Andrei Barbu, Alexander Bridge, Zachary Burchill, Dan Coroian, Sven Dickinson, Sanja Fidler, Aaron Michaux, Sam Mussman, Siddharth Narayanaswamy, Dhaval Salvi, and others. 2012a. Video in sentences out. In *UAI 2012*.

Andrei Barbu, Aaron Michaux, Siddharth Narayanaswamy, and Jeffrey Mark Siskind. 2012b. Simultaneous object detection, tracking, and event recognition. In *ACS 2012*.

Kobus Barnard, Pinar Duygulu, David Forsyth, Nando De Freitas, David M. Blei, and Michael I. Jordan. 2003. Matching words and pictures. *J. Mach. Learn. Res.* 3 (2003), 1107–1135.

Kobus Barnard and David Forsyth. 2001. Learning the semantics of words and pictures. In *Proceedings of the 8th IEEE International Conference on Computer Vision, 2001 (ICCV 2001)*, Vol. 2. IEEE, 408–415.

Marco Baroni. 2016. Grounding distributional semantics in the visual world. *Lang. Ling. Compass* 10, 1 (2016), 3–13.

Francisco Barranco, Cornelia Fermüller, and Yiannis Aloimonos. 2014. Contour motion estimation for asynchronous event-driven cameras. *Proc. IEEE* 102, 10 (2014), 1537–1556.

Daniel Barrett, Andrei Barbu, N. Siddharth, and Jeffrey Siskind. 2016. Saying what you're looking for: Linguistics meets video search. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 38, 10 (Oct. 2016).

Jonathan Barron and Jitendra Malik. 2015. Shape, illumination, and reflectance from shading. *IEEE Trans. Pattern Anal. Mach. Intell.* 37, 8 (2015), 1670–1687.

Herbert Bay, Andreas Ess, Tinne Tuytelaars, and Luc Van Gool. 2008. Speeded-up robust features (SURF). *Comput. Vis. Image Understand.* 110, 3 (2008), 346–359.

Herbert Bay, Tinne Tuytelaars, and Luc Van Gool. 2006. Surf: Speeded up robust features. In *Computer Vision–ECCV 2006*. Springer, 404–417.

Michael Beetz, Suat Gedikli, Jan Bandouch, Bernhard Kirchlechner, Nico von Hoyningen-Huene, and Alexander Perzylo. 2007. Visually tracking football games based on TV broadcasts. In *Proceedings of the 20th International Joint Conference on Artificial Intelligence (IJCAI)*.

Peter N. Belhumeur, João P. Hespanha, and David J. Kriegman. 1997. Eigenfaces vs. fisherfaces: Recognition using class specific linear projection. *IEEE Trans. Pattern Anal. Mach. Intell.* 19, 7 (1997), 711–720.

Mikhail Belkin and Partha Niyogi. 2003. Laplacian eigenmaps for dimensionality reduction and data representation. *Neur. Comput.* 15, 6 (2003), 1373–1396.

Islam Beltagy, Stephen Roller, Pengxiang Cheng, Katrin Erk, and Raymond J. Mooney. 2015. Representing meaning with a combination of logical form and vectors. *arXiv preprint arXiv:1505.06816* (2015).

Yoshua Bengio, Aaron Courville, and Pierre Vincent. 2013. Representation learning: A review and new perspectives. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 8 (2013), 1798–1828.

Yoshua Bengio, Réjean Ducharme, Pascal Vincent, and Christian Janvin. 2003. A neural probabilistic language model. *J. Mach. Learn. Res.* 3 (2003), 1137–1155.

Yoshua Bengio, Hugo Larochelle, Pascal Lamblin, Dan Popovici, Aaron Courville, Clarence Simard, Jerome Louradour, and Dumitru Erhan. 2007. Deep architectures for baby AI. (2007).

A. Berg, J. Deng, and L. Fei-Fei. 2010. Large scale visual recognition challenge (ILSVRC), 2010. Retrieved from http://www. image-net.org/challenges/LSVRC (2010).

Tamara Berg and Alexander C. Berg. 2009. Finding iconic images. In *Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, 2009 (CVPR Workshops 2009)*. IEEE, 1–8.

Tamara L. Berg, Alexander C. Berg, Jaety Edwards, Michael Maire, Ryan White, Yee-Whye Teh, Erik Learned-Miller, and David A. Forsyth. 2004. Names and faces in the news. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR'04)*, Vol. 2. IEEE, II–848.

Tamara L. Berg, Alexander C. Berg, and Jonathan Shih. 2010. Automatic attribute discovery and characterization from noisy web data. In *Computer Vision–ECCV 2010*. Springer, 663–676.

Tamara L. Berg, David Forsyth, and others. 2006. Animals on the web. In *Proceedings of the 2006 IEEE Computer Society Conference on Computer Vision and Pattern Recognition*, Vol. 2. IEEE, 1463–1470.

Raffaella Bernardi, Ruket Cakici, Desmond Elliott, Aykut Erdem, Erkut Erdem, Nazli Ikizler-Cinbis, Frank Keller, Adrian Muscat, and Barbara Plank. 2016. Automatic description generation from images: A survey of models, datasets, and evaluation measures. *J. Artif. Intell. Res.* 55 (2016), 409–442.

David M. Blei and Michael I. Jordan. 2003. Modeling annotated data. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Informaion Retrieval*. ACM, 127–134.

David M. Blei, Andrew Y. Ng, and Michael I. Jordan. 2003. Latent Dirichlet allocation. *J. Mach. Learn. Res.* 3 (2003), 993–1022.

Benjamin S. Bloom and others. 1956. Taxonomy of educational objectives. Vol. 1: Cognitive domain. McKay, New York, NY (1956), 20–24.

Alexander M. Bronstein, Michael M. Bronstein, and Ron Kimmel. 2005. Three-dimensional face recognition. *Int. J. Comput. Vis.* 64, 1 (2005), 5–30.

Elia Bruni, Gemma Boleda, Marco Baroni, and Nam-Khanh Tran. 2012. Distributional semantics in technicolor. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 136–145.

Elia Bruni, Nam-Khanh Tran, and Marco Baroni. 2014. Multimodal distributional semantics. *J. Artif. Intell. Res.* 49 (2014), 1–47.

Donna Byron, Alexander Koller, Jon Oberlander, Laura Stoia, and Kristina Striegnitz. 2007. Generating instructions in virtual environments (GIVE): A challenge and an evaluation testbed for NLG. (2007).

Angelo Cangelosi. 2006. The grounding and sharing of symbols. *Pragm. Cogn.* 14, 2 (2006), 275–285.

Andrew Carlson, Justin Betteridge, Bryan Kisiel, Burr Settles, Estevam R. Hruschka Jr, and Tom M. Mitchell. 2010. Toward an architecture for never-ending language learning. In *AAAI*, Vol. 5. 3.

Marisa Carrasco. 2011. Visual attention: The past 25 years. *Vis. Res.* 51, 13 (2011), 1484–1525.

Joao Carreira and Cristian Sminchisescu. 2010. Constrained parametric min-cuts for automatic object segmentation. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3241–3248.

Angel X. Chang, Manolis Savva, and Christopher D. Manning. 2014. Semantic parsing for text to 3d scene generation. *ACL 2014* (2014), 17.

