An Analysis of Action Recognition Datasets for Language and Vision Tasks

Spandana Gella and Frank Keller
Institute for Language, Cognition and Computation
School of Informatics, University of Edinburgh
10 Crichton Street, Edinburgh EH8 9AB
S.Gella@sms.ed.ac.uk, keller@inf.ed.ac.uk

Abstract
A large amount of recent research has focused on tasks that combine language and vision, resulting in a proliferation of datasets and methods. One such task is action recognition, whose applications include image annotation, scene understanding and image retrieval. In this survey, we categorize the existing approaches based on how they conceptualize this problem and provide a detailed review of existing datasets, highlighting their diversity as well as advantages and disadvantages. We focus on recently developed datasets which link visual information with linguistic resources and provide a fine-grained syntactic and semantic analysis of actions in images.

1 Introduction
Action recognition is the task of identifying the action being depicted in a video or still image. The task is useful for a range of applications such as generating descriptions, image/video retrieval, surveillance, and human–computer interaction. It has been widely studied in computer vision, often on videos (Nagel, 1994; Forsyth et al., 2005), where motion and temporal information provide cues for recognizing actions (Taylor et al., 2010). However, many actions are recognizable from still images, see the examples in Figure 1. Due to the absence of motion cues and temporal features (Iki zler et al., 2008) action recognition from stills is more challenging. Most of the existing work can be categorized into four tasks: (a) action classification (AC); (b) determining human–object interaction (HOI); (c) visual verb sense disambiguation (VSD); and (d) visual semantic role labeling (VSRL). In Figure 2 we illustrate each of these tasks and show how they are related to each other.

Until recently, action recognition was studied as action classification on small-scale datasets with a limited number of predefined actions labels (Iki zler et al., 2008; Gupta et al., 2009; Yao and Fei Fei, 2010; Everingham et al., 2010; Yao et al., 2011). Often the labels in action classification tasks are verb phrases or a combination of verb and object such as playing baseball, riding horse. These datasets have helped in building models and understanding which aspects of an image are important for classifying actions, but most methods are not scalable to larger numbers of actions (Ramanathan et al., 2015). Action classification models are trained on images annotated with mutually exclusive labels, i.e., the assumption is that only a single label is relevant for a given image. This ignores the fact that actions such as holding bicycle and riding bicycle can co-occur in the same image. To address these issues and also to understand the range of possible interactions between humans and objects, the human–object interaction (HOI) detection task has been proposed, in which all possible interactions between a human and a given object have to be identified (Le et al., 2014; Chao et al., 2015; Lu et al., 2016).

However, both action classification and HOI detection do not consider the ambiguity that arises when verbs are used as labels, e.g., the verb play has multiple meanings in different contexts. On the other hand, action labels consisting of verb-object pairs can miss important generalizations:
Figure 2: Categorization of action recognition tasks in images

riding horse and riding elephant both instantiate the same verb semantics, i.e., riding animal. Thirdly, existing action labels miss generalizations across verbs, e.g., the fact that fixing bike and repairing bike are semantically equivalent, in spite of the use of different verbs. These observations have led authors to argue that actions should be analyzed at the level of verb senses. Gella et al. (2016) propose the new task of visual verb sense disambiguation (VSD), in which a verb–image pair is annotated with a verb sense taken from an existing lexical database (OntoNotes in this case). While VSD handles distinction between different verb senses, it does not identify or localize the objects that participate in the action denoted by the verb. Recent work (Gupta and Malik, 2015; Yatskar et al., 2016) has filled this gap by proposing the task of visual semantic role labeling (VSRL), in which images are labeled with verb frames, and the objects that fill the semantic roles of the frame are identified in the image.

In this paper, we provide a unified view of action recognition tasks, pointing out their strengths and weaknesses. We survey existing literature and provide insights into existing datasets and models for action recognition tasks.

