Combining Visual Recognition and Computational Linguistics

Linguistic Knowledge for Visual Recognition and Natural Language Descriptions of Visual Content

Thesis for obtaining the title of
Doctor of Engineering Science
(Dr.-Ing.)
of the Faculty of Natural Science and Technology I
of Saarland University

by
Marcus Rohrbach, M.Sc.

Saarbrücken
March 2014

Day of Colloquium	26th of March, 2014
Dean of the Faculty	Univ.-Prof. Dr. Mark Groves Saarland University, Germany

Examination Committee

Chair	Prof. Dr. Matthias Hein Saarland University, Germany
Reviewer, Advisor	Prof. Dr. Bernt Schiele Max Planck Institute for Informatics, Germany Saarland University, Germany
Reviewer	Prof. Dr. Manfred Pinkal Saarland University, Germany
Reviewer	Prof. Fei-Fei Li, Ph.D. Stanford University
Academic Assistant	Dr. Björn Andres Max Planck Institute for Informatics, Germany

ABSTRACT

Extensive efforts are being made to improve visual recognition and semantic understanding of language. However, surprisingly little has been done to exploit the mutual benefits of combining both fields. In this thesis we show how the different fields of research can profit from each other.

First, we scale recognition to 200 unseen object classes and show how to extract robust semantic relatedness from linguistic resources. Our novel approach extends zero-shot to few shot recognition and exploits unlabeled data by adopting label propagation for transfer learning.

Second, we capture the high variability but low availability of composite activity videos by extracting the essential information from text descriptions. For this we recorded and annotated a corpus for fine-grained activity recognition. We show improvements in a supervised case but we are also able to recognize unseen composite activities.

Third, we present a corpus of videos and aligned descriptions. We use it for grounding activity descriptions and for learning how to automatically generate natural language descriptions for a video. We show that our proposed approach is also applicable to image description and that it outperforms baselines and related work.

In summary, this thesis presents a novel approach for automatic video description and shows the benefits of extracting linguistic knowledge for object and activity recognition as well as the advantage of visual recognition for understanding activity descriptions.

ZUSAMMENFASSUNG

Trotz umfangreicher Anstrengungen zur Verbesserung der die visuelle Erkennung und dem automatischen Verständnis von Sprache, ist bisher wenig getan worden, um diese beiden Forschungsbereiche zu kombinieren. In dieser Dissertation zeigen wir, wie beide voneinander profitieren können.

Als erstes skalieren wir Objekterkennung zu 200 ungesehen Klassen und zeigen, wie man robust semantische Ähnlichkeiten von Sprachressourcen extrahiert. Unser neuer Ansatz kombiniert Transfer und halbüberwachten Lernverfahren und kann so Daten ohne Annotation ausnutzen und mit keinen als auch mit wenigen Trainingsbeispielen auskommen.

Zweitens erfassen wir die hohe Variabilität aber geringe Verfügbarkeit von Videos mit zusammengesetzten Aktivitäten durch Extraktion der wesentlichen Informationen aus Textbeschreibungen. Wir verbessern überwachtes Training als auch die Erkennung von ungesehenen Aktivitäten.

Drittens stellen wir einen parallelen Datensatz von Videos und Beschreibungen vor. Wir verwenden ihn für Grounding von Aktivitätsbeschreibungen und um die automatische Generierung natürlicher Sprache für ein Video zu erlernen. Wir zeigen, dass sich unsere Ansatz auch für Bildbeschreibung einsetzten lässt und das er bisherige Ansätze übertrifft.

Zusammenfassend stellt die Dissertation einen neuen Ansatz zur automatische Videobeschreibung vor und zeigt die Vorteile von sprachbasierten Ähnlichkeitsmaßen für die Objekt- und Aktivitätserkennung als auch umgekehrt.

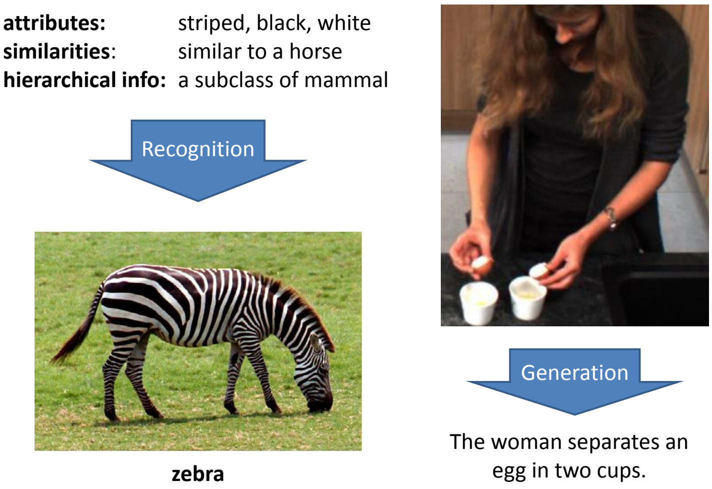
CONTENTS

INTRODUCTION 1

Contents

Our two most important means of communication as humans are the visual and linguistic channel. Humans can easily relate and convert between both channels. For example, given a linguistic description of objects, humans can visually recognize these objects, even if they might not have seen them before. Descriptions given in the form of hierarchical categories (e.g., *a mammal*), attributes (*striped, black, and white*), or similarities, (*similar to a horse*), allow humans to recognize visual categories. For the above examples most humans would be able to recognise the *zebra* shown in Figure 1.1(a). Furthermore, people can not only use linguistic information to guide their visual recognition but also generate descriptions of activities and objects they have seen. For example a human could easily describe the video depicted in Figure 1.1(b) with *"The woman separates an egg in two cups."*.

While humans are proficient in such tasks, automatically recognizing an object based on text-mined information or describing human activities in video with a sentence requires answering core research questions of computer vision and computational linguistics. We have witnessed computer vision research to advance to a mature field achieving impressive performance for automatically detecting people (e.g. Benenson *et al.*, 2013), their poses (e.g. Yang and Ramanan, 2013; Pishchulin *et al.*, 2013) and actions (e.g. Wang *et al.*, 2013), and discriminating large number of objects (e.g. Dean *et al.*, 2013). However, in the case when training data is sparse visual recognition approaches significantly drop in performance, e.g. for recognition of unseen categories, fine-grained recognition, or composite activity recognition. Furthermore, the output of such systems is in most cases semantic labels which might be useful for automated processing but difficult to communicate to humans, who communicate using natural language.


attributes: striped, black, white
similarities: similar to a horse
hierarchical info: a subclass of mammal
Recognition
zebra
Generation
The woman separates an egg in two cups.

(a) Category-level, text-mined information for recognition of novel classes

(b) Sentence generated for a specific video.

Figure 1.1: Examples for textual descriptions and visual content.

On the other side, computational linguistics faces the challenge to understand external references from text representations alone while they might be visually observable. However, it has devised sophisticated tools for extracting semantic similarity from text corpora (e.g. Szarvas *et al.*, 2011) and automatically translating between different languages (e.g. Koehn, 2010).

Inspired by the abilities of humans to extract knowledge from visual and linguistic data, motivated by the potential benefits of combining both fields, and driven by the need to communicate our visual recognition to humans, this thesis investigates three directions how to combine and convert between visual and linguistic information. The first direction is *Visual knowledge transfer using linguistic semantic relatedness*. Here we mine hierarchical information, as well as object-attribute and in-between-object associations from diverse linguistic resources such as Wikipedia, WordNet, and the World Wide Web to improve object recognition, focusing on the challenging setting in visual recognition where there are no or only few instance labels available for a certain category. This is presented in Chapters 3, 4, and 7.

In the second direction, *Script data for activity recognition*, we leverage the large variation of executing composite activities by collecting textual instructions how to perform complex activities. This allows improving recognition of composite activities in videos by focusing on the relevant parts and also recognizing novel activities only based on textual descriptions. The collection of the activity recognition dataset and visual recognition approach is described in Chapter 5 and the script data based recognition in Chapter 6.

For the third direction we examine *Natural language descriptions of visual content* which are specific to a certain video or image. This allows us to ground the text in the video and use video-features to estimate similarities of activity descriptions. Using the videos aligned with textual descriptions we approach the challenging task of automatically describing a video or image with natural sentence description. We propose an approach which allows learning this conversion fully from data rather than manually specifying rules or templates on the visual or linguistic side.

It is clear that these three directions cannot cover all potential interactions between visual recognition and computational linguistics. Other directions include generating visual content from language descriptions (e.g. Zitnick *et al.*, 2013; Liang *et al.*, 2013) and jointly modeling visual and linguistic information which we discuss as direction for future work in Chapter 10. However, we believe that the chosen directions approach challenging topics and are of high relevance in computer vision and computational linguistics. As we show in Chapter 2 the topics discussed in this thesis have received increased interest in recent years and have been picked up in both communities during the course of this thesis. Still, we believe the research looking at the interaction between the two modalities is just at the beginning and we discuss possible directions in the last chapter.

In the following we first analyze the challenges and how we approach them with respect to these three parts. Afterwards we discuss the respective contributions of this thesis and at the end of this chapter we provide an outline of this thesis.

1.1 CHALLENGES FOR COMBINING VISUAL AND LINGUISTIC MODALITIES

One of the fundamental differences between the visual and the linguistic modality is the level of abstraction. The basic data unit of the visual modality is a (photographic) image or video which always shows a specific instance of a category, or even more precisely a certain instance for a specific viewpoint, lighting, pose, time etc. For example Figure 1.1(a) shows one specific instance of the category *zebra* from a side view, eating grass. In contrast to this, the basic semantic unit of the linguistic modality are words (which are strings of characters or phonemes for spoken language, but we will restrict ourselves to written linguistic expressions in this thesis). Although a word might *refer* to a specific instance, the word, i.e. the string, always *represents* a category of objects, activities, or attributes, abstracting from a specific instance. Although we will leave out named entities such as *Eifel Tower* in the further discussion, an image of the Eifel Tower is still more specific, e.g. in viewpoint, than the concept Eifel Tower. Interestingly this difference, instance versus category level representation, is also what defines the core challenges in visual recognition and is also an important topic in computational linguistics. In visual recognition we are interested in defining or learning models which abstract over a specific image or video to understand the visual characteristic of a category. In computational linguistics, when automatically parsing a text we frequently face the inverse challenge of trying to identify intra and the extra linguistic references (anaphora resolution / grounding) of a word or phrase. These problems arise because words typically represent concepts rather than instances and because anaphors, synonyms, hypernyms, or metaphorical expressions are used to refer to the identical object in the real world.