Yu-Wei Chao, Zhan Wang, Yugeng He, Jiaxuan Wang, and Jia Deng. 2015. HICO: A benchmark for recognizing human-object interactions in images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1017–1025.

Ciprian Chelba, Tomas Mikolov, Mike Schuster, Qi Ge, Thorsten Brants, Phillipp Koehn, and Tony Robinson. 2014. One billion word benchmark for measuring progress in statistical language modeling. In *Proceedings of the 15th Annual Conference of the International Speech Communication Association*.

Anthony Chemero. 2003. An outline of a theory of affordances. *Ecological Psychology* 15, 2 (2003), 181–195.

David L. Chen and William B. Dolan. 2011. Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the 49th Annual Meeting of the Association for Computational Linguistics: Human Language Technologies—Volume 1*. Association for Computational Linguistics, 190–200.

David L. Chen and Raymond J. Mooney. 2008. Learning to sportscast: A test of grounded language acquisition. In *Proceedings of the 25th International Conference on Machine Learning*. ACM, 128–135.

Xinlei Chen, Hao Fang, Tsung-Yi Lin, Ramakrishna Vedantam, Saurabh Gupta, Piotr Dollar, and C. Lawrence Zitnick. 2015a. Microsoft COCO captions: Data collection and evaluation server. *arXiv preprint arXiv:1504.00325* (2015).

Xinlei Chen, Ashish Shrivastava, and Arpan Gupta. 2013. Neil: Extracting visual knowledge from web data. In *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 1409–1416.

Zhigang Chen, Wei Lin, Qian Chen, Xiaoping Chen, Si Wei, Hui Jiang, and Xiaodan Zhu. 2015b. Revisiting word embedding for contrasting meaning. In *Proceedings of ACL*.

Kyunghyun Cho, Bart van Merriënboer, Dzmitry Bahdanau, and Yoshua Bengio. 2014. On the properties of neural machine translation: Encoder–decoder approaches. *Syntax Sem. Struct. Stat. Transl.* (2014), 103.

Myung Jin Choi, Antonio Torralba, and Alan S. Willsky. 2012. Context models and out-of-context objects. *Pattern Recogn. Lett.* 33, 7 (2012), 853–862.

Tat-Seng Chua, Jinhui Tang, Richang Hong, Haojie Li, Zhiping Luo, and Yantao Zheng. 2009. NUS-WIDE: A real-world web image database from national university of Singapore. In *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 48.

Stephen Clark and Stephen Pulman. 2007. Combining symbolic and distributional models of meaning. In *AAAI Spring Symposium: Quantum Interaction*. 52–55.

Michael D. Cohen and Paul Bacdayan. 1994. Organizational routines are stored as procedural memory: Evidence from a laboratory study. *Organiz. Sci.* 5, 4 (1994), 554–568.

Nadav Cohen, Or Sharir, and Amnon Shashua. 2016. On the expressive power of deep learning: A tensor analysis. In *Proceedings of the 29th Annual Conference on Learning Theory*. 698–728.

Silvia Coradeschi, Amy Loutfi, and Britta Wrede. 2013. A short review of symbol grounding in robotic and intelligent systems. *KI-Künstliche Intell.* 27, 2 (2013), 129–136.

Silvia Coradeschi and Alessandro Saffiotti. 2000. Anchoring symbols to sensor data: Preliminary report. In *AAAI/IAAI*. 129–135.

Nelson Cowan. 2008. What are the differences between long-term, short-term, and working memory? *Progr. Brain Res.* 169 (2008), 323–338.

Trevor Darrell. 2010. Learning Representations for Real-world Recognition. Retrieved from http://www.eecs.berkeley.edu/~trevor/colloq.pdf UCB EECS Colloquium [Accessed: 2015 11 1].

Pradipto Das, Chenliang Xu, Richard Doell, and Jason Corso. 2013. A thousand frames in just a few words: Lingual description of videos through latent topics and sparse object stitching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2634–2641.

Hal Daumé III. 2007. Frustratingly easy domain adaptation. *ACL 2007* (2007), 256.

Hal Daumé III, John Langford, and Daniel Marcu. 2009. Search-based structured prediction. *Mach. Learn.* 75, 3 (2009), 297–325.

Emily L. Denton, Wojciech Zaremba, Joan Bruna, Yann LeCun, and Rob Fergus. 2014. Exploiting linear structure within convolutional networks for efficient evaluation. In *Advances in Neural Information Processing Systems*. 1269–1277.

Jesse Dodge, Amit Goyal, Xufeng Han, Alyssa Mensch, Margaret Mitchell, Karl Stratos, Kota Yamaguchi, Yejin Choi, Hal Daumé III, Alexander C. Berg, and others. 2012. Detecting visual text. In *Proceedings of the 2012 Conference of the North American Chapter of the Association for Computational Linguistics: Human Language Technologies*. Association for Computational Linguistics, 762–772.

Jeffrey Donahue, Lisa Anne Hendricks, Sergio Guadarrama, Marcus Rohrbach, Subhashini Venugopalan, Kate Saenko, and Trevor Darrell. 2015. Long-term recurrent convolutional networks for visual recognition and description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 2625–2634.

Alexey Dosovitskiy, Philipp Fischery, Eddy Ilg, Caner Hazirbas, Vladimir Golkov, Patrick van der Smagt, Daniel Cremers, Thomas Brox, and others. 2015. Flownet: Learning optical flow with convolutional networks. In *Proceedings of the 2015 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2758–2766.

Susan T. Dumais. 2007. LSA and information retrieval: Getting back to basics. *Handb. Latent Semant. Anal.* (2007), 293–321.

Hugh Durrant-Whyte and Tim Bailey. 2006. Simultaneous localization and mapping: Part I. *IEEE Robot. Autom. Mag.* 13, 2 (2006), 99–110.

Pinar Duygulu, Kobus Barnard, Joao F. G. de Freitas, and David A. Forsyth. 2002. Object recognition as machine translation: Learning a lexicon for a fixed image vocabulary. In *Computer Vision ECCV 2002*. Springer, 97–112.

Aleksandrs Ecins, Cornelia Fermuller, and Yiannis Aloimonos. 2014. Shadow free segmentation in still images using local density measure. In *Proceedings of the 2014 IEEE International Conference on Computational Photography (ICCP)*. IEEE, 1–8.

Aleksandrs Ecins, Cornelia Fermuller, and Yiannis Aloimonos. 2016. Cluttered scene segmentation using the symmetry constraint. In *Proceedings of the International Conference in Robotics and Automation (ICRA)*.

H. Eichenbaum. 2008. Memory. *Scholarpedia* 3, 3 (2008), 1747.

Desmond Elliott and Frank Keller. 2013. Image description using visual dependency representations. In *EMNLP*. 1292–1302.

Desmond Elliott and Frank Keller. 2014. Comparing automatic evaluation measures for image description. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics: Short Papers*, Vol. 452. 457.

Oren Etzioni, Michele Banko, and Michael J. Cafarella. 2006. Machine reading. In *AAAI*, Vol. 6. 1517–1519.

Mark Everingham, Luc Van Gool, Christopher K. I. Williams, John Winn, and Andrew Zisserman. 2010. The pascal visual object classes (voc) challenge. *Int. J. Comput. Vis.* 88, 2 (2010), 303–338.