2 Datasets for Action Recognition

We give an overview of commonly used datasets for action recognition tasks in Table 1 and group them according to subtask. We observe that the number of verbs covered in these datasets is often smaller than the number of action labels reported (see Table 1, columns #V and #L) and in many cases the action label involves object reference. A few of the first action recognition datasets such as the Ikizler and Willow datasets (Ikizler et al., 2008; Delaïtre et al., 2010) had action labels such as throwing and running; they were taken from the sports domain and exhibited diversity in camera view point, background and resolution. Then datasets were created to capture variation in human poses in the sports domain for actions such as tennis serve and cricket bowling; typically features based on poses and body parts were used to build models (Gupta et al., 2009). Further datasets were created based on the intuition that object information helps in modeling action recognition (Li and Fei-Fei, 2007; Ikizler-Cinbis and Sclaroff, 2010), which resulted in the use of action labels such as riding horse or riding bike (Everingham et al., 2010; Yao et al., 2011). Not only were most of these datasets domain specific, but the labels were also manually selected and mutually exclusive, i.e., two actions cannot co-occur in the same image. Also, most of these datasets do not localize objects or identify their semantic roles.

2.1 Identifying Visual Verbs and Verb Senses

The limitations with early datasets (small scale, domain specificity, and the use of ad-hoc labels that combine verb and object) have been recently addressed in a number of broad-coverage datasets that offer linguistically motivated labels. Often these datasets use existing linguistic resources such as VerbNet (Schuler, 2005), OntoNotes (Hovy et al., 2006) and FrameNet (Baker et al., 1998) to classify verbs and their senses. This allows for a more general, semantically motivated treatment of verbs and verb phrases, and also takes into account that not all verbs are depictable. For example, abstract verbs such as presuming and acquiring are not depictable at all, while other verbs have both depictable and non-depictable senses: play is non-depictable in playing with emotions, but depictable in playing instrument and playing sport. The process of identifying depictable verbs or verb senses is used by Ronchi and Perona (2015), Gella et al. (2016) and Yatskar et al. (2016) to identify visual verbs, visual verb senses, and the semantic roles of the participating objects respectively. In all the cases the process of identifying visual verbs or senses is carried out by human annotators via crowd-sourcing platforms. Visualness labels for 935 OntoNotes verb senses corresponding to 154 verbs is provided by Gella et al. (2016), while Yatskar et al. (2016) provides visualness labels for 9683 FrameNet verbs.
2.2 Datasets Beyond Action Classification

Over the last few years tasks that combine language and vision such as image description and visual question answering have gained much attention. This has led to the creation of new, large datasets such as MSCOCO (Chen et al., 2015) and the VQA dataset (Antol et al., 2015). Although these datasets are not created for action recognition, a number of attempts have been made to use the verbs present in image descriptions to annotate actions. The COCO-a, VerSe and VCCOCO-SRL datasets all use the MSCOCO image descriptions to annotate fine-grained aspects of interaction and semantic roles.

**HICO:** The HICO dataset has 47.8k images annotated with 600 categories of human-object interactions with 111 verbs applying to 80 object categories of MSCOCO. It is annotated to include diverse interactions for objects and has an average of 6.5 distinct interactions per object category. Unlike other HOI datasets such as TUHOI which label interactions as verbs and ignore senses, the HOI categories of HICO are based on WordNet verb senses. The HICO dataset also has multiple annotations per object and it incorporates the information that certain interactions such as riding a bike and holding a bike often co-occur. However, it fails to include annotations to distinguish between multiple senses of a verb.

**Visual Genome:** The dataset created by Krishna et al. (2016) has dense annotations of objects, attributes, and relationships between objects. The Visual Genome dataset contains 105k images with 40k unique relationships between objects. Unlike other HOI datasets such as HICO, visual genome relationships also include prepositions, comparative and prepositional phrases such as near and taller than, making the visual relationship task more generic than action recognition. Krishna et al. (2016) combine all the annotations of objects, relationships, and attributes into directed graphs known as scene graphs.

**COCO-a:** Ronchi and Perona (2015) present Visual VerbNet (VVN), a list of 140 common visual verbs manually mined from English VerbNet (Schuler, 2005). The coverage of visual verbs in this dataset is not complete, as many visual verbs such as dive, perform and shoot are not included. This also highlights a bias in this dataset as the authors relied on occurrence in MSCOCO as a verification step to consider a verb as visual. They annotated 10k images containing human subjects with one of the 140 visual verbs, for 80 MSCOCO objects. This dataset has better coverage of human-object interactions than the HICO dataset despite of missing many visual verbs.