Understanding that the visual and linguistic modalities have different levels of abstraction is important when trying to combine both modalities. In Chapters 4 and 6 we use linguistic knowledge at category rather than instance level for visual knowledge transfer, i.e. we use linguistic knowledge at the level where its most expressive that is at level of its basic representation. In Chapter 8 we exploit visual representations at a point where linguistic knowledge is less powerful: estimating similarities between very concrete and specific composite activities is very challenging using textual information alone, e.g. deciding that "*slicing a zucchini*" is more similar to "*cutting a cucumber*" than to "*peeling a zucchini*". Grounding these activities in video and exploiting visual similarity allows us to model this relationship much better. Finally, in Chapter 9, when describing visual input with natural language, we put the point of interaction, our semantic representation, at the concept level, i.e. we recognize the category of activity and objects and leave concrete realization of sentences to a language model rather than inferring it from the visual representation.

In the following subsections we identify the specific challenges of the tasks we want to solve and also discuss how we attack those challenges in this thesis.

1.1.1 Extracting linguistic knowledge for object recognition

Visual recognition of objects in natural images, i.e. identifying the category of the shown instance, is a challenging problem due to typically different viewpoints, backgrounds, varying pose, and frequently high intra-class variation, but low inter-class variation. However, in the following we want to focus on the specific challenges of using linguistic knowledge for visual recognition.

Integration of linguistic knowledge. The first question is how to integrate the linguistic knowledge into the visual recognition process. As we discussed above it is important to notice that the basic linguistic representation is on concept rather than instance level. It seems to be more powerful to use knowledge about concepts rather than individual instances.

In this thesis we thus use semantic relatedness mined from linguistic resources to estimate similarity between object categories and link them to visual object classifiers. We discuss an attribute and direct similarity model we employ as well as our extensions in Chapter 4.

Semantic meaning vs. linguistic representation. The linguistic modality uses words as representation but they are not identical to a single semantic concept. A word or phrase can have multiple meanings, i.e. represent polysemous (related) as well as homonymous (unrelated) concepts, but also a semantic concept can be represented by multiple synonym words. We thus have a one-to-many relationship in both directions. On the visual side we frequently work with classifiers which are tagged with a single word rather than a full meaning definition. We thus have only a loose linkage between visual classifiers und words used in the text corpora.

We approach these problems mainly by three steps: First, we never rely on a single association, but multiple to recover from few erroneous associations. Second, most of our measures are co-occurrence based, which are typically dominated by the dominant meaning of a word. Finally, for our large scale experiment, we work on ImageNet (Deng *et al.*, 2009) which associates categories to synsets nodes (sets of synonym words representing a meaning) in the WordNet hierarchy (Fellbaum, 1998), which are associated to a specific meaning. We thus can rely on exact calculations in the WordNet-based measures and use all words associated to a synset to improve our word-based co-occurrence statistics. More details can be found in Chapter 3.

Robustness of semantic relatedness measures. Estimating semantic relatedness from linguistic resources is noisy as terms are not equally covered in a corpus, especially as we require co-occurrence statistics of terms.

We examine this problem and suggest several ways to combine measures and explore additional resources in Chapter 3. Combining multiple measures and resources allows compensating specific deficits of a single resource.

Large scale recognition. Scaling recognition to several hundred classes and extracting linguistic knowledge about them leads to several challenges. Scaling to large number of classes typically requires distinguishing between more fine-grained categories, i.e. decreasing the inter-class difference. These fine grained categories are frequently represented by compounds in language (e.g. "chain saw", "red beech ") and/or by specific vocabulary (e.g. "stupa", "calceolaria") with no or very limited occurrence in linguistic corpora. This is especially true for mining relations which require co-occurrence of two words.

We attack these challenges by mining relations of compounds rather than single words and by relying only on the strongest associations which are more robust. We also introduce measures based on text snippets returned by search engines which are better capable of covering the required associations but ensure high precision as shown in Chapter 4.

Visual appearance of novel classes. There is a limit of information which can be transferred via category level associations. The exact visual appearance can never be fully recovered for the novel categories. Additionally the semantic relatedness estimated using linguistic resources is not identical to visual similarity or more precisely the visual descriptor similarity, which we ideally would like to know to optimally transfer information.

To attack this problem we show in Chapter 7 how to integrate few labeled instances in our approach and in this way extend our zero-shot recognition to few-shot learning. Furthermore we additionally exploit the instance similarity to better capture visual appearance of the novel classes.

1.1.2 Script data for activity recognition

While object categories are a well-studied problem in computer vision, recognizing activities is a more recently studied problem. This also holds for extracting semantic similarity between objects represented by noun-phrases versus activities represented as verb phrases from linguistic resources.

Datasets for activity videos and descriptions. While several activity recognition datasets have been proposed recently, most of them focus on single actions. However, we want to explore videos with sequences of multiple activities which we can relate to multi-sentence scripts. To make reasonable connections to the script data we want to have videos which allow robust recognition but operate in a realistic scenario.

We opted for a kitchen scenario as it allows to record data easily ourselves (in contrast to an industrial or medical scenario) and allows to vary the complexity from simple ingredient preparation to complex dishes. In Chapters 5 and 6 we present two datasets with a large number of activities performed by a diverse set of participants using diverse ingredients and tools in a non-scripted fashion. We also recorded and annotated a multi-view human-pose dataset and multi-view object recognition dataset we released publicly (Amin *et al.*, 2013; Susanto *et al.*, 2012). Additionally we collected cooking instructions (script data) with Amazon Mechanical Turk for the same scenarios.

Visual features for fine grained recognition. Recognizing fine details of activities and objects in video is a challenging task. Many activities are based on subtle hand motions with low inter-class variability. The tools used are in most cases occluded by hands and the cooked ingredients are frequently non-ridged and strongly change appearance during the cooking process.

In this thesis we analyze holistic and pose-based features and their combination in Chapter 5. During the course of this thesis we have also shown how to improve pose-representation by using multiple cameras (Amin *et al.*, 2013) and we explored the benefit of multiple depth cameras for object recognition (Susanto *et al.*, 2012).

Variability of composite activities. Composite activities such as preparing a dish are human activities which last over several minutes and consist of many basic-level activities and object interactions. To recognize them in video is challenging due to their high variability and limited available training data.

In Chapter 6 we propose to exploit script data for this task. We decompose the composite activities into attributes of basic-level activities and objects which can be learned easier. To capture the complexity and variability we mine alternative variants from script data which are easy to collect.

1.1.3 Natural language descriptions of visual content

For the third direction of the thesis we shift the focus of category-level knowledge to description specific to a certain video or image instance. While this allows for very attractive tasks, such as learning how to automatically describe images and video with natural language, we can no longer rely on strong statistics of category wide consistent similarities but have to reason about specific instances individually.

Aligned video with textual descriptions. To reason and describe specific activity instances, we need a dataset for learning and evaluating our methods and approaches. We require that sentences are aligned to the specific video snippet they describe.

In Chapter 8 we describe how we collected and aligned the dataset with the help of Amazon Mechanical Turk. We also discuss the detailed properties and released it publicly.

Estimating similarity of activity descriptions. Estimating the similarity between noun phrases is a well-studied problem which we extensively exploit in this thesis. Estimating similarity between different activities is a far less studied problem and turns out to be significantly harder as it requires to understand subtle motion differences, e.g. "*peeling a carrot*" requires very different motion and tools compared to "*peeling a pineapple*.

We approach this challenge by grounding the descriptions in video and combining text-based semantic relatedness with video-based features in Chapter 8.

Automatic verbalization of visual information. Even if we know what is depicted in the visual data, it is not clear how and what to verbalize i.e. describe. Many questions have to be answered. Which words to use, which things are relevant to describe, which level of detail should be chosen, and also which complexity of linguistic expressions should be employed.

We refrain from setting these parameters manually but propose to learn them from example descriptions in Chapter 9. We also thoroughly discuss other related approaches in Chapter 2.

Intermediate semantic representation. As we discussed in the previous paragraphs our approach uses an intermediate semantic representation to describe visual content. The challenge is to find a semantic representation which fulfils the following two aspects: First, it needs to be rich enough to allow producing the desired descriptions. Second, it should be feasible to estimate the semantic representation from the visual content.

To address this challenge we compare three different intermediate representations in Chapter 9. The best representation consists of the activity (verb) and main objects (nouns) and leaves it to the language model to fill in typical adjectives and adverbs. This basically means we only rely on the prior rather than trying to detect adjectives and adverbs as this would likely lead to introducing more noise than being beneficial. To understand the limits of this semantic representation we also run the language generation part of the model on the ground truth semantic representation.

1.2 CONTRIBUTIONS OF THE THESIS

After discussing in the previous section how this thesis contributes to the individual challenges in this field, we summarize the contributions with respect to the three direction of the thesis. In Chapter 2 we relate our contributions to prior work and in Chapter 10, as part of the conclusions of this thesis, we discuss the contributions with respect to the individual chapters.

We also follow two more general goals with this thesis. First, we point out several options for interaction between the visual and linguistic modality, examine how to best exploit each other's strength, and then contribute novel approaches which implement these findings and excel state-of-the-art performance. Second, to foster research looking at the interactions between visual recognition and computational linguistics, we collected, annotated, and publicly released several datasets combined with software and intermediate results.

1.2.1 Contributions to visual knowledge transfer using linguistic semantic relatedness

To integrate linguistic information into visual knowledge transfer we build upon our earlier work (Rohrbach *et al.*, 2010) which allows for unsupervised attribute-based and direct transfer. The essential information for semantic knowledge transfer are the associations between attributes and object categories which we mine from language to replace manual supervision. In this thesis we identify the deficits of the existing approaches and consequently extend the set of semantic relatedness measures to more robust measures. By automatically extending the attribute inventory and combining diverse semantic relatedness measures, we show that we can improve knowledge transfer to the level of manual supervised transfer.

Using automatically mined associations we were the first to scale zero-shot recognition to a truly large scale setting of 200 unseen test classes and we evaluate semantic knowledge transfer to in a supervised setting of 1,000 object classes in ImageNet. We contribute an extensive evaluation of different transfer approaches (attribute, hierarchical, and direct knowledge transfer), and different linguistic semantic relatedness measures, different visual features, and learning approaches. To scale to this large scale setting we suggest essential extensions and technical modification to existing transfer approaches, semantic relatedness measures, and learning algorithms.