Rui Fang, Changsong Liu, Lanbo She, and Joyce Y. Chai. 2013. Towards situated dialogue: Revisiting referring expression generation. In *EMNLP*. 392–402.

Ali Farhadi. 2011. Designing Representational Architectures in Recognition. University of Illinois at Urbana-Champaign. Champaign, IL, USA.

Ali Farhadi, Ian Endres, and Derek Hoiem. 2010. Attribute-centric recognition for cross-category generalization. In *Proceedings of the 2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2352–2359.

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

Ali Farhadi, Mohsen Hejrati, Mohammad Amin Sadeghi, Peter Young, Cyrus Rashtchian, Julia Hockenmaier, and David Forsyth. 2010. Every picture tells a story: Generating sentences from images. In *Computer Vision–ECCV 2010*. Springer, 15–29.

Manaal Faruqui, Jesse Dodge, Sujay K. Jauhar, Chris Dyer, Eduard Hovy, and Noah A. Smith. 2014. Retrofitting word vectors to semantic lexicons. *arXiv preprint arXiv:1411.4166* (2014).

Manaal Faruqui, Yulia Tsvetkov, Dani Yogatama, Chris Dyer, and Noah Smith. 2015. Sparse overcomplete word vector representations. *arXiv preprint arXiv:1506.02004* (2015).

Li Fei-Fei, Rob Fergus, and Pietro Perona. 2007. Learning generative visual models from few training examples: An incremental bayesian approach tested on 101 object categories. *Comput. Vis. Image Underst.* 106, 1 (2007), 59–70.

S. L. Feng, Raghavan Manmatha, and Victor Lavrenko. 2004. Multiple Bernoulli relevance models for image and video annotation. In *Proceedings of the 2004 IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 2004 (CVPR'04).*, Vol. 2. IEEE, II–1002.

Francis Ferraro, Nasrin Mostafazadeh, Ting-Hao Huang, Lucy Vanderwende, Jacob Devlin, Michel Galley, and Margaret Mitchell. 2015. A survey of current datasets for vision and language research. In *Proceedings of the 2015 Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, Lisbon, Portugal, 207–213.

Ronald A. Fisher. 1936. The use of multiple measurements in taxonomic problems. *Ann. Eugen.* 7, 2 (1936), 179–188.

Daryl Fougnie. 2008. The relationship between attention and working memory. *New Res. Short-term Mem.* (2008), 1–45.

D. F. Fouhey, A. Gupta, and A. Zisserman. 2016. 3D shape attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Jianlong Fu, Jinqiao Wang, Xin-Jing Wang, Yong Rui, and Hanqing Lu. 2015. What visual attributes characterize an object class? In *Computer Vision–ACCV 2014*. Springer, 243–259.

Haoyuan Gao, Junhua Mao, Jie Zhou, Zhiheng Huang, Lei Wang, and Wei Xu. 2015. Are you talking to a machine? Dataset and methods for multilingual image question. In *Advances in Neural Information Processing Systems*. 2296–2304.

D. Garcia-Gasulla, J. Béjar, U. Cortés, E. Ayguadé, and J. Labarta. 2015. Extracting visual patterns from deep learning representations. *arXiv preprint arXiv:1507.08818* (2015).

Peter Gärdenfors. 2014. *The Geometry of Meaning: Semantics Based on Conceptual Spaces*. MIT Press.

Konstantina Garoufi. 2014. Planning-based models of natural language generation. *Lang. Ling. Compass* 8, 1 (2014), 1–10.

Konstantina Garoufi and Alexander Koller. 2010. Automated planning for situated natural language generation. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 1573–1582.

Konstantina Garoufi, Maria Staudte, Alexander Koller, and Matthew W. Crocker. 2016. Exploiting listener gaze to improve situated communication in dynamic virtual environments. *Cognitive Science* 40, 7 (2016), 1671–1703.

Dan Garrette, Katrin Erk, and Raymond Mooney. 2014. A formal approach to linking logical form and vector-space lexical semantics. In *Computing Meaning*. Springer, 27–48.

Ross Girshick. 2015. Fast r-cnn. In *Proceedings of the IEEE International Conference on Computer Vision*. 1440–1448.

Ross Girshick, Jeff Donahue, Trevor Darrell, and Jitendra Malik. 2014. Rich feature hierarchies for accurate object detection and semantic segmentation. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 580–587.

Yunchao Gong, Qifa Ke, Michael Isard, and Svetlana Lazebnik. 2014. A multi-view embedding space for modeling internet images, tags, and their semantics. *Int. J. Comput. Vis.* 106, 2 (2014), 210–233.

Kristen Grauman and Bastian Leibe. 2010. *Visual Object Recognition*. Number 11. Morgan & Claypool Publishers.

Douglas Greenlee. 1978. Semiotic and significs. *Int. Stud. Philos.* 10 (1978), 251–254.

Sergio Guadarrama, Niveda Krishnamoorthy, Girish Malkarnenkar, Sarad Venugopalan, Randy Mooney, Trevor Darrell, and Kate Saenko. 2013. Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shot recognition. In *Proceedings of the 2013 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 2712–2719.

Gutemberg Guerra-Filho and Yiannis Aloimonos. 2007. A language for human action. *Computer* 40, 5 (2007), 42–51.

Abhinav Gupta. 2009. Beyond nouns and verbs. (2009).

Saurabh Gupta, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2015. Indoor scene understanding with RGB-D images: Bottom-up segmentation, object detection and semantic segmentation. *Int. J. Comput. Vis.* 112, 2 (2015), 133–149.

Saurabh Gupta and Jitendra Malik. 2015. Visual semantic role labeling. *arXiv preprint arXiv:1505.04474* (2015).

Xintong Han, Bharat Singh, Vlad I. Morariu, and Larry S. Davis. 2015. Fast automatic video retrieval using web images. *arXiv preprint arXiv:1512.03384* (2015).

Emily M. Hand and Rama Chellappa. 2016. Attributes for improved attributes: A multi-task network for attribute classification. *arXiv preprint arXiv:1604.07360* (2016).

Bharath Hariharan, Pablo Arbeláez, Ross Girshick, and Jitendra Malik. 2014. Simultaneous detection and segmentation. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Stevan Harnad. 1990. The symbol grounding problem. *Physica D* 42, 1 (1990), 335–346.

Zellig S. Harris. 1954. Distributional structure. *Word* 10, 2–3 (1954), 146–162.

Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. 2016. Deep residual learning for image recognition. In *CVPR*.

Geremy Heitz and Daphne Koller. 2008. Learning spatial context: Using stuff to find things. In *Computer Vision–ECCV 2008*. Springer, 30–43.

Geoffrey E. Hinton. 1984. Distributed representations. Technical Report: Carnegie Melon University.

Sepp Hochreiter and Jürgen Schmidhuber. 1997. Long short-term memory. *Neur. Comput.* 9, 8 (1997), 1735–1780.

Micah Hodosh, Peter Young, and Julia Hockenmaier. 2013. Framing image description as a ranking task: Data, models and evaluation metrics. *J. Artif. Intell. Res.* (2013), 853–899.

Thomas Hofmann. 1999. Probabilistic latent semantic indexing. In *Proceedings of the 22nd Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 50–57.

Bernhard Hommel, Jochen Müsseler, Gisa Aschersleben, and Wolfgang Prinzb. 2001. The theory of event coding (TEC): A framework for perception and action planning. *Behav. Brain Sci.* 24 (2001), 849–937.