**VerSe:** Gella et al. (2016) created a dataset of 3.5k images sampled from the MSCOCO and TUHOI datasets and annotated it with 90 verbs and their OntoNotes senses to distinguish different verb senses using visual context. This is the first dataset that aims to annotate all visual senses

### Table 1: Comparison of various existing action recognition datasets.

<table>
<thead>
<tr>
<th>Dataset</th>
<th>Task</th>
<th>#L</th>
<th>#V</th>
<th>Obj</th>
<th>Imgs</th>
<th>Des</th>
<th>Sen</th>
<th>Cln</th>
<th>ML</th>
<th>Resource</th>
<th>Example Labels</th>
</tr>
</thead>
<tbody>
<tr>
<td>Ikizler (Ikizler et al., 2008)</td>
<td>AC</td>
<td>6</td>
<td>6</td>
<td>0</td>
<td>467</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>running, walking</td>
<td></td>
</tr>
<tr>
<td>Sports Dataset (Gupta et al., 2009)</td>
<td>AC</td>
<td>6</td>
<td>6</td>
<td>4</td>
<td>300</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>tennis serve, cricket bowling</td>
<td></td>
</tr>
<tr>
<td>Willow (Delaitre et al., 2010)</td>
<td>AC</td>
<td>7</td>
<td>6</td>
<td>5</td>
<td>986</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>riding bike, photographing</td>
<td></td>
</tr>
<tr>
<td>PPMI (Yao and Fei-Fei, 2010)</td>
<td>AC</td>
<td>24</td>
<td>2</td>
<td>12</td>
<td>4.8k</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>play guitar, hold violin</td>
<td></td>
</tr>
<tr>
<td>Stanford 40 Actions (Yao et al., 2011)</td>
<td>AC</td>
<td>40</td>
<td>33</td>
<td>31</td>
<td>9.5k</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>cut vegetables, ride horse</td>
<td></td>
</tr>
<tr>
<td>PASCAL 2012 (Everingham et al., 2015)</td>
<td>AC</td>
<td>11</td>
<td>9</td>
<td>6</td>
<td>4.5k</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>riding bike, riding horse</td>
<td></td>
</tr>
<tr>
<td>89 Actions (Le et al., 2013)</td>
<td>AC</td>
<td>89</td>
<td>36</td>
<td>19</td>
<td>2k</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>ride bike, fix bike</td>
<td></td>
</tr>
<tr>
<td>MPII Human Pose (Andriluka et al., 2014)</td>
<td>AC</td>
<td>410</td>
<td>66</td>
<td>40.5k</td>
<td>N</td>
<td>Y</td>
<td>N</td>
<td>N</td>
<td>riding car, hair styling</td>
<td></td>
<td></td>
</tr>
<tr>
<td>TUHOI (Le et al., 2014)</td>
<td>HOI</td>
<td>2974</td>
<td>189</td>
<td>10.8k</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>sit on chair, play with dog</td>
<td></td>
<td></td>
</tr>
<tr>
<td>COCO-a (Ronchi and Perona, 2015)</td>
<td>HOI</td>
<td>140</td>
<td>10k</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>VerbNet walk bike, hold bike</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Google Images (Ramanathan et al., 2015)</td>
<td>AC</td>
<td>2880</td>
<td>–</td>
<td>102k</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>N</td>
<td>riding horse, riding camel</td>
<td></td>
<td></td>
</tr>
<tr>
<td>HICO (Chao et al., 2015)</td>
<td>HOI</td>
<td>111</td>
<td>80</td>
<td>47k</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>WordNet ride#v#1, hold#v#2 bike</td>
<td></td>
</tr>
<tr>
<td>VCCOCO-SRL (Gupta and Malik, 2015)</td>
<td>VSSL</td>
<td>–</td>
<td>26</td>
<td>48</td>
<td>10k</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>Y</td>
<td>verb: hit, instr. bat; obj: ball</td>
<td></td>
</tr>
<tr>
<td>imSitu (Yatskar et al., 2016)</td>
<td>VSSL</td>
<td>–</td>
<td>504</td>
<td>126k</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>FrameNet verb: ride; agent: girl#n#2</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual Genome (Krishna et al., 2016)</td>
<td>VSD</td>
<td>163</td>
<td>90</td>
<td>3.5k</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>OntoNotes ride.v.01, play.v.02</td>
<td></td>
<td></td>
</tr>
<tr>
<td>VerSe (Gella et al., 2016)</td>
<td>VSD</td>
<td>163</td>
<td>90</td>
<td>3.5k</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>Y</td>
<td>man playing frisbee</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Visual Genome (Krishna et al., 2016)</td>
<td>VRD</td>
<td>42.3k</td>
<td>–</td>
<td>108k</td>
<td>N</td>
<td>Y</td>
<td>Y</td>
<td>N</td>
<td>man playing frisbee</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
of a verb. However, the total number of images annotated and number of images for some senses is relatively small, which makes it difficult to use this dataset to train models. The authors further divided their 90 verbs into motion and non-motion verbs according to Levin (1993) verb classes and analyzed visual ambiguity in the task of visual sense disambiguation.