To exploit similarities in the unlabeled data distribution we propose our novel *Propagated Semantic Transfer* approach which extends semantic knowledge transfer to the transductive setting by adapting label propagation – previously only used for semi-supervised learning. The approach not only allows for zero-shot recognition but also smoothly integrates labels for novel classes (few-shot recognition). As the local neighborhood structure is essential for exploiting unlabeled data, we propose to map the data into a low dimensional semantic output space using the trained attribute and category models. This significantly improves the neighbourhood structure and final recognition performance. We validate our approach on three challenging datasets for two different applications, namely on *Animals with Attributes* and *ImageNet* for image classification and on our dataset for activity recognition and show significant improvements over related work as well as to baselines using only knowledge transfer or only label propagation.

1.2.2 Contributions to script data for activity recognition

Based on the observation of the unavailability of an appropriate video dataset for activity detection and multi-sentence descriptions we record, annotate, and release two novel activity recognition datasets. The first, *MPII Cooking Activities*, provides a classification and detection benchmark for fine-grained activity recognition on long, challenging cooking sequences consisting of diverse, complex dishes. We evaluate several video descriptors and activity recognition approaches. On the one hand we benchmark the state-of-the-art *dense trajectories* (Wang *et al.*, 2011) and on the other hand, we propose two approaches based on body pose tracks. We also provide an annotated body pose dataset which we extend to multiple cameras in a follow up work (Amin *et al.*, 2013).

The second, *MPII Composite Cooking Activities*, focuses on recognition of entire dishes and contains 256 videos rather than basic-level activities along with independently collected cooking instruction (script data). We contribute an approach to use text-based script data for handling the large variability of composite activity recognition by selecting relevant attributes. We do not only improve performance in the supervised case but also can transfer to unseen composite cooking activities. We achieve this by decomposing composite activities into a flexible attribute representation. We show that using co-occurrence and temporal activity context can help recognizing the challenging basic-level activities.

1.2.3 Contributions to natural language descriptions of visual content

As the basis for our work to analyze and automatically generate descriptions for video instances, we collect the TACoS corpus which contains natural language descriptions for each video with sentence-level alignment composite activity video dataset. The dataset is publicly available and we expect the corpus to be a valuable resource for computational semantics and moreover helpful for a variety of purposes, including video understanding and generation of text from videos which we explore in this thesis.

We provide a *gold-standard dataset* for the evaluation of similarity models for action verbs and phrases. We compute semantic similarity by combining visual and textual relatedness we demonstrate the impact of grounded information provided by video. While the visual similarity models outperform text-based models, the performance of combined models even approaches the upper bound indicated by inter-annotator agreement.

Our novel two-step approach for *video description* firsts learns a mapping from video to an intermediate semantic representation of activities and objects. In the second step our system learns how to translate the semantic representation to a natural language description. It is the first approach for video description which learns both, visual recognition and description, from a parallel corpus of videos, semantic representation, and sentences, rather than relying on retrieval or manually defined templates. Using automatic as well as human evaluation, the proposed approach outperforms several baseline methods inspired by previous work.

Furthermore, our approach is applicable to *image description* and we show that our approach compares favorably to related work on the Pascal-sentence dataset (Farhadi *et al.*, 2010b).

1.3 OUTLINE OF THE THESIS

In this section we summarize the chapters of the thesis and relate them to each other. We also note the respective publications and collaborations with other researches.

This thesis aims to combine aspects from computer vision and computational linguistics. However, we would like to note that the bulk of the thesis is written from a computer vision angle, especially Chapters 3 4, 5, 6, 7, and 9, while Chapter 8 is written from the angle of computational linguistics.

Chapter 2: Related work. This chapter surveys related work which combines visual recognition and computational linguistics with a focus on the three directions of the thesis *Visual knowledge transfer using linguistic semantic relatedness*, *Script data for activity recognition*, and *Natural language descriptions of visual content*. We discuss how these works relate to the approaches and contributions presented in this thesis. A discussion of related work specific to the following chapters is provided within each chapter.

Chapter 3: Combining Language Sources for Knowledge Transfer. In this chapter we introduce the approaches for attribute- and direct similarity-based knowledge transfer for object recognition using semantic relatedness from linguistic resources. They are based on our work in (Rohrbach *et al.*, 2010) and also used in Chapters 4 and 7. However, the focus of this chapter is to improve the robustness of linguistic semantic relatedness to the level of manual defined associations.

The content of this chapter was presented in the First International Workshop on Parts and Attributes (PnA2010) in conjunction with ECCV 2010 with the title

Combining Language Sources and Robust Semantic Relatedness for Attribute-Based Knowledge Transfer (Rohrbach *et al.*, 2012c). Marcus Rohrbach was the lead author of this paper. It is a follow up work on (Rohrbach *et al.*, 2010), which was part of Marcus Rohrbach's Master Thesis and based on a collaboration with the Ubiquitous Knowledge Processing Lab at TU Darmstadt.

Chapter 4: Knowledge Transfer in a Large-Scale Setting. In this chapter we scale the approaches from Chapter 3 to large scale recognition. We examine zero-shot recognition for 200 classes, but also compare semantic knowledge transfer in a supervised setting to standard one-vs-all classification. This chapter examines which visual features, linguistic semantic relatedness, and learning approaches are still applicable at this scale and what has to be adapted.

The content of this chapter corresponds to the CVPR 2011 publication *Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting* (Rohrbach *et al.*, 2011). Marcus Rohrbach was the lead author of this paper.

Chapter 5: Fine Grained Cooking Activity Detection. In this chapter we shift from object detection to activity recognition. We present a large dataset of fine-grained cooking activities which we recorded and annotated. We compare holistic and pose-based activity recognition approaches and examine how they perform in this scenario for activity classification and detection. The dataset and findings are the bases for the following chapters, which integrate it with language resources.

The content of this chapter corresponds to the CVPR 2012 publication *A Database for Fine Grained Activity Detection of Cooking Activities* (Rohrbach *et al.*, 2012a). Marcus Rohrbach was the lead author of this paper, while Sikandar Amin contributed the pose-estimation part. Sikandar Amin's internship at MPI Informatics was co-supervised by Marcus Rohrbach and Mykhaylo Andriluka.

Chapter 6: Script Data for Recognition of Composite Activities. Based on the experiences from Chapter 5, we extend the dataset with recordings of 41 composite cooking tasks. Using script data, which consists of independently collected instruction of these tasks, we can capture the variability of the data better than using a purely vision-based approach, but also recognize the tasks without visual examples. Our approach is inspired by the attribute representation we used for object recognition in Chapters 3 and 4.

The content of this chapter corresponds to the ECCV 2012 publication *Script data for attribute-based recognition of composite activities* (Rohrbach *et al.*, 2012b). The paper is based on a collaboration with the Computation Linguistics Department (CoLi) at Saarland University. Marcus Rohrbach was the lead author of this paper while Michaela Regneri from CoLi contributed the script data collection and extraction.

Chapter 7: Transfer Learning in a Transductive Setting. In this chapter we show how to exploit the visual similarity of images or videos of the novel categories to improve knowledge transfer for object and activity recognition as introduced in Chapters 3, 4, and 6. Our proposed approach also allows integrating a few labeled instances if available, i.e. we show how we extend our knowledge transfer approaches from zero- to few-shot recognition.

The content of this chapter corresponds to the NIPS 2013 publication *Transfer Learning in a Transductive Setting* (Rohrbach *et al.*, 2013a). Marcus Rohrbach was the lead author of this paper.

Chapter 8: Grounding Action Descriptions in Videos. In this chapter we change from category specific information or descriptions to instance specific descriptions. More specifically we collect and align descriptions which describe the videos we introduced in Chapter 6. This allows us to ground descriptions in video and enrich their representation with video features which significantly improves similarity estimation of descriptions.

The content of this chapter corresponds to the TACL 2013 publication *Grounding Action Descriptions in Videos* (Regneri *et al.*, 2013). The lead author of this paper was Michaela Regneri from CoLi. Marcus Rohrbach contributed the vision side of the work and collaborated in combining both modalities.

Chapter 9: Translating Video to Natural Language Descriptions. Based on the corpus presented in the previous chapter, we show how to automatically describe video with natural language sentences. More specifically we learn a visual recognition model of activities and objects and then learn a statistical machine translation model to translate this intermediate representation of activities and objects to natural language descriptions. We also show that this approach is applicable to image description.

The content of this chapter corresponds to the ICCV 2013 publication *Translating video content to natural language descriptions* (Rohrbach *et al.*, 2013b). Marcus Rohrbach was the lead author of this paper. It is based on the continued collaboration with CoLi. Wei Qiu has contributed the implementation for the statistical machine translation as part of his Master Thesis, which was co-supervised by Marcus Rohrbach.

Chapter 10: Conclusions and future perspectives. This chapter concludes the thesis by summarizing the contributions and highlighting their current limitations and possible directions to overcome them. We provide an outlook on our ongoing and future work and discuss future directions for the field of research which brings visual and linguistic modalities together.

RELATED WORK 2

Contents

The work of this thesis lies in between two areas of computer science typically researched and developed independently: computer vision and computational linguistics. This chapter aims to present work in combining both fields focusing on the directions chosen in this thesis.

In this chapter we discuss the most recent developments as well as seminal works and relate them to the contributions of this thesis in the conclusion of each section. The following chapters also discuss related work, but targeted to the respective topic of the respective chapter.

2.1 VISUAL KNOWLEDGE TRANSFER USING LINGUISTIC SEMANTIC RELATEDNESS

The first direction of the thesis is concerned with transferring knowledge using semantic relatedness mined from linguistic knowledge bases and corpora. Consequently we first discuss approaches for visual knowledge transfer (Section 2.1.1) and then examine approaches using linguistic semantic relatedness for visual recognition (Section 2.1.2). We conclude this section with discussing the relations to the work in this thesis (Section 2.1.3).

2.1.1 Visual Knowledge Transfer

Transferring knowledge between visual categories has become an important research direction for scalable recognition. We distinguish the achievements in zero-shot and few-shot recognition, which – despite solving a similar challenge – are typically approached differently as discussed below.

2.1.1.1 *Zero-shot recognition*

The challenge of zero-shot recognition is to recognize unseen visual object categories, i.e. without any training exemplars of the unseen class. This requires additional semantic information to identify the unseen category and relate it to known visual concepts. This challenge has recently received increased interest for object (e.g. Lampert *et al.*, 2009; Mensink *et al.*, 2012; Frome *et al.*, 2013) and activity recognition (e.g. Fu *et al.*, 2013; Guadarrama *et al.*, 2013a).