Thanarat Horprasert, David Harwood, and Larry S. Davis. 1999. A statistical approach for real-time robust background subtraction and shadow detection. In *IEEE ICCV*, Vol. 99. 1–19.

Gary B. Huang, Manu Ramesh, Tamara Berg, and Erik Learned-Miller. 2007. *Labeled Faces in the Wild: A Database for Studying Face Recognition in Unconstrained Environments*. Technical Report. Technical Report 07-49, University of Massachusetts, Amherst.

Mark J. Huiskes and Michael S. Lew. 2008. The MIR flickr retrieval evaluation. In *Proceedings of the 1st ACM International Conference on Multimedia Information Retrieval*. ACM, 39–43.

Julian Jaynes. 2000. *The Origin of Consciousness in the Breakdown of the Bicameral Mind*. Houghton Mifflin Harcourt.

Jiwoon Jeon, Victor Lavrenko, and Raghavan Manmatha. 2003. Automatic image annotation and retrieval using cross-media relevance models. In *Proceedings of the 26th Annual International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 119–126.

Justin Johnson, Andrej Karpathy, and Li Fei-Fei. 2016. DenseCap: Fully convolutional localization networks for dense captioning. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Benjamin Johnston, Fangkai Yang, Rogan Mendoza, Xiaoping Chen, and Mary-Anne Williams. 2008. Ontology based object categorization for robots. In *Practical Aspects of Knowledge Management*. Springer, 219–231.

Armand Joulin, Laurens van der Maaten, Allan Jabri, and Nicolas Vasilache. 2015. Learning visual features from large weakly supervised data. *arXiv preprint arXiv:1511.02251* (2015).

Alap Karapurkar. 2008. Modeling human activities. Scholarly Paper Archive, Department of Computer Science, University of Maryland, College Park, MD, 20742.

Andrej Karpathy and Li Fei-Fei. 2015a. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3128–3137.

Andrej Karpathy and Li Fei-Fei. 2015b. Deep visual-semantic alignments for generating image descriptions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ryan Kiros, Yukun Zhu, Ruslan R. Salakhutdinov, Richard Zemel, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015. Skip-thought vectors. In *Advances in Neural Information Processing Systems*. 3276–3284.

Atsuhiro Kojima, Takeshi Tamura, and Kunio Fukunaga. 2002. Natural language description of human activities from video images based on concept hierarchy of actions. *Int. J. Comput. Vis.* 50, 2 (2002), 171–184.

Alexander Koller and Matthew Stone. 2007. Sentence generation as a planning problem. *ACL 2007* (2007), 336.

Ranjay Krishna, Yuke Zhu, Oliver Groth, Justin Johnson, Kenji Hata, Joshua Kravitz, Stephanie Chen, Yannis Kalanditis, Li-Jia Li, David A. Shamma, Michael Bernstein, and Li Fei-Fei. 2016. Visual genome: Connecting language and vision using crowdsourced dense image annotations. *International Journal of Computer Vision* (2016), 45.

Niveda Krishnamoorthy, Girish Malkarnenkar, Raymond Mooney, Kate Saenko, and Sergio Guadarrama. 2013. Generating natural-language video descriptions using text-mined knowledge. *NAACL HLT 2013* (2013), 10.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E. Hinton. 2012. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing Systems*. 1097–1105.

German Kruszewski, Denis Paperno, and Marco Baroni. 2015. Deriving boolean structures from distributional vectors. *Trans. Assoc. Comput. Ling.* 3 (2015), 375–388.

Gaurav Kulkarni, Visruth Premraj, Vicente Ordonez, Sudipta Dhar, Siming Li, Yejin Choi, Alexander C. Berg, and Tamara Berg. 2013. Babytalk: Understanding and generating simple image descriptions. *IEEE Trans. Pattern Anal. Mach. Intell.* 35, 12 (2013), 2891–2903.

Ankit Kumar, Ozan Irsoy, Jonathan Su, James Bradbury, Robert English, Brian Pierce, Peter Ondruska, Ishaan Gulrajani, and Richard Socher. 2016. Ask me anything: Dynamic memory networks for natural language processing. In *ICML*.

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

Polina Kuznetsova, Vicente Ordonez, Alexander C. Berg, Tamara L. Berg, and Yejin Choi. 2012. Collective generation of natural image descriptions. In *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1*. Association for Computational Linguistics, 359–368.

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

Kevin Lai and Dieter Fox. 2010. Object recognition in 3D point clouds using web data and domain adaptation. *Int. J. Robot. Res.* 29, 8 (2010), 1019–1037.

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

Victor Lavrenko, R. Manmatha, and Jiwoon Jeon. 2003. A model for learning the semantics of pictures. In *Advances in Neural Information Processing Systems*. None.

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

Dieu-Thu Le, Jasper Uijlings, and Raffaella Bernardi. 2014. Tuhoi: Trento universal human object interaction dataset. *V&L Net 2014* (2014), 17.

Yann LeCun, Léon Bottou, Yoshua Bengio, and Patrick Haffner. 1998. Gradient-based learning applied to document recognition. *Proc. IEEE* 86, 11 (1998), 2278–2324.

Chee Wee Leong and Rada Mihalcea. 2011. Going beyond text: A hybrid image-text approach for measuring word relatedness. In *IJCNLP*. 1403–1407.

Stephen C. Levinson. 2001. Pragmatics. In *International Encyclopedia of Social and Behavioral Sciences: Vol. 17*. Pergamon, 11948–11954.

Omer Levy and Yoav Goldberg. 2014a. Dependencybased word embeddings. In *Proceedings of the 52nd Annual Meeting of the Association for Computational Linguistics*, Vol. 2. 302–308.

Omer Levy and Yoav Goldberg. 2014b. Neural word embedding as implicit matrix factorization. In *Advances in Neural Information Processing Systems*. 2177–2185.

Li-Jia Li and Li Fei-Fei. 2007. What, where and who? Classifying events by scene and object recognition. In *Proceedings of the IEEE 11th International Conference on Computer Vision*. IEEE, 1–8.

Siming Li, Girish Kulkarni, Tamara L. Berg, Alexander C. Berg, and Yejin Choi. 2011. Composing simple image descriptions using web-scale n-grams. In *Proceedings of the 15th Conference on Computational Natural Language Learning*. Association for Computational Linguistics, 220–228.

Xirong Li, Tiberio Uricchio, Lamberto Ballan, Marco Bertini, Cees G. M. Snoek, and Alberto Del Bimbo. 2016. Socializing the semantic gap: A comparative survey on image tag assignment, refinement, and retrieval. *ACM Comput. Surv.* 49, 1 (2016), 14.

Chin-Yew Lin. 2004. Rouge: A package for automatic evaluation of summaries. In *Text Summarization Branches Out: Proceedings of the ACL-04 Workshop*, Vol. 8.

Changsong Liu and Joyce Yue Chai. 2015. Learning to mediate perceptual differences in situated human-robot dialogue. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence*. AAAI Press, 2288–2294.

Changsong Liu, Lanbo She, Rui Fang, and Joyce Y. Chai. 2014a. Probabilistic labeling for efficient referential grounding based on collaborative discourse. In *ACL (2)*. 13–18.