**VCOCO-SRL:** Gupta and Malik (2015) annotated a dataset of 16k person instances in 10k images with 26 verbs and associated objects in the scene with the semantic roles for each action. The main aim of the dataset is to build models for visual semantic role labeling in images. This task involves identifying the actions depicted in an image, along with the people and objects that instantiate the semantic roles of the actions. In the VCOCO-SRL dataset, each person instance is annotated with a mean of 2.8 actions simultaneously.

**imSitu:** Yatskar et al. (2016) annotated a large dataset of 125k images with 504 verbs, 1.7k semantic roles and 11k objects. They used FrameNet verbs, frames and associated objects or scenes with roles to develop the dataset. They annotate every image with a single verb and the semantic roles of the objects present in the image. VCOCO-SRL is the dataset most similar to imSitu, however VCOCO-SRL includes localization information of agents and all objects and provides multiple action annotations per image. On the other hand, imSitu is the dataset that covers highest number of verbs, while also omitting many commonly studied polysemous verbs such as *play*.

### 2.3 Diversity in Datasets

With the exception of a few datasets such as COCO-a, VerSe, imSitu all action recognition datasets have manually picked labels or focus on covering actions in specific domains such as sports. Alternatively, many datasets only cover actions relevant to specific object categories such as musical instruments, animals and vehicles. In the real world, people interact with many more objects and perform actions relevant to a wide range of domains such as personal care, household activities, or socializing. This limits the diversity and coverage of existing action recognition datasets. Recently proposed datasets partly handle this issue by using generic linguistic resources to extend the vocabulary of verbs in action labels. The diversity issue has also been highlighted and addressed in recent video action recognition datasets (Caba Heilbron et al., 2015; Sigurdsson et al., 2016), which include generic household activities. An analysis of various image description and question answering datasets by Ferraro et al. (2015) shows the bias in the distribution of word categories. Image description datasets have a higher distribution of nouns compared to other word categories, indicating that the descriptions are object specific, limiting their usefulness for action-based tasks.

### 3 Relevant Language and Vision Tasks

Template based description generation systems for both videos and images rely on identifying subject–verb–object triples and use language modeling to generate or rank descriptions (Yang et al., 2011; Thomason et al., 2014; Bernardi et al., 2016). Understanding actions also plays an important role in question answering, especially when the question is pertaining to an action depicted in the image. There are some specifically curated question answering datasets which target human activities or relationships between a pair of objects (Yu et al., 2015). Mallya and Lazebnik (2016) have shown that systems trained on action recognition datasets could be used to improve the accuracy of visual question answering systems that handle questions related to human activity and human–object relationships. Action recognition datasets could be used to learn actions that are visually similar such as *interacting with panda* and *feeding a panda* or *tickling a baby* and *calming a baby*, which cannot be learned from text alone (Ramanathan et al., 2015). Visual semantic role labeling is a crucial step for grounding actions in the physical world (Yang et al., 2016).