Lampert *et al.* (2009, 2013) propose a probabilistic framework to recognize unseen classes based on an intermediate level of semantic attributes. Semantic attributes are concepts, which are properties with human understandable meaning, such as parts, colors, or appearance. The idea is that they are valid across categories and thus allow transferring knowledge. Following this idea, Lampert *et al.* propose to integrate human knowledge in the recognition process of unseen classes by providing category-level class-attribute associations. In their Direct Attribute Prediction (DAP) model they learn attribute classifiers from the known classes. In their Indirect Attribute Prediction (IAP) model, they build attribute classifiers by combining the probabilities of all associated known classes. For recognizing an instance of an unseen class they combine the valid attributes of the unseen class.

In our earlier work (Rohrbach *et al.*, 2010) we introduce zero-shot recognition by directly exploiting the similarity between classes, i.e. to recognize an unseen class we combine the classifiers of the most similar known classes. While IAP from Lampert *et al.* combine object classifiers trained on known classes according to attribute activations, our direct similarity model combines object classifiers according to similarity of object classes, which is frequently easier and more robust to estimate than class-attribute associations.

An alternative direction is to phrase zero-shot learning as an embedding problem. Learning an embedding function, i.e. a transformation, from the image space to a space where the unseen classes are known, allows recognizing unseen visual categories according to the similarity in the second space. In the task of handwritten character recognition Larochelle *et al.* (2008) learn the transformation from a handwritten character to a typed representation. To recognize unseen classes they require a typed character pixel representation for recognizing an unseen handwritten character. During training the transformation for known classes is learned. Novel categories can then be identified based on the similarity in the transformed space. While Larochelle *et al.* rely on an alternative visual representation, Socher *et al.* (2013) and Frome *et al.* (2013) learn the embedding into a linguistic word space which models the semantic similarity between concepts. Socher *et al.* (2013) learn a neural network model, which maps linguistics words into a 50-dimensional space where the distance represents the semantic similarity between words (Huang *et al.*, 2012). The representation from Huang *et al.* learns local word context as well as document context from Wikipedia. For the images the authors learn an embedding into this word-space, which allows them to recognize novel classes by projecting novel images into this semantic space and to find the closest prototypical linguistic word in this space. Frome *et al.* (2013) follow a very similar idea but scale recognition to several thousand unseen classes. To construct a few hundred dimensional semantic space they use the conceptually similar skip-gram model (Mikolov *et al.*, 2013) leaned on Wikipedia. Akata *et al.* (2013) embed the class labels in an attribute space instead of embedding images into another space. For this they use a modified version of the ranking objective function from Weston *et al.* (2010, WSABIE). For zero-shot learning the label-embeddings of the unseen classes cannot be learned and are thus set to the fixed prior information i.e. class-attribute association.

The evaluation of zero-shot recognition is typically restricted to novel classes. However, this setup is somewhat artificial as it assumes that there are no known instances among the test instances. While we were the first in (Rohrbach *et al.*, 2010) to show the increased difficulty when instances of the known classes are added as distractors to the recognition evaluation, Socher *et al.* (2013) and Frome *et al.* (2013) aim to recognize both, known and novel classes. Socher *et al.* (2013) employ novelty detection (see below for a more detailed discussion on this) to distinguish between known and novel classes. Frome *et al.* (2013) do not rely on novelty detection to recognize novel classes, but show that their image embedding in the semantic space is capable to achieve state-of-the-art performance for the known classes and novel classes in a large scale setting.

Palatucci *et al.* (2009) analyze zero-shot learning theoretically and are able to estimate the probability of correctly predicting an unseen class, assuming a Probably Approximately Correct learner (PAC learner) for estimating the intermediate representation and nearest neighbor classification of the unseen classes. They validate their approach on the task of neural decoding of novel thoughts. Semantic relatedness is mined using co-occurrence in the Google Trillion-Word-Corpus and, alternatively, by asking humans to estimate attributes for the concept words of interest. The latter provides much more reliable predictions but also significantly larger supervision.

Additionally to transferring knowledge one can exploit the unlabeled instances to improve recognition, assuming a transductive setting. For zero-shot recognition, Fu *et al.* (2012, 2013) first project instances of novel classes into a semantic attribute space based on predefined class-attribute associations. To exploit the test-data distribution they perform a single round of self-training by averaging over the k-nearest neighbors. Averaging is done on a discriminatively trained, latent attribute representation. The final prediction is made by finding the closest prototype, which is also projected in the latent attribute representation in the same way. In their experiments the prototype is provided by human annotations.

Zero-shot recognition has some similarities to novelty detection. While novelty detection separates instances of unknown categories from instances of known categories, without identifying the unknown categories, zero shot recognition is able to identify to which unseen class a certain instance belongs, however, zero-shot recognition requires additional external information to do so. Novelty or outlier detection can be formulated e.g. as a one-class problem and modeled by Gaussian Processes (Kemmler *et al.*, 2010), by extending the multi-class classification problem with outlier rejection (Tax and Duin, 2008), or by using multi-class null space projections (Bodesheim *et al.*, 2013). As shown by Socher *et al.* (2013) one can recognize known and unseen classes by combining outlier detection with zero-shot recognition, i.e. applying zero-shot recognition only to the outliers. Given the still limited absolute performance of both directions, an integrated approach solving both problems together might be more promising as suggested by Frome *et al.* (2013). Related to novelty detection is the open set problem introduced by Scheirer *et al.* (2012). Here also the number of unseen/unknown classes is not known. In case of only one known and one unknown class the open set problem is equivalent to the 1-class problem (Schölkopf *et al.*, 2001).

2.1.1.2 *Few-shot recognition*

Few-shot recognition tries to learn a *novel* object class model from a single or only very few labeled instances. There are two ways this problem is typically approached. On the one hand transferring the information from known classes to novel classes and, on the other hand, propagating knowledge from the few instances to unlabeled instances in a transductive setting using the structure of the data.

First, we discuss approaches which transfer knowledge from known classes. Thrun (1996) explores several approaches to exploit the data from known classes: learning a representation which helps to distinguish similar and dissimilar concepts; metric learning, which learns a distance function between instances; learning with hints, which learns a neural network for the novel categories in parallel with the known classes (the hints); and a neural network to learn a distance function and tangents on the set of known classes, which uses the explanation-based neural network learning algorithm (EBNN) and the Tangent-prop algorithm (Simard *et al.*, 1991). In an experiment to distinguish 2 novel object classes with the help of 5 known classes, Thrun shows that all approaches of knowledge transfer, expect his implementation of learning with hints, show significant improvements compared to baselines only relying on the instances of novel classes.

Bart and Ullman (2005b) propose to recognize a novel class by a single exemplar is similar to Thrun's idea of learning a representation. They retrieve the most similar known classes and use their respective stacked classifiers scores as new representation. Classification is achieved by nearest neighbor on the stacked classifier score vector. Recent approaches also use attributes as an improved representation. Sharmanska *et al.* (2012) augment the attribute representation from Lampert *et al.* (2009) by non-semantic features, which are learned using an autoencoder model on the few available samples, classification is also performed with nearest neighbor.

(Fink, 2004) pick up the idea of metric learning on known classes for the extreme case of only a single training sample of the novel class. Fink uses a kernel based metric learning for one-shot character recognition. Mensink *et al.* (2012) scale metric learning to 200 novel classes of ImageNet. They use a Nearest Class Mean classifier, which they show to perform better than k-Nearest Neighbor (kNN) classifier and close to SVM for the many instance setting.

A different direction is followed by Li *et al.* (2006) who adapt the parameters in a Bayesian formulation. Using a generative model they transfer knowledge from prior models of the known classes adapting parameters according to one or few samples from the novel classes. This results in a posterior probabilistic model for each class. On their dataset of 101 image classes (Caltech 101) they show that their Bayesian approach outperforms Maximum Likelihood (ML) and Maximum A Posteriori (MAP). Stark *et al.* (2009) also transfer mode parameters, but in a explicit way of shape contour parameters, in a part-based shape model for visual object categories. Their model allows for partial or full transfer but what to transfer has to be manually specified.

Second, we discuss approaches in a transductive setting, which exploit the unlabeled instances in addition to the few labeled instances. This can be achieved using semi-supervised learning techniques (Chapelle *et al.*, 2006). An extensive discussion and categorization of semi-supervised approaches can be found in (Chapelle *et al.*, 2006) and (Ebert, 2012). For semi-supervised recognition approaches to work, the data either has to cluster in a way that instances of the same cluster are likely of the same class (cluster assumption) or the data lies on a low-dimensional manifold (manifold assumption). The cluster assumption is the base for many widely used approaches including transductive SVM (TSVM) used for text classification (Joachims, 1999) but also for visual recognition. In this thesis we rely on graph based methods which use the manifold assumption. The most common approach for graph-based methods is *label propagation*. Based on a graph between all instances which encodes their similarity, labels are propagated from labeled instances to unlabeled instances (Zhu *et al.*, 2003). While Zhu *et al.* used fixed original labels, Zhou *et al.* (2004) allow a change of the original labels by using a normalized graph Laplacian. Label propagation has been used in many applications, including part-of-speech tagging (Subramanya *et al.*, 2010), image classification (Ebert *et al.*, 2010), and activity recognition (Stikic *et al.*, 2011).

Fu *et al.* (2012, 2013) exploit knowledge transfer from known classes and exploit unlabeled instances. For few-shot learning they first learn a latent attribute representation, which is initialized with manually defined attributes. In a second step they learn a classifier on the attribute space with few labels. Compared to manually defined attributes their latent attributes are beneficial if the number of attributes is small. However, on the full set of manually defined attributes of the AwA and the multi-modal USAA dataset (Fu *et al.*, 2012) latent attribute representation performs worse. Nevertheless, both attribute representations significantly outperform a direct variant which does not exploit the unlabeled instances.

2.1.2 Linguistic semantic relatedness for object recognition

Visual attributes are human nameable properties, which are typically shared across object classes. Being nameable they lend themselves for a connection point to linguistic knowledge. Delezoide *et al.* (2008) use web co-occurrence statistics to associate visual object categories with context attributes. Given a segmentation between foreground objects and background scene (context attributes) they separately learn a foreground and background model. In a Bayesian formulation the conditional probability between objects (animals in their case) and context (e.g. *forest*, *savanna*, or *ice*) is set according to web co-occurrence statistics. They use the dice coefficient (Dice, 1945), which has shown to be an appropriate measure of semantic relatedness for web co-occurrence statistics (Kilgarriff and Grefenstette, 2003). They distinguish four different variants: standard web search, restricting counts to words appearing close to each other in the web, and Flickr Text and Flickr Tags as corpus. They show that the restriction to close by words is beneficial and that Flickr Text performs best among the four variants. However, in a supervised setting, learning the associations between objects and context attributes from image data outperforms all the web-corpus based approaches.