Jingen Liu, Benjamin Kuipers, and Silvio Savarese. 2011. Recognizing human actions by attributes. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 3337–3344.

Tie-Yan Liu. 2009. Learning to rank for information retrieval. *Found. Trends Inform. Retriev.* 3, 3 (2009), 225–331.

Xiaobai Liu, Yibiao Zhao, and Song-Chun Zhu. 2014b. Single-view 3d scene parsing by attributed grammar. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 684–691.

Jonathan Long, Evan Shelhamer, and Trevor Darrell. 2015. Fully convolutional networks for semantic segmentation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3431–3440.

David G. Lowe. 2004. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vis.* 60, 2 (2004), 91–110.

James MacQueen and others. 1967. Some methods for classification and analysis of multivariate observations. In *Proceedings of the 5th Berkeley Symposium on Mathematical Statistics and Probability*, Vol. 1. 281–297.

Ameesh Makadia, Vladimir Pavlovic, and Sanjiv Kumar. 2008. A new baseline for image annotation. In *Computer Vision–ECCV 2008*. Springer, 316–329.

Alexis Maldonado, Humberto Alvarez, and Michael Beetz. 2012. Improving robot manipulation through fingertip perception. In *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 2947–2954.

Jitendra Malik, Pablo Arbeláez, João Carreira, Katerina Fragkiadaki, Ross Girshick, Georgia Gkioxari, Saurabh Gupta, Bharath Hariharan, Abhishek Kar, and Shubham Tulsiani. 2016. The three Rs of computer vision: Recognition, reconstruction and reorganization. *Pattern Recogn. Lett.* 72 (2016), 4–14.

Mateusz Malinowski, Marcus Rohrbach, and Mario Fritz. 2015. Ask your neurons: A neural-based approach to answering questions about images. In *Proceedings of the IEEE International Conference on Computer Vision*. 1–9.

Jonathan Malmaud, Jonathan Huang, Vivek Rathod, Nick Johnston, Andrew Rabinovich, and Kevin Murphy. 2015. What's cookin'? Interpreting cooking videos using text, speech and vision. In *NAACL 2015*.

Matthew Marge, Claire Bonial, Brendan Byrne, Taylor Cassidy, A. William Evans, Susan G. Hill, and Clare Voss. 2016. Applying the wizard-of-oz technique to multimodal human-robot dialogue. In *Proceedings of RO-MAN (To appear)*.

David Marr. 1982. *Vision: A Computational Investigation Into the Human Representation and Processing of Visual Information*. Henry Holt and Co., Inc., New York, NY.

David R. Martin, Charless C. Fowlkes, and Jitendra Malik. 2004. Learning to detect natural image boundaries using local brightness, color, and texture cues. *IEEE Trans. Pattern Anal. Mach. Intell.* 26, 5 (2004), 530–549.

Cynthia Matuszek*, Nicholas FitzGerald*, Luke Zettlemoyer, Liefeng Bo, and Dieter Fox. 2012. A joint model of language and perception for grounded attribute learning. In *Proceedings of the 2012 International Conference on Machine Learning*. Edinburgh, Scotland.

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

Nikolaos Mavridis. 2015. A review of verbal and non-verbal human–robot interactive communication. *Robot. Auton. Syst.* 63 (2015), 22–35.

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

Julian McAuley, Rahul Pandey, and Jure Leskovec. 2015a. Inferring networks of substitutable and complementary products. In *Proceedings of the 21th ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*. ACM, 785–794.

Julian McAuley, Christopher Targett, Qinfeng Shi, and Anton van den Hengel. 2015b. Image-based recommendations on styles and substitutes. In *Proceedings of the 38th International ACM SIGIR Conference on Research and Development in Information Retrieval*. ACM, 43–52.

Jon D. Mcauliffe and David M. Blei. 2008. Supervised topic models. In *Advances in Neural Information Processing Systems*. 121–128.

Brian McMahan and Matthew Stone. 2015. A Bayesian model of grounded color semantics. *Trans. Assoc. Comput. Ling.* 3 (2015), 103–115.

Ken McRae, George S. Cree, Mark S. Seidenberg, and Chris McNorgan. 2005. Semantic feature production norms for a large set of living and nonliving things. *Behav. Res. Methods* 37, 4 (2005), 547–559.

Chet Meyers and Thomas B. Jones. 1993. *Promoting Active Learning. Strategies for the College Classroom*. ERIC.

Tomas Mikolov, Ilya Sutskever, Kai Chen, Greg S. Corrado, and Jeff Dean. 2013. Distributed representations of words and phrases and their compositionality. In *Advances in Neural Information Processing Systems*. 3111–3119.

George A. Miller. 1995. WordNet: A lexical database for English. *Commun. ACM* 38, 11 (1995), 39–41.

Marvin Minsky. 2006. The emotion machine. New York: Pantheon (2006).

Margaret Mitchell, Xufeng Han, Jesse Dodge, Alyssa Mensch, Amit Goyal, Alex Berg, Kota Yamaguchi, Tamara Berg, Karl Stratos, and Hal Daumé III. 2012. Midge: Generating image descriptions from computer vision detections. In *Proceedings of the 13th Conference of the European Chapter of the Association for Computational Linguistics*. Association for Computational Linguistics, 747–756.

Saif M. Mohammad, Bonnie J. Dorr, Graeme Hirst, and Peter D. Turney. 2013. Computing lexical contrast. *Comput. Ling.* 39, 3 (2013), 555–590.

Raymond J. Mooney. 2008. Learning to connect language and perception. In *AAAI*. 1598–1601.

Raymond J. Mooney. 2013. Grounded Language Learning. (7 2013). 27th AAAI Conference on Artificial Intelligence, Washington 2013 Retrieved November 2, 2015 from http://videolectures.net/aaai2013_mooney_language_learning/.

Yasuhide Mori, Hironobu Takahashi, and Ryuichi Oka. 1999. Image-to-word transformation based on dividing and vector quantizing images with words. In *First International Workshop on Multimedia Intelligent Storage and Retrieval Management*. Citeseer, 1–9.

Frederic Morin and Yoshua Bengio. 2005. Hierarchical probabilistic neural network language model. In *Proceedings of the International Workshop on Artificial Intelligence and Statistics*. Citeseer, 246–252.

Charles William Morris. 1938. Foundations of the theory of signs. (1938).

Venkatesh N. Murthy, Subhransu Maji, and R. Manmatha. 2015. Automatic image annotation using deep learning representations. In *Proceedings of the 5th ACM on International Conference on Multimedia Retrieval*. ACM, 603–606.

Austin Myers, Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. 2015. Affordance detection of tool parts from geometric features. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*.

Douglas L. Nelson, Cathy L. McEvoy, and Thomas A. Schreiber. 2004. The university of south Florida free association, rhyme, and word fragment norms. *Behav. Res. Methods Instrum. Comput.* 36, 3 (2004), 402–407.

Alexander Novikov, Dmitry Podoprikhin, Anton Osokin, and Dmitry Vetrov. 2015. Tensorizing neural networks. In *Advances in Neural Information Processing Systems 28 (NIPS)*.

Aude Oliva and Antonio Torralba. 2001. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vis.* 42, 3 (2001), 145–175.