### 4 Action Recognition Models

Most of the models proposed for action classification and human–object interaction tasks rely on identifying higher-level visual cues present in the image, including human bodies or body parts (Ikizer et al., 2008; Gupta et al., 2009; Yao et al., 2011; Andriluka et al., 2014), objects (Gupta et al., 2009), and scenes (Li and Fei-Fei, 2007). Higher-level visual cues are obtained through low-level features extracted from the image such as Scale Invariant Feature Transforms (SIFT), Histogram of Oriented Gradients (HOG), and Spatial Envelopes (Gist) features (Lowe, 1999; Dalal and Triggs, 2005).
2005). These are useful in identifying key points, detecting humans, and scene or background information in images, respectively. In addition to identifying humans and objects, the relative position or angle between a human and an object is useful in learning human–object interactions (Le et al., 2014). Most of the existing approaches rely on learning supervised classifiers over low-level features to predict action labels.

More recent approaches are based on end-to-end convolutional neural network architectures which learn visual cues such as objects and image features for action recognition (Chao et al., 2015; Zhou et al., 2016; Mallya and Lazebnik, 2016). While most of the action classification models rely solely on visual information, models proposed for human–object interaction or visual relationship detection sometimes combine human and object identification (using visual features) with linguistic knowledge (Le et al., 2014; Krishna et al., 2016; Lu et al., 2016). Other work on identifying actions, especially methods that focus on relationships that are infrequent or unseen, utilize word vectors learned on large text corpora as an additional source of information (Lu et al., 2016). Similarly, Gella et al. (2016) show that embeddings generated from textual data associated with images (object labels, image descriptions) is useful for visual verb sense disambiguation, and is complementary to visual information.

5 Discussion

Linguistic resources such as WordNet, OntoNotes, and FrameNet play a key role in textual sense disambiguation and semantic role labeling. The visual action disambiguation and visual semantic role labeling tasks are extensions of their textual counterparts, where context is provided as an image instead of as text. Linguistic resources therefore have to play a key role if we are to make rapid progress in these language and vision tasks. However, as we have shown in this paper, only a few of the existing datasets for action recognition and related tasks are based on linguistic resources (Chao et al., 2015; Gella et al., 2016; Yatskar et al., 2016). This is despite the fact that the WordNet noun hierarchy (for example) has played an important role in recent progress in object recognition, by virtue of underlying the ImageNet database, the de-facto standard for this task (Russakovsky et al., 2015). The success of ImageNet for objects has in turn helped NLP tasks such as bilingual lexicon induction (Vulić et al., 2016). In our view, language and vision datasets that are based on the WordNet, OntoNotes, or FrameNet verb sense inventories can play a similar role for tasks such as action recognition or visual semantic role labeling, and ultimately be useful also for more distantly related tasks such as language grounding.

Another argument for linking language and vision datasets with linguistic resources is that this enables us to deploy the datasets in a multilingual setting. For example a polysemous verb such as ride in English has multiple translations in German and Spanish, depending on the context and the objects involved. Riding a horse is translated as reiten in German and cabalgar in Spanish, whereas riding a bicycle is translated as fahren in German and pedalear in Spanish. In contrast, some polysemous verb (e.g., English play) are always translated as the same verb, independent of sense (spielen in German). Such sense mappings are discoverable from multilingual lexical resources (e.g., BabelNet,Navigli and Ponzetto 2010), which makes it possible to construct language and vision models that are applicable to multiple languages. This opportunity is lost if language and vision dataset are constructed in isolation, instead of using existing linguistic resources.

6 Conclusions

In this paper, we have shown the evolution of action recognition datasets and tasks from simple ad-hoc labels to the fine-grained annotation of verb semantics. It is encouraging to see the recent increase in datasets that deal with sense ambiguity and annotate semantic roles, while using standard linguistic resources. One major remaining issue with existing datasets is their limited coverage, and the skewed distribution of verbs or verb senses. Another challenge is the inconsistency in annotation schemes and task definitions across datasets. For example Chao et al. (2015) used WordNet senses as interaction labels, while Gella et al. (2016) used the more coarse-grained OntoNotes senses. Yatskar et al. (2016) used FrameNet frames for semantic role annotation, while Gupta and Malik (2015) used manually curated roles. If we are to develop robust, domain independent models, then we need to standardize annotation schemes and use the same linguistic resources across datasets.
References


Vignesh Ramanathan, Congcong Li, Jia Deng, Wei Han, Zhen Li, Kunlong Gu, Yang Song, Samy Bengio, Chuck Rosenberg, and Li Fei-Fei. 2015. Learning semantic relationships for better action retrieval in images. In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. pages 1100–1109.