Tzoukermann *et al.* (2011) examine different measures to estimate the similarity between activities and tools and recognize them in art and crafts TV shows. To estimate similarity they use Wikipedia activity articles and check which nouns are present; the nouns are verified according to WordNet. They also used the hit-count based Google similarity distance (Cilibrasi and Vitanyi, 2007) and adapted it, similar to Delezoide *et al.* (2008), to nearby words.

One challenge in using attributes is to define them. While human labels or definitions are expensive, information mined from linguistic sources can be obtained automatically. In Rohrbach *et al.* (2010) we propose to mine attribute names from part-of relations in the WordNet hierarchy. An alternative to mine attributes would be to use WordNet's synset definitions as proposed by Russakovsky and Fei-Fei (2010). Berg *et al.* (2010b) use multiple instance learning to identify and localize attributes from web shopping images with captions, here the attributes are not shared across classes. Duan *et al.* (2012) discover and localize discriminative attributes for fine-grained recognition. While the discovery step is performed automatically using a latent CRF, human interaction is required to name the semantic attributes.

Aytar *et al.* (2008) show the benefit of web co-occurrence statistics for retrieving videos of unseen concepts (zero-shot). For this they expand all words in query phrase with synonyms obtained from the WordNet; then they use classifiers of similar concepts according to pointwise mutual information using web mined data (PMI-IR) (Turney, 2001). This formulation is similar to the *log* of the dice coefficient (Dice, 1945). They show that this approach yields better results than the WordNet-based Lin measure (Lin, 1998), performs similar to visual co-occurrence and close to supervised trained concept classifiers.

An alternative to the above described statistical approaches for mining information from textual corpora is to directly extract semantic information from text. Wang *et al.* (2009b) leverage descriptions of butterflies to classify unseen butterfly categories. For each category they rely on a single description which is automatically part-of-speech tagged. Afterwards a predefined template is filled to determine attributes including wing color and spot color using key words. By recognizing spot and wing color, Wang *et al.* show that their approach can distinguish 10 categories of pre-segmented butterflies nearly as good as using the ground truth template. Elhoseiny *et al.* (2013) also exploit category level text descriptions to identify relevant attributes. However, in contrast ot Wang *et al.* and similar to the zero-shot approaches described in Section 2.1.1.1, Elhoseiny *et al.* (2013) use classifiers trained on the known classes. Representing the text descriptions with tf*idf (term frequency times inverse document frequency) vectors for relevant encyclopedic entries, they compare a regression, a domain adaptation, and a newly proposed constrained optimization formulation to learn a function from the textual vector to the visual classifier space. On two fine-grained visual recognition datasets, CU200 Birds (Welinder *et al.*, 2010) and Oxford Flower-102 (Nilsback and Zisserman, 2008), they show the benefit of their constraint optimization approach. Approaches using text descriptions for activity recognition can be found in Section 2.2.2.

Another option to include linguistic knowledge in the visual recognition pipeline is to learn an embedding function from images to a semantic word space as done for zero-shot recognition by Socher *et al.* (2013) and Frome *et al.* (2013), see Section 2.1.1.1 for more details.

There is also a line of work which jointly models visual items with linguistic words. This includes Barnard *et al.* (2003) who examine different multi-modal probability distributions to jointly model weakly supervised tags and images regions and Li *et al.* (2009) who jointly model image scenes and tags. However, in this line of work words are merely used as labels or ids without exploiting (linguistic) semantic information between them. This also holds for approaches that build a visual hierarchy from images (Li *et al.*, 2010c), or learn a visual ontology (Chen *et al.*, 2013).

2.1.3 Relations to our work

In Chapters 3, 4, and 7 we build upon the attribute-based knowledge transfer approach from Lampert *et al.* (2009) and on the direct-similarity based knowledge transfer approach we introduced in (Rohrbach *et al.*, 2010). In contrast to Lampert *et al.* (2009) and similar to (Rohrbach *et al.*, 2010) we rely on automatically mined linguistic semantic relatedness rather than manual supervision.

Similar to Delezoide *et al.* (2008) and Tzoukermann *et al.* (2011) we restrict co-occurrence web search to nearby words in Chapter 3. While this yields improved performance, we show that it is crucial to combine different linguistic measures to obtain robust semantic relatedness. We also explore the benefit of an extended attribute inventory, but, in contrast to creating latent attributes (e.g. Sharmanska *et al.*, 2012), we rely on clustering of attributes. This retains the semantic property of the attributes, which is necessary for incorporating linguistic knowledge.

Mensink *et al.* (2012) and Frome *et al.* (2013) show impressive results for scaling zero-shot recognition to a large number of classes by using metric learning and embedding in a semantic word space, respectively. In Chapter 4 we evaluate different semantic knowledge transfer approaches in a similar setting. When compared in the same scenario of 200 unseen classes of the ILSVRC10 (Berg *et al.*, 2010a) our approaches perform on-par with Mensink *et al.* (2012) and outperform Frome *et al.* (2013) as shown in (Frome *et al.*, 2013).

In Chapter 7 we present our novel Propagated Semantic Transfer approach that extends our zero-learning approaches to few-shot recognition. While Mensink *et al.* (2012) use metric learning to improve nearest neighbor quality in this setting, we propose to exploit stacked attribute or object classifiers. Furthermore, Propagated Semantic Transfer also exploits unlabeled instances of the novel classes by using label-propagation where again the nearest neighbor quality is essential to build a good graph structure as shown by Ebert *et al.* (2010).

Apart from our Propagated Semantic Transfer, Fu *et al.* (2013) present the only work which exploits knowledge transfer and unlabeled data for zero- and few-shot recognition. While our Propagated Semantic Transfer approach can seamlessly integrate labels for novel classes, Fu *et al.* handle zero-shot and few-shot slightly differently and we show that our approach performs better on the AwA dataset (Lampert *et al.*, 2009).

2.2 SCRIPT DATA FOR ACTIVITY RECOGNITION

In this section we shortly review recent approaches to activity recognition (Section 2.2.1) and then discuss works which use textual information for improved recognition of activities (Section 2.2.2). Compared to the previous section, the focus shifts from object to activity recognition, and, on the linguistic side, from mainly co-occurrence based approaches to approaches exploiting textual descriptions. However, we note that this distinction is not a clear cut and some works could be grouped to both sections, e.g. Elhoseiny *et al.* (2013) discussed in the previous section use text descriptions for object recognition, and many works which focus on activity recognition also exploit object detectors. In these cases we choose the section with respect to the more prominent aspects.

Video activity recognition datasets which have descriptions associated with them are discussed in Section 2.3.1 and an overview of the different video activity recognition datasets without textual descriptions can be found in Section 5.2.1 and Table 5.1.

2.2.1 Advances in activity recognition

Activity recognition for still images has been advanced e.g. by jointly modeling people and objects (Yao and Li, 2012) or scenes and objects (Li and Li, 2007). In the following we focus on recognizing activities in video, distinguishing three aspects:

Holistic features for activity recognition, exploiting body pose, and modelling the temporal structure of activities.

2.2.1.1 *Holistic features for activity recognition*

To create a discriminative feature representation of a video, many approaches first detect space-time interest points (Chakraborty *et al.*, 2011; Laptev, 2005) or sample them densely (Wang *et al.*, 2009a) and then extract diverse descriptors in the image-time volume, such as histograms of oriented gradients (HOG) and flow (HOF) (Laptev *et al.*, 2008) or local trinary patterns (Yeffet and Wolf, 2009).

Messing *et al.* (2009) found improved performance by tracking Harris3D interest points (Laptev, 2005). The state-of-the-art *dense trajectory* approach from Wang *et al.* (2011, 2013) uses this idea: It tracks dense feature points and extract strong video features around these tracks, namely Histogram of Oriented Gradients (HOG), Histogram of Oriented Flow (HOF), and Motion Boundary Histograms (MBH, Dalal *et al.*, 2006). They report state-of-the art results on KTH (Schuldt *et al.*, 2004), UCF YouTube (Liu *et al.*, 2009), Hollywood2 (Marszalek *et al.*, 2009), and UCF sports (Rodriguez *et al.*, 2008). Recently, Wang and Schmid (2013) improved their dense trajectory approach by removing background flow and by ensuring that detected humans do not contribute to the background motion estimation. Additionally they replace the BoW encoding with Fisher vectors.

An alternative to manually defined features is deep learning with convolutional neural networks, which has shown significant performance gains for object classification (Krizhevsky *et al.*, 2012) and object detection (Girshick *et al.*, 2013). Baccouche *et al.* (2011) automatically learn spatio-temporal features by extending convolutional neural networks to 3D and show encouraging results on the simple KTH dataset. Taylor *et al.* (2010) employ a convolutional gated Restricted Boltzmann Machine to learn spatio-temporal features on consecutive video frames pairs. On top they add a deep architecture performing 3-D convolution and several pooling steps. The top layer consists of a fully connected net, trained to classify the activities. Le *et al.* (2011) focus on learning the low level features and use standard k-means vector quantization and χ^2-kernel SVMs on top. For learning the low level features they adapt the Independent Subspace Analysis with convolution and stacking. Their network uses 10 consecutive frames, which allows them to capture features such as moving edges.

Given the difficulty of classifying activities in realistic video, assessing the quality how humans perform a certain activity is hardly ever done. An exception is Bettadapura *et al.* (2013), who aim to assess how well surgeon activities are executed with respect to seven metrics. They aim to automatically estimate expert assessments of time and motion of the activity or instrument handling.

2.2.1.2 *Body pose for activity recognition*

Human body poses and their motion frequently characterize human activities and interactions. This has been exploited in Microsoft's Kinect, which uses human pose as a game controller but relies on a depth sensor to recognize human pose (Shotton *et al.*, 2011). Sung *et al.* (2011) use this depth information to estimate pose and distinguish 12 activities. Pose-based activity recognition appears to work particularly well for images with little clutter and fully visible people as in the gesture dataset from Singh and Nevatia (2011). Estimates of people poses were also used as auxiliary information for activity recognition in single images (Yang *et al.*, 2010).