Vicente Ordonez, Jia Deng, Yejin Choi, Alexander C. Berg, and Tamara L. Berg. 2013. From large scale image categorization to entry-level categories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Vicente Ordonez, Girish Kulkarni, and Tamara L. Berg. 2011. Im2text: Describing images using 1 million captioned photographs. In *Advances in Neural Information Processing Systems*. 1143–1151.

Kishore Papineni, Salim Roukos, Todd Ward, and Wei-Jing Zhu. 2002. BLEU: A method for automatic evaluation of machine translation. In *Proceedings of the 40th Annual Meeting on Association for Computational Linguistics*. Association for Computational Linguistics, 311–318.

Devi Parikh. 2009. Modeling context for image understanding: When, for what, and how? (2009).

Devi Parikh and Kristen Grauman. 2011. Relative attributes. In *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 503–510.

Seyoung Park, Bruce Xiaohan Nie, and Song-Chun Zhu. 2016. Attribute and-or grammar for joint parsing of human attributes, part and pose. *arXiv preprint arXiv:1605.02112* (2016).

Katerina Pastra and Yiannis Aloimonos. 2012. The minimalist grammar of action. *Philos. Trans. Roy. Soc. B: Biol. Sci.* 367, 1585 (2012), 103–117.

Jeffrey Pennington, Richard Socher, and Christopher D. Manning. 2014. Glove: Global vectors for word representation. *Proceedings of the Empiricial Methods in Natural Language Processing (EMNLP 2014)* 12 (2014), 1532–1543.

Jean Piaget. 2013. *Play, Dreams and Imitation in Childhood*. Vol. 25. Routledge.

Tony Plate. 1997. A common framework for distributed representation schemes for compositional structure. *Connectionist Systems for Knowledge Representation and Deduction* (1997), 15–34.

Robert Pless and Richard Souvenir. 2009. A survey of manifold learning for images. *IPSJ Trans. Comput. Vis. Appl.* 1 (2009), 83–94.

J. Pont-Tuset, P. Arbelaez, J. Barron, F. Marques, and J. Malik. 2016. Multiscale combinatorial grouping for image segmentation and object proposal generation. *IEEE Trans. Pattern Anal. Mach. Intelli.* (2016).

Hoifung Poon and Pedro Domingos. 2009. Unsupervised semantic parsing. In *Proceedings of the 2009 Conference on Empirical Methods in Natural Language Processing: Volume 1*. Association for Computational Linguistics, 1–10.

Cecilia Quiroga-Clare. 2003. Language ambiguity: A curse and a blessing. *Transl. J.* 7, 1 (2003).

Gabriel A. Radvansky and Jeffrey M. Zacks. 2014. *Event Cognition*. Oxford University Press.

Mengye Ren, Ryan Kiros, and Richard Zemel. 2015. Exploring models and data for image question answering. In *Advances in Neural Information Processing Systems*. 2953–2961.

Giacomo Rizzolatti and Laila Craighero. 2004. The mirror-neuron system. *Annu. Rev. Neurosci.* 27 (2004), 169–192.

Anna Rohrbach, Marcus Rohrbach, Wei Qiu, Annemarie Friedrich, Manfred Pinkal, and Bernt Schiele. 2014. Coherent multi-sentence video description with variable level of detail. In *Pattern Recognition*. Springer, 184–195.

Anna Rohrbach, Marcus Rohrbach, Niket Tandon, and Bernt Schiele. 2015. A dataset for movie description. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3202–3212.

Stephen Roller and Sabine Schulte Im Walde. 2013. A multimodal LDA model integrating textual, cognitive and visual modalities. In *Proceedings of the 2013 Conference on Empirical Methods in Natural Language Processing*. 1146–1157.

Sascha Rothe and Hinrich Schütze. 2015. AutoExtend: Extending word embeddings to embeddings for synsets and lexemes. In *Proceedings of the ACL*.

Sam T. Roweis and Lawrence K. Saul. 2000. Nonlinear dimensionality reduction by locally linear embedding. *Science* 290, 5500 (2000), 2323–2326.

Deb Roy. 2005. Grounding words in perception and action: Computational insights. *Trends Cogn. Sci.* 9, 8 (2005), 390.

Bryan C. Russell, Antonio Torralba, Kevin P. Murphy, and William T. Freeman. 2008. LabelMe: A database and web-based tool for image annotation. *Int. J. Comput. Vis.* 77, 1-3 (2008), 157–173.

Fereshteh Sadeghi, C. Lawrence Zitnick, and Ali Farhadi. 2015. VISALOGY: Answering visual analogy questions. In *Advances in Neural Information Processing Systems (NIPS-15)*.

Mohammad Amin Sadeghi and Ali Farhadi. 2011. Recognition using visual phrases. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1745–1752.

Karin Kipper Schuler. 2005. VerbNet: A broad-coverage, comprehensive verb lexicon (January 1, 2005). Dissertations available from ProQuest. Paper AAI3179808. http://repository.upenn.edu/dissertations/AAI3179808.

Roy Schwartz, Roi Reichart, and Ari Rappoport. 2015. Symmetric pattern based word embeddings for improved word similarity prediction. *CoNLL 2015* (2015), 258.

Nishant Shukla, Caiming Xiong, and Song-Chun Zhu. 2015. A unified framework for human-robot knowledge transfer. In *Proceedings of the 2015 AAAI Fall Symposium Series*.

Narayanaswamy Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. 2014. Seeing what you're told: Sentence-guided activity recognition in video. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 732–739.

Carina Silberer, Vittorio Ferrari, and Mirella Lapata. 2013. Models of semantic representation with visual attributes. In *ACL (1)*. 572–582.

Karen Simonyan and Andrew Zisserman. 2014. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556* (2014).

Bharat Singh, Xintong Han, Zhe Wu, Vlad I. Morariu, and Larry S. Davis. 2015. Selecting relevant web trained concepts for automated event retrieval. In *Proceedings of the IEEE International Conference on Computer Vision*. 4561–4569.

Jeffrey Mark Siskind. 2001. Grounding the lexical semantics of verbs in visual perception using force dynamics and event logic. *J. Artif. Intell. Res.* 15 (2001), 31–90.

Richard Socher, Andrej Karpathy, Quoc V. Le, Christopher D. Manning, and Andrew Y. Ng. 2014. Grounded compositional semantics for finding and describing images with sentences. *Trans. Assoc. Comput. Ling.* 2 (2014), 207–218.

Richard Socher, Cliff C. Lin, Chris Manning, and Andrew Y. Ng. 2011. Parsing natural scenes and natural language with recursive neural networks. In *Proceedings of the 28th International Conference on Machine Learning (ICML-11)*. 129–136.

Nitish Srivastava and Ruslan Salakhutdinov. 2014. Multimodal learning with deep Boltzmann machines. *J. Mach. Learn. Res.* 15 (2014), 2949–2980.

Mark Steedman. 1996. Surface structure and interpretation. (1996).

Sainbayar Sukhbaatar, Jason Weston, Rob Fergus, and others. 2015. End-to-end memory networks. In *Advances in Neural Information Processing Systems*. 2440–2448.

Douglas Summers-Stay, Ching L. Teo, Yezhou Yang, Cornelia Fermüller, and Yiannis Aloimonos. 2012. Using a minimal action grammar for activity understanding in the real world. In *Proceedings of the 2012 IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 4104–4111.

Douglas Alan Summers-Stay. 2013. Productive vision: Methods for automatic image comprehension. (2013).