Raptis and Sigal (2013) model an activity video as sequence of four temporally discriminative keyframes. In each key frame they extract poselets (Bourdev and Malik, 2009), adapted to include motion information and learned with weak supervision. The approach shows state-of-the-art performance on the UT Interaction dataset (Ryoo and Aggarwal, 2009) and implicitly localizes the activities.

Jhuang *et al.* (2013) study the benefits of pose estimation for activity recognition on a subset of the HMDB dataset (Kuehne *et al.*, 2011). They show that ground truth pose, estimated over time can significantly outperform the holistic dense-trajectory features (Wang *et al.*, 2013); this is also true for estimated pose using (Yang and Ramanan, 2013) but only on a subset where the full body is visible.

2.2.1.3 *Temporal structure for activity recognition*

A simple temporal structure is encoded in the template-based Action MACH from Rodriguez *et al.* (2008) who generalize the image MACH filter to the temporal dimension. Brendel and Todorovic (2011) model temporal and spatial structure by segmenting the space-temporal volume, which is matched to the trained activity model.

As the exact temporal activity structure is typical latent, Niebles *et al.* (2010) model activities as a temporal composition of primitive actions and discriminatively learn such models. Similar to the parts-modeling in DPM (Felzenszwalb *et al.*, 2010) for object recognition, temporal segments are learned in a data-driven manner, but varied over time rather than image location. While Niebles *et al.* fix anchor points and the length of the temporal segments before training, Tang *et al.* (2012) learn all parameters from data using a variable-duration hidden Markov model and show that this improves recognition on several activity recognition benchmarks.

One challenge in activity recognition is to combine the features from different cues, such as object, video, and scene descriptors. Tang *et al.* (2013) approach this problem by learning an AND/OR graph structure, which combines different features at its nodes. They propose an efficient inference technique and show quantitatively that this approach outperforms conventional multiple kernel learning. Gupta *et al.* (2009) also employ an AND/OR-graph, but use it to define co-occurring and consecutive actions in sport videos.

2.2.2 Natural language text for activity recognition

Ni *et al.* (2011) learn activity classifiers with weak supervision. On a large set of 6.5 million YouTube videos they extract nouns and verbs on the associated user provided descriptions using part-of-speech tagging with OpenNLP[1]. After restricting objects to "physical entities" and verbs to activities according to the WordNet hierarchy, a set of about 30,000 event categories remained. After training the event categories they select a subset which works reliable on a validation set. When stacking the classifier outputs with the original features they show improved performance on two activity benchmarks.

Teo *et al.* (2012) also estimate the relation between activities and object in a video from a linguistic corpus. More specifically, to recognize activities, Teo *et al.* mine co-occurrence of activities and objects in the newswire Gigaword Corpus. They leverage synonyms and hyponyms of the activity and object terms from WordNet to increase coverage. This allows them to leverage object detections, which are tools in their kitchen scenario, for recognizing the activities. They compare an unsupervised EM approach without activity labels, as well as a semi-supervised and a supervised model which exploits videos with activity labels. In all cases they significantly improve when including object detector which are linked to activities based on linguistic co-occurrence statistics.

A similar idea is pursued by Motwani and Mooney (2012), who mine verbs from descriptions of the video snippets in the MSVD dataset (Chen and Dolan, 2011). First they cluster the verbs in the descriptions to discover activity classes and produce a labeled training set. The clustered labels are than used to learn activity classifiers using spatio-temporal features. Furthermore they exploit object detectors by mining the correlations between activity verbs and object categories. They generate dependency parse trees with the Stanford Parser (de Marneffe *et al.*, 2006) for mining verbs and their direct object. They show that combining activity recognition with the object detectors is beneficial. Their dependency parsing approach to estimate verb-object associations shows also better performance compared to only looking at part-of-speech tags as in (Ni *et al.*, 2011).

To distinguish between seven scenes of video snippets, such as *actions*, *close ups*, and *traffic*, Zhang *et al.* (2011) rely only on text descriptions collected for each video. To estimate the importance of a term for a scene, they use tf*idf and MAP estimates (Hazen *et al.*, 2007). Combining yields close to perfect (98%) recognition accuracy on their dataset.

Ramanathan *et al.* (2013) infer actions (e.g. *kiss*, *blowing candles*) and roles (e.g. *broom*, *priest*, *birthday child*) for YouTube videos with weakly labeled captions. As captions frequently do not contain the precise action and role, Ramanathan *et al.* mine a large number of YouTube descriptions and propose to use a topic model to estimate the semantic relatedness between an action/role and a description. In order to assign actions and roles to spatiotemporally-localized human tracklets they use Posterior Regularization and require that semantically related actions and roles are associated. Their evaluation shows that their semantic relatedness measure based on YouTube description increases performance for the case when descriptions are available during training, and show significant improvements in a scenario where descriptions are available during testing as well.

[1] http://opennlp.apache.org/

Socher and Fei-Fei (2010) propose a semi-supervised approach which jointly annotates and segments images exploiting a large text corpus of textual description and using kernelized canonical correlation analysis (Hardoon *et al.*, 2004), which maps visual to textual words by projecting them into a latent meaning space.

2.2.3 Relations to our work

Most of the activity recognition approaches and datasets have been evaluated on full-body motion or challenging web or movie datasets but not on fine-grained motions with low inter-class variability. We therefore evaluate the holistic dense trajectory approach from Wang *et al.* (2011) and a pose-based approach on our MPII Cooking dataset in Chapter 5. Our pose-based approach encodes trajectories of body joints using features motivated from the sensor-based activity recognition community (Zinnen *et al.*, 2009). The features are also similar to features used in (Jhuang *et al.*, 2013) for encoding body pose. In contrast to the conclusions from Jhuang *et al.* (2013), we find that body-pose features alone perform clearly below the holistic dense trajectory features. This is most likely due to the fact that we need to distinguish very subtle differences in pose and joint movement to recognize our cooking activities. In (Amin *et al.*, 2013) we improve the important and challenging hand position and leverage multiple cameras to handle self-occlusion.

Learned features are an alternative to these hand-crafted features for activity recognition. While Le *et al.* (2011) show slight improvements over the state-of-the-art it is likely that significant performance jumps, as realized for object recognition, are still ahead for activity recognition as the unavailability of very large activity recognition datasets seems to limit the potential of convolutional neural networks. While this is out of scope of this thesis we believe this is an interesting direction for future work as discussed in Chapter 10.

In Chapter 6 we exploit cooking instructions (script data) to extract which activities, tools, and ingredients are relevant for a certain dish (composite activity). For this we compare co-occurrence statistics with tf*idf, which has also been used by Zhang *et al.* (2011) and Elhoseiny *et al.* (2013) to extract relevant concepts for video scene and object recognition. We find that tf*idf better discriminates different dishes and improves performance in most cases.

While the temporal structure seems an important component to recognize activities, so far mainly the short term structure of short video clips has been explored (e.g. Gupta *et al.*, 2009; Brendel and Todorovic, 2011; Tang *et al.*, 2012). In Chapter 6 we exploit temporal co-occurrence and context of short actions and their participating object. For long term composite activities we only rely on occurrence statistics of its components due to the high variability in temporal order. Nevertheless, we believe that the temporal structure of scripts (Regneri *et al.*, 2010) might form a good prior for the temporal structure of videos and vise-versa as discussed in the future work section in Chapter 10.

2.3 natural language descriptions of visual content

In this section we present parallel corpora which align visual information with natural language descriptions (Section 2.3.1). These corpora allow, on the one hand, studying how to ground linguistic expression in images and videos (related work is discussed in Section 2.3.2) and, on the other hand, learning how to generate natural language descriptions of images and video (Section 2.3.3). In Section 2.3.4 we discuss how our contributions are related to the discussed prior work.

2.3.1 Parallel corpora of visual content and descriptions

In the following we give a short overview of datasets which contain images or videos paired with natural language descriptions.

2.3.1.1 *Images with descriptions*

UIUC Pascal Sentence dataset (Rashtchian *et al.*, 2010; Farhadi *et al.*, 2010b). The dataset consists of 1,000 images from the PASCAL Visual Object Classes (VOC) challenge (Everingham *et al.*, 2009) paired with five descriptive sentences collected with Amazon's Mechanical Turk (AMT). The 1,000 images are sampled equally from 20 classes in the VOC challenge. The difficulty of this dataset is the high variability of the visual data and the diversity of descriptions. Compared to its complexity the dataset is rather small.

SBU Captioned Photo dataset (Ordonez *et al.*, 2011). This dataset is a collection of 1 million photographs from Flickr with associated captions which users provided with their photographs. The photographs are selected so that their caption contains a minimum number of words, as well as the 2 terms used for querying the image and a prepositional word in the hope that descriptions contain spatial relations. Despite this filtering, descriptions tend to be very noisy; one reason for this is that users caption their photos frequently with additional information not necessarily visible in the image. For this corpus Kuznetsova *et al.* (2013) improve the linguistic fluency and increase the consistency between the text and visual content using visual object classifiers.

Abstract Scenes dataset (Zitnick and Parikh, 2013). The dataset consists of 1,002 scenes depicted with clip arts with associated descriptions. The dataset was created to understand the semantics within images as well as between images and sentence descriptions while removing the challenge of visual recognition. All 1,002 scenes are similar in that they are outdoors with same background and consist of 2 children with different poses and facial expressions as well as 56 different objects, including trees, toys, food, and animals. To create the dataset Zitnick and Parikh asked subjects on AMT to generate clip art images for a scene described by one or two sentences. For each scene 10 different clip arts were created by different subjects and every scene is afterwards described with 6 different sentences, again by different subjects.

2.3.1.2 *Videos with descriptions*

Movie scripts, closed captions, and audio descriptions. Movie or television productions frequently contain associated linguistic information. There are movie scripts which are either pre- or post-production screenplays; subtitles/closed captions contain the transcribed spoken language; and audio descriptions describe the visual content and what is happening for visually impaired people. All of them provide only weak supervision, because closed captions frequently are not about what is visible and movie scripts as well as audio descriptions are not well aligned with video. Close captions and movie scripts have so far only been used for retrieving weak labels for activities or during test time as additional information. Audio descriptions have to our knowledge not been used in computer vision. Gupta and Mooney (2010) use closed captions to retrieve and learn activity classifiers in sport videos where the closed captions provide a description of what is happening. Sapp *et al.* (2011a) use close captions of arts and crafts TV shows for activity and tool recognition. Laptev *et al.* (2008) and Cour *et al.* (2008) retrieved activities from movie scripts, focusing on the alignment the video to the scripts with the help of subtitles and Bojanowski *et al.* (2013) use movie scripts at test time for improved activity and actor recognition.

Microsoft Video Description corpus (MSVD, Chen and Dolan, 2011). MSVD consist of 2,089 video segments and 122K multilingual short descriptions of one sentence length; the majority of sentences, 85K, are in English. The videos are segments of YouTube videos showing single activities of usually 10 second or less. The corpus was collected with AMT, aimed for linguistic translation and paraphrasing tasks independent of the video. In addition to this original intention the English part of the corpus has been used for video description (Krishnamoorthy *et al.*, 2013; Guadarrama *et al.*, 2013a).

YouCook (Das *et al.*, 2013b). The YouCook dataset focuses on longer composite activity videos of several minutes. The 88 videos downloaded from YouTube depict 6 different cooking styles. Each video has several multi-sentence descriptions of at least three sentences.

TRECVID Multimedia Event Recounting (MER). MER is a sub-challenge of the TRECVID challenge (Over *et al.*, 2012). The 2012 challenge required the participants to provide a textual description for 30 videos of 6 different categories. Unfortunately the dataset is not publicly available, only participants have temporally access to it with strong restrictions. Das *et al.* (2013b) used the dataset and collect multi-sentence ground truth descriptions.

2.3.2 Grounding linguistic expressions in images or videos

Grounding natural language means finding a mapping or reference of linguistic entities to the external world. While the focus of this section is automatically grounding linguistic expressions in images and video, related research can be found in psychology and robotics. Of psychological interest is to understand how humans represent meaning and language (Pecher and Zwaan, 2005; Glenberg, 2002) and relate it to visual and other sensory input as well as understanding the process of language acquisition (Howell *et al.*, 2005). In robotics the challenge is to relate natural language to the robots internal representation and actions. Guadarrama *et al.* (2013b) present a robot that interprets natural language instructions referring to commands (verbs) which it consequently executes, perceived objects (nouns), and their spatial relations (prepositions). In the following we discuss works which ground language in visual content, first noun phrases and adjectives, and then verb phrases and sentences.

2.3.2.1 *Grounding noun phrases and adjectives*

Leong and Mihalcea (2011) exploit image similarity to create improved word-similarity models. Using the annotated images in ImageNet they compute image similarity based on Sift bag-of-word histograms. They obtain best results when they combine visual similarity with text-based metrics (Lin, 1998; Gabrilovich and Markovitch, 2007). Similarly, Bruni *et al.* (2011) create a visually grounded semantic model which concatenates a visual and linguistic feature. In contrast to Leong and Mihalcea, Bruni *et al.* use the ESP game data set (von Ahn and Dabbish, 2004) as labeled image source and use the more structured text-based model from Baroni and Lenci (2010). Bruni *et al.*'s improvements for using image data are not as pronounced as in (Leong and Mihalcea, 2011), most likely due to the more noisy labeled visual data.

A related line of work tries to determine how well words can be grounded in images. Barnard and Yanai (2006) rate the "visualness" of language entities using mutual information between keywords and image regions. To estimate whether nouns and adjectives are visually related, Boiy *et al.* (2008) compare statistics of visual descriptions, e.g. of flowers or paintings, to statistics of general corpora which typically do not contain visual information. Motivated by the large number of weakly supervised captioned images, Dodge *et al.* (2012) show that given a large dataset of 48k images with captions they can mine visual nouns and adjectives.

Steyvers (2010), Silberer and Lapata (2012) avoid to directly extract information from visual channels but rely on a proxy, namely *feature norms* (McRae *et al.*, 2005) which represent objects as attribute vectors. The object-attribute associations are based on human judgments, who were asked to list important attributes for a given object. This is similar to the data presented in Osherson *et al.* (1991) and used by Lampert *et al.* (2009) to recognize unseen image categories, see Chapter 3. Silberer and Lapata use these feature norms as a proxy for perceptual information and show that a topic model, which discovers a latent representation, improves over simple concatenation of linguistic and perceptual modality. It would be interesting to see if such a model can also be applied to information extracted from images or videos.

Bergsma and Van Durme (2011) aim to build a bilingual lexicon, i.e. learn the translation between words of different languages, with the help of image similarity. This is achieved by using queries of web images using words from the different languages and then estimate word translation based on image similarity.

2.3.2.2 *Grounding verb phrases and sentences*

Zitnick *et al.* (2013) aim to understand the relation between visual scenes, sentence descriptions, and semantic meaning. On their Abstract Scenes Datasets (Zitnick and Parikh, 2013), which contains clip art images and corresponding sentence descriptions, Zitnick *et al.* extract one or more subject-verb-object triples from the descriptive sentence. This representation is similar to Farhadi *et al.* (2010b) who extract such triples from real images, and similar to Kulkarni *et al.* (2011) who relate and describe all objects pairs in an image. Zitnick *et al.* use the triples as basis to analyze spatial relations between subject and object in the clip art images. It also allows them to retrieve different verbalization for the same object as well as typical facial expression and poses for verb phrases. They also show that they can generate novel scenes for a new description based on a CRF which models all objects and relations between them. Non-realistic image data is also used by Orkin and Roy (2009) who employ a computer game in virtual restaurant to learn how to ground chat dialogues in mouse clicks. The resulting data has also been used to model determiner meaning (Reckman *et al.*, 2011). It remains unclear how these computer graphics generalizes to real images and videos.

Mathe *et al.* (2008) propose an approach to estimate the semantics of motion verbs in form of 2d spatial displacement features. From 8-10 videos for each action, they estimate the semantics of *put, push, pull,* and *touch*. Their approach relies on manually labeled object and human pose, which could be automatically recognized, but already with the manual annotations the approach is unable to discriminate *push* and *pull*.

Yu and Siskind (2013) propose an approach to directly learn the correspondence between sentence components and videos. Nouns, verbs, prepositions, adjectives, and adverbs are grounded in video features, which consist of object detections and their velocity, as well as the pairwise features distance, size ratio, and relative horizontal position. In a restricted setting of 4 objects and limited grammar for sentences, they show that they can automatically describe new videos based on their learned correspondences. Other approaches which automatically generate natural language descriptions are discussed in detail in the next section.

2.3.3 Natural language generation from images and video

Generating descriptions of visual content can be roughly divided in four different categories according to: (1) generating descriptions by using manually defined rules or templates, (2) retrieving existing descriptions from similar visual content, (3) learning a language model from a training corpus to generate descriptions, or (4) generating summaries of descriptions with associated images or videos.

2.3.3.1 *Manually defined rules or templates*

Most works for video description harness manually specified rules and templates to generate language from an intermediate semantic representation, which is extracted from the visual content. This scheme allows precise production and is sufficient for most approaches which operate in a limited domain and complexity and when the variety of recognized and described elements is small or the generated language structure is simple. It is therefore not surprising that the first works on describing videos with natural language descriptions fall in this area. Nagel (1988) discusses some of these first efforts: for example, the Naos system (Neumann and Novak, 1983) generates case frames of street surveillance videos of a few video frames, identifying bounding boxes and their tracks and then describing their relations. Case frames represent the essential parts in a natural sentence, such as predicate, agent, location, and object. Nagel joint efforts with Zimmermann *et al.* (1987) to allow for a dialog system about visual scenes. Case frames are still used more than a decade later by Kojima *et al.* (2002), who build a concept hierarchy of actions which is represented by hierarchical case frames. In their work the hierarchy of case frames is manually defined and associated with different body, hand, and head movements in simple video scenarios.

Current work (Hanckmann *et al.*, 2012; Barbu *et al.*, 2012) has moved to realistic videos (here the DARPA Mind's eye corpus which depicts 48 different verbs) and extracts actions, body-pose, objects, and their tracks. Using a set of templates they generate text for their semantic representation. Similarly, Khan *et al.* (2011) uses templates to describe videos on the TREC Video summarization task.

Khan and Gotoh (2012) recognize humans and their activities. While they show only results on a limited set of six activities, they are able to recognize age, gender, and human emotions based on facial features (Maglogiannis *et al.*, 2009) as well as simple human interactions and spatial relations between objects and/or humans. They combine this diverse set of information with a context free grammar to describe videos. The description contains, in contrast to most other works, adjectives and sentence conjunctions, which are not based on a background language model but on visual recognition.

In an attempt to exploit the audio channel of many videos, Tan *et al.* (2011) learn audio-visual concepts and generate a video description for three different activities using rules to combine action, scene, and audio concepts with glue words.

Yang *et al.* (2011) are one of the first who use an external text corpus to improve generation. On the UIUC Pascal Sentence data set they recognize objects (DPM, Felzenszwalb *et al.*, 2010) and scenes (GIST, Torralba *et al.*, 2003). Then they add activities and prepositions according to a language model estimated from the newswire Gigaword corpus. A sentence is constructed according to a template using a Hidden Markov Model (HMM), which finds the best sentence given object detections and language model.

Krishnamoorthy *et al.* (2013) predict multiple subject-verb-object (SVO) triples for a subset of the MSVD corpus (Chen and Dolan, 2011) where the pre-trained DPM models are able to detect subjects and objects. The motion descriptors from Laptev *et al.* (2008) are used for recognizing the activity verbs. The detected activity verbs are further expanded to include similar verbs according to the WordNet hierarchy. For content planning Krishnamoorthy *et al.* combine the highest scoring object detections and verbs in all possible variations and re-score these SVO triples according to a large scale SVO language model. For surface realization, the best suited triple is used. Then multiple sentences are generated based on different templates which are again scored against an n-gram language model. While Krishnamoorthy *et al.* restrict themselves to videos related the VOC categories, Guadarrama *et al.* (2013a) scale this approach to the full MSVD corpus. To allow reasonable predictions in this more challenging setting they use stronger visual features (Li *et al.*, 2010b; Wang *et al.*, 2013) and trade-off the confidence along a classifier hierarchy to generate more abstract description in case of uncertainty, similar to the *"hedging your bets"* idea proposed for large scale object recognition by Deng *et al.* (2012).