Yuyin Sun, Liefeng Bo, and Dieter Fox. 2013. Attribute based object identification. In *Proceedings of the 2013 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2096–2103.

Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich. 2015. Going deeper with convolutions. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1–9.

Kai Sheng Tai, Richard Socher, and Christopher D. Manning. 2015. Improved semantic representations from tree-structured long short-term memory networks. In *ACL*.

Yaniv Taigman, Ming Yang, Marc'Aurelio Ranzato, and Lars Wolf. 2014. Deepface: Closing the gap to human-level performance in face verification. In *Proceedings of the 2014 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1701–1708.

Makarand Tapaswi, Martin Bäuml, and Rainer Stiefelhagen. 2015. Book2movie: Aligning video scenes with book chapters. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1827–1835.

Makarand Tapaswi, Yukun Zhu, Rainer Stiefelhagen, Antonio Torralba, Raquel Urtasun, and Sanja Fidler. 2016. MovieQA: Understanding stories in movies through question-answering. In *IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ben Taskar, Vassil Chatalbashev, Daphne Koller, and Carlos Guestrin. 2005. Learning structured prediction models: A large margin approach. In *Proceedings of the 22nd International Conference on Machine Learning*. ACM, 896–903.

Stefanie Tellex, Ross Knepper, Adrian Li, Daniela Rus, and Nicholas Roy. 2014. Asking for help using inverse semantics. *Proceedings of Robotics: Science and Systems, Berkeley, USA* (2014).

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

Joshua B. Tenenbaum, Vin De Silva, and John C. Langford. 2000. A global geometric framework for nonlinear dimensionality reduction. *Science* 290, 5500 (2000), 2319–2323.

Ching L. Teo, Cornelia Fermüller, and Yiannis Aloimonos. 2015. Fast 2D border ownership assignment. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 5117–5125.

Ching L. Teo, Yezhou Yang, Hal Daumé III, Cornelia Fermüller, and Yiannis Aloimonos. 2012. Towards a Watson that sees: Language-guided action recognition for robots. In *Proceedings of the 2012 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 374–381.

Jesse Thomason, Subhashini Venugopalan, Sergio Guadarrama, Kate Saenko, and Raymond Mooney. 2014. Integrating language and vision to generate natural language descriptions of videos in the wild. In *Proceedings of the 25th International Conference on Computational Linguistics (COLING)*.

Jesse Thomason, Shiqi Zhang, Raymond Mooney, and Peter Stone. 2015. Learning to interpret natural language commands through human-robot dialog. In *Proceedings of the 24th International Joint Conference on Artificial Intelligence (IJCAI)*.

Sebastian Thrun, Wolfram Burgard, and Dieter Fox. 2005. *Probabilistic Robotics*. MIT Press.

Joseph Tighe and Svetlana Lazebnik. 2010. Superparsing: Scalable nonparametric image parsing with superpixels. In *European Conference on Computer Vision*. Springer, 352–365.

Atousa Torabi, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Using descriptive video services to create a large data source for video annotation research. *arXiv preprint arXiv:1503.01070* (2015).

Antonio Torralba, Alexei Efros, and others. 2011. Unbiased look at dataset bias. In *Proceedings of the 2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 1521–1528.

Anne-Marie Tousch, Stéphane Herbin, and Jean-Yves Audibert. 2012. Semantic hierarchies for image annotation: A survey. *Pattern Recogn.* 45, 1 (2012), 333–345.

Zhuowen Tu, Xiangrong Chen, Alan L. Yuille, and Song-Chun Zhu. 2005. Image parsing: Unifying segmentation, detection, and recognition. *Int. J. Comput. Vis.* 63, 2 (2005), 113–140.

Joseph Turian, Lev Ratinov, and Yoshua Bengio. 2010. Word representations: A simple and general method for semi-supervised learning. In *Proceedings of the 48th Annual Meeting of the Association for Computational Linguistics*. Association for Computational Linguistics, 384–394.

Matthew Turk and Alex Pentland. 1991. Eigenfaces for recognition. *J. Cogn. Neurosci.* 3, 1 (1991), 71–86.

Jasper R. R. Uijlings and Vittorio Ferrari. 2015. Situational object boundary detection. In *Proceedings of the 2015 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 4712–4721.

Laurens J. P. van der Maaten, Eric O. Postma, and H. Jaap van den Herik. 2009. Dimensionality reduction: A comparative review. *J. Mach. Learn. Res.* 10, 1–41 (2009), 66–71.

Bernard Vauquois. 1968. Structures profondes et traduction automatique. Le système du CETA. *Rev. Roum. Ling.* 13, 2 (1968), 105–130.

Ramakrishna Vedantam, C. Lawrence Zitnick, and Devi Parikh. 2015. Cider: Consensus-based image description evaluation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 4566–4575.

Subhashini Venugopalan, Marcus Rohrbach, Jeffrey Donahue, Raymond Mooney, Trevor Darrell, and Kate Saenko. 2015a. Sequence to sequence-video to text. In *Proceedings of the IEEE International Conference on Computer Vision*. 4534–4542.

Subhashini Venugopalan, Huijuan Xu, Jeff Donahue, Marcus Rohrbach, Raymond Mooney, and Kate Saenko. 2015b. Translating videos to natural language using deep recurrent neural networks. In *NAACL HLT*.

Oriol Vinyals, Alexander Toshev, Samy Bengio, and Dumitru Erhan. 2015. Show and tell: A neural image caption generator. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 3156–3164.

Luis Von Ahn and Laura Dabbish. 2004. Labeling images with a computer game. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. ACM, 319–326.

Matthew R. Walter, Matthew E. Antone, Ekapol Chuangsuwanich, Andrew Correa, Randall Davis, Luke Fletcher, Emilio Frazzoli, Yuli Friedman, James R. Glass, Jonathan P. How, Jeong Hwan Jeon, Sertac Karaman, Brandon Luders, Nicholas Roy, Stefanie Tellex, and Seth J. Teller. 2015. A situationally aware voice-commandable robotic forklift working alongside people in unstructured outdoor environments. *J. Field Robot.* 32, 4 (2015), 590–628. DOI: http://dx.doi.org/10.1002/rob.21539

Chong Wang, David Blei, and Fei-Fei Li. 2009. Simultaneous image classification and annotation. In *Proceedings of the 2009 IEEE Conference on Computer Vision and Pattern Recognition (CVPR'09)*. IEEE, 1903–1910.

Meng Wang, Bingbing Ni, Xian-Sheng Hua, and Tat-Seng Chua. 2012. Assistive tagging: A survey of multimedia tagging with human-computer joint exploration. *ACM Comput. Surv.* 44, 4 (2012), 25.

Ronald J. Williams. 1988. On the use of backpropagation in associative reinforcement learning. In *Proceedings of the IEEE International Conference on Neural Networks, 1988*. IEEE, 263–270.

Qi Wu, Peng Wang, Chunhua Shen, Anton van den Hengel, and Anthony Dick. 2015b. Ask me anything: Free-form visual question answering based on knowledge from external sources. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhirong Wu, Shuran Song, Aditya Khosla, Fisher Yu, Linguang Zhang, Xiaoou Tang, and Jianxiong Xiao. 2015a. 3d shapenets: A deep representation for volumetric shapes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*. 1912–1920.

Caiming Xiong, Stephen Merity, and Richard Socher. 2016. Dynamic memory networks for visual and textual question answering. In *ICML*.

Huijuan Xu, Subhashini Venugopalan, Vasili Ramanishka, Marcus Rohrbach, and Kate Saenko. 2015b. A multi-scale multiple instance video description network. *arXiv preprint arXiv:1505.05914* (2015).

Kelvin Xu, Jimmy Ba, Ryan Kiros, Kyunghyun Cho, Aaron Courville, Ruslan Salakhudinov, Rich Zemel, and Yoshua Bengio. 2015a. Show, attend and tell: Neural image caption generation with visual attention. In *Proceedings of the 32nd International Conference on Machine Learning*. 2048–2057.

Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang. 2015c. A large-scale car dataset for fine-grained categorization and verification. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yezhou Yang, Cornelia Fermuller, and Yiannis Aloimonos. 2013. Detection of manipulation action consequences (MAC). In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE, 2563–2570.

Yezhou Yang, Cornelia Fermuller, Yiannis Aloimonos, and Eren Erdal Aksoy. 2015. Learning the semantics of manipulation action. *The 53rd Annual Meeting of the Association for Computational Linguistics (ACL)*. Vol. 1. Association for Computational Linguistics, 676–686.

Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. 2015a. Neural self talk: Image understanding via continuous questioning and answering. *arXiv preprint arXiv:1512.03460* (2015).

Yezhou Yang, Yi Li, Cornelia Fermuller, and Yiannis Aloimonos. 2015b. Robot learning manipulation action plans by "watching" unconstrained videos from the world wide web. In *Proceedings of the 29th AAAI Conference on Artificial Intelligence (AAAI-15)*.

Yezhou Yang, Ching Lik Teo, Hal Daumé III, and Yiannis Aloimonos. 2011. Corpus-guided sentence generation of natural images. In *Proceedings of the Conference on Empirical Methods in Natural Language Processing*. Association for Computational Linguistics, 444–454.

Yezhou Yang, Ching L. Teo, Cornelia Fermuller, and Yiannis Aloimonos. 2013. Robots with language: Multi-label visual recognition using NLP. In *Proceedings of the 2013 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 4256–4262.

Bangpeng Yao, Xiaoye Jiang, Aditya Khosla, Andy Lai Lin, Leonidas Guibas, and Li Fei-Fei. 2011. Human action recognition by learning bases of action attributes and parts. In *Proceedings of the 2011 International Conference on Computer Vision*. IEEE, 1331–1338.

Li Yao, Atousa Torabi, Kyunghyun Cho, Nicolas Ballas, Christopher Pal, Hugo Larochelle, and Aaron Courville. 2015. Describing videos by exploiting temporal structure. In *Proceedings of the IEEE International Conference on Computer Vision*. 4507–4515.

Mark Yatskar, Luke Zettlemoyer, and Ali Farhadi. 2016. Situation recognition: Visual semantic role labeling for image understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Serena Yeung, Olga Russakovsky, Greg Mori, and Li Fei-Fei. 2016. End-to-end learning of action detection from frame glimpses in videos. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dani Yogatama, Manaal Faruqui, Chris Dyer, and Noah A. Smith. 2014. Learning word representations with hierarchical sparse coding. *arXiv preprint arXiv:1406.2035* (2014).

Nivasan Yogeswaran, Wenting Dang, William Taube Navaraj, Dhayalan Shakthivel, Saleem Khan, Emre Ozan Polat, Shoubhik Gupta, Hadi Heidari, Mohsen Kaboli, Leandro Lorenzelli, and others. 2015. New materials and advances in making electronic skin for interactive robots. *Adv. Robot.* 29, 21 (2015), 1359–1373.

Peter Young, Alice Lai, Micah Hodosh, and Julia Hockenmaier. 2014. From image descriptions to visual denotations: New similarity metrics for semantic inference over event descriptions. *Trans. Assoc. Comput. Ling.* 2 (2014), 67–78.

Haonan Yu, N. Siddharth, Andrei Barbu, and Jeffrey Mark Siskind. 2015b. A compositional framework for grounding language inference, generation, and acquisition in video. *J. Artif. Intell. Res.* (2015), 601–713.

Haonan Yu and Jeffrey Mark Siskind. 2013. Grounded language learning from video described with sentences. In *ACL (1)*. 53–63.

Haonan Yu, Jiang Wang, Zhiheng Huang, Yi Yang, and Wei Xu. 2015c. Video paragraph captioning using hierarchical recurrent neural networks. *arXiv preprint arXiv:1510.07712* (2015).

Licheng Yu, Eunbyung Park, Alexander C. Berg, and Tamara L. Berg. 2015a. Visual madlibs: Fill in the blank description generation and question answering. In *Proceedings of the IEEE International Conference on Computer Vision*. 2461–2469.

Xiaodong Yu, Cornelia Fermuller, Ching Lik Teo, Yezhou Yang, and Yiannis Aloimonos. 2011. Active scene recognition with vision and language. In *Proceedings of the 2011 IEEE International Conference on Computer Vision (ICCV)*. IEEE, 810–817.

Konstantinos Zampogiannis, Yezhou Yang, Cornelia Fermuller, and Yiannis Aloimonos. 2015. Learning the spatial semantics of manipulation actions through preposition grounding. In *Proceedings of the 2015 IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 1389–1396.

John M. Zelle and Raymond J. Mooney. 1996. Learning to parse database queries using inductive logic programming. In *Proceedings of the National Conference on Artificial Intelligence*. 1050–1055.

Luke S. Zettlemoyer and Michael Collins. 2005. Learning to map sentences to logical form: Structured classification with probabilistic categorial grammars. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*.

Dengsheng Zhang, Md Monirul Islam, and Guojun Lu. 2012. A review on automatic image annotation techniques. *Pattern Recogn.* 45, 1 (2012), 346–362.

Rong Zhao and William I. Grosky. 2002. Bridging the semantic gap in image retrieval. *Distributed Multimedia Databases: Techniques and Applications* (2002), 14–36.

Wenyi Zhao, Rama Chellappa, P. Jonathon Phillips, and Azriel Rosenfeld. 2003. Face recognition: A literature survey. *ACM Comput. Surv.* 35, 4 (2003), 399–458.

Shuai Zheng, Sadeep Jayasumana, Bernardino Romera-Paredes, Vibhav Vineet, Zhizhong Su, Dalong Du, Chang Huang, and Philip H. S. Torr. 2015. Conditional random fields as recurrent neural networks. In *Proceedings of the IEEE International Conference on Computer Vision*. 1529–1537.

Yuke Zhu, Oliver Groth, Michael Bernstein, and Li Fei-Fei. 2016. Visual7W: Grounded question answering in images. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*.

Yukun Zhu, Ryan Kiros, Rich Zemel, Ruslan Salakhutdinov, Raquel Urtasun, Antonio Torralba, and Sanja Fidler. 2015a. Aligning books and movies: Towards story-like visual explanations by watching movies and reading books. In *Proceedings of the IEEE International Conference on Computer Vision*. 19–27.

Yuke Zhu, Ce Zhang, Christopher Ré, and Li Fei-Fei. 2015b. Building a large-scale multimodal knowledge base for visual question answering. *arXiv preprint arXiv:1507.05670* (2015).

Received February 2016; revised July 2016; accepted October 2016