Several authors have also started to generate multiple sentences. Gupta *et al.* (2009) learn AND/OR graphs to capture the causal relationships of actions given visual and textual data. During test time they find the most fitting graph to produce template-based, multi-sentence descriptions. Khan *et al.* (2011) produce multiple sentences and use paraphrasing and merging to get the minimum number of sentences needed. Using a simple template, Tan *et al.* (2011) generate a sentence every 10 seconds based on concept detection. For consistency they recognize a high level event and remove inconsistent concepts. Rather than using content-independent equal-length segments, Das *et al.* (2013b) segment the video based on the similarity of concept detections in neighboring frames. To generate sentences they combine two recognition approaches: the lower level recognition approach jointly models textual and visual words in a topic model (Das *et al.*, 2013a) and the higher level recognition approach combines the concept predictions in a tripartite graph of subject, tools, and objects. Each valid triple in the tripartite graph has a manually defined verb and template associated with it. After a verification step between low- and high-level recognition they retrieve the most likely training sentence. This makes this approach also fall into the retrieval category (see Section 2.3.3.2).

Templates have also been used to describe images. Kulkarni *et al.* (2011) extract objects and their attributes as well as their spatial prepositions from images. These entities are modeled in a Conditional Random Field (CRF). From the CRF predictions they generate descriptions based on simple templates; each relation in the CRF is transformed into a single sentence. Kulkarni *et al.* also explore using an n-gram language model to generate sentences (which falls into the third category), however this yields rather noisy and significantly worse results.

2.3.3.2 *Retrieval*

The second group of approaches reduces the generation process to retrieving sentences from a training corpus based on locally (Ordonez *et al.*, 2011) or globally (Farhadi *et al.*, 2010b) similar images. Using a Markov Random Field, Farhadi *et al.* (2010b) learn an intermediate semantic representation of object, action, and scenes, which is mapped to image and sentence space. On the *UIUC Pascal Sentence data set* with 1,000 images and 5,000 sentences they retrieve the closest sentence from training set by comparing it to the predicted semantic representation using a learned semantic representation-sentence mapping. To allow matching the semantic representation to out-of-vocabulary words used in the sentences, the mapping exploits the distance in the WordNet hierarchy using the Lin measure (Lin *et al.*, 2007) for objects and scenes, and co-occurrence statistics for verbs in image caption data.

To describe an image, Ordonez *et al.* (2011) retrieve the closest caption in the *SBU Captioned Photo Dataset* with over one million photographs. In a greedy approach they first select the 100 most similar images based on global image features. Then they run more costly image descriptor to recognize the objects, actions, scenes, people, and stuff. On the training set they learn weighting using linear regression or SVM, optimizing for BLEU score. They show that it is important to use such a large dataset for sentence retrieval to produce close image matches and thus precise captions.

2.3.3.3 *Language models for generation*

The third line of work, which also includes the approach proposed in this thesis, goes beyond retrieving existing descriptions by learning a language model to compose novel descriptions.

Yao *et al.* (2010) describe images via AND-OR-graphs, and then generate natural language using a small handcrafted grammar. The image is described with Web Ontology Language (OWL), which is then realized in language using a functional description and structure grammar, which ensures correct realization, e.g. person-number agreement. They apply their approach in a surveillance and road description task.

Two recent approaches use an aligned corpus of images and descriptions as a basis for generating novel descriptions for images using state-of-the art language generation techniques. Kuznetsova *et al.* (2012) retrieve candidate phrases from the SBU Captioned Photo dataset based on object, scene, and region recognition. Using an Integer Linear Programming formulation for content planning and surface realization they construct the most relevant and linguistically coherent descriptions. On the same dataset, Mitchell *et al.* (2012) use the visual recognition system of Kulkarni *et al.* (2011) to learn predicting sets of nouns and their order. They add necessary prepositions, predicates, and determiners to form syntactically well-formed phrases.

2.3.3.4 *Summaries of descriptions with associated images or videos*

Assuming the availability of text associated with the image at test time one can effectively use summarization techniques which benefit from visual content.

Starting from image tags, Aker and Gaizauskas (2010) mine several descriptions of the tagged buildings or locations using web search. Using multi-document summarization, they generate a single coherent image description. While Aker and Gaizauskas rely only on textual information, Feng and Lapata (2010) jointly represent text and images of news articles in a bag-of-words model to generate captions. They show that a topic model of textual and visual words significantly improves over text-only representations. They generate caption sentences not only by extracting sentences from the source news article, but also by combining the most relevant textual phrases according to visual and textual features. Long distance dependencies and syntactic structure are modeled using attachment probabilities between different phrases.

More recently, Kuznetsova *et al.* (2013) aim to improve the precision and visual relevance of user generated captions for images. They compress the sentences by optimizing linguistic fluency and consistency of the text and visual content. The visual content is estimated by object classifiers trained on ImageNet (Deng *et al.*, 2009) and the optimization is performed using dynamic programming and beam search. On the SBU Captioned Photo dataset of one million image-caption pairs, their refined descriptions are significantly more precise and relevant than the original captions.

2.3.4 Relations to our work

In this section we relate prior work to the three main contributions, namely our TACoS corpus, our grounding approach, and our approach for describing videos and images with natural language.

Similar to the MSVD corpus (Chen and Dolan, 2011) and the YouCook dataset (Das *et al.*, 2013b), our TACoS corpus contains videos with associated descriptions, as we will introduce in Chapter 8. TACoS is based on the videos of the MPII Composites dataset (Chapter 6) which is visually less varied than related dataset for descriptions. However, a key difference to the other datasets is, that it is annotated with an intermediate semantic representation for all time intervals, which allows supervised training of the visual recognition and language generation. Apart from MSVD, our dataset is also significantly larger with respect to videos, descriptions, and annotations as well as depicted and described activities and objects. The variability and size of MSVD is impressive, but it misses annotations and requires additional external training data to learn visual classifiers. It is questionable if the data is sufficient to learn reliable models, so far generated descriptions on the MSVD corpus are in most cases very noisy (Guadarrama *et al.*, 2013a).

While several works have explored grounding nouns in images (e.g. Leong and Mihalcea, 2011; Bruni *et al.*, 2011), grounding activity description in real video is less researched; exceptions are Mathe *et al.* (2008) and Yu and Siskind (2013). While Mathe *et al.* rely on annotated object and poses, Yu and Siskind detect objects in videos and estimate their spatial relations and motion. In contrast to Yu and Siskind which distinguishes only 4 objects, our work estimates the similarity between descriptive sentences, which are about a large range of activities and a large number of handled objects. For this challenge we propose the ASiM benchmark on TACoS and compare several visual activity recognition and text-based similarity models.

A few works have been proposed to ground motion verbs in video (Mathe *et al.*, 2008; Yu and Siskind, 2013), however, it is unclear how these approaches would scale to more challenging activities as they rely on optimal recognition (Mathe *et al.*, 2008) or reliable 2D geometric motion (Yu and Siskind, 2013).

Our approach to describe videos and images with sentences is presented in Chapter 9 and falls into the third category as it uses a language model to generate sentences. In contrast to other video description approaches we learn both, the visual recognition step and the language generation step from a parallel corpus rather than relying on manually defined sentence templates (e.g. Barbu *et al.*, 2012; Guadarrama *et al.*, 2013a) or retrieving sentences from the training corpus Das *et al.* (e.g. 2013b). Similarly to Guadarrama *et al.* (2013a) our translation approach weights resulting sentences according to a language model but we rely on our domain specific TACoS corpus rather than a general corpus. With respect to language generation our work is closest to the image description works of Kuznetsova *et al.* (2013) and Mitchell *et al.* (2012). However, while Kuznetsova *et al.* (2013) use hand-crafted constraints for content planning and surface realization solved by an Integer Linear Program and Mitchell *et al.* (2012) use a Tree-adjoining-grammar (TAG)-like natural language generation approach, we employ a co-occurrence based statistical machine translation (SMT) approach. We show that our SMT-based generation approach can achieve close to human performance given ground truth visual recognition and is applicable to image description. Similar to Kulkarni *et al.* (2011) our approach also relies on a CRF to predict an intermediate semantic representation, however, we model the relations between activities and their participants (tools, objects, and location) in videos, while the CRF model from Kulkarni *et al.* focuses on spatial relations between objects in an image.

3 COMBINING LANGUAGE SOURCES AND ROBUST SEMANTIC RELATEDNESS FOR KNOWLEDGE TRANSFER

Contents

In this chapter we use semantic relatedness minded from linguistic knowledge bases to enable knowledge transfer between visual object classes. Knowledge transfer between object classes has been identified as an important tool for scalable visual recognition. In contrast to previous work including our own (Rohrbach *et al.*, 2010), in this chapter, we explicitly aim to design robust semantic relatedness measures and to combine different language sources for attribute-based knowledge transfer. On the challenging Animals with Attributes (AwA) data set, we report largely improved attribute-based zero-shot object class recognition performance that matches the performance of human supervision.

In Chapter 4 we evaluate the approaches introduced in this chapter in a large scale setting. In Chapter 7 we show how to exploit the unlabeled data of the unseen classes and how to extend our approaches from zero-shot to the few-shot setting, i.e. benefiting from sample images if they are available.

3.1 INTRODUCTION

While remarkable recognition performance has been reported on a wide variety of object classes, scaling recognition to large numbers of classes remains a key challenge, mostly because of the prohibitive amount of required training data. Knowledge transfer between object classes has been advocated to reduce the amount of required training data by re-using acquired information in the context of related, but previously unknown recognition tasks (zero-shot recognition). Knowledge transfer on the level of attribute-based object class models has received particular attention (Ferrari and Zisserman, 2007; Lampert *et al.*, 2009; Wang and Forsyth, 2009). In (Rohrbach *et al.*, 2010) we proposed to combine attribute-based object class models with information mined automatically from linguistic knowledge bases, thereby avoiding any kind of human supervision. While we could show first promising results, only standard semantic relatedness measures were employed thereby limiting their robustness for visual object class recognition. At the same time, we suggested an alternative model for knowledge transfer (Rohrbach *et al.*, 2010), bypassing the intermediate layer of attributes. While this direct similarity-based model (Fink, 2004) exhibited superior performance for zero-shot recognition compared to the attribute-based model, it generalized significantly worse for a more realistic testing scenario in which training and test classes cannot be assumed disjoint.

The main objective of our work is therefore to explicitly adapt semantic relatedness to the specific task of attribute-based object class recognition, to improve the robustness and reliability of inter-class knowledge transfer. The first important tool for this task is the combination of different semantic relatedness measures and language sources, where we can benefit from their complementary strengths, compensating their weaknesses. The second important tool is to expand a given attribute inventory by additional attributes, in order to solidify the basis upon which class-level decisions are taken. Both tools aim at replacing individual semantic relatedness estimates taken between a pair of concepts by several measurements to increase robustness against errors.

The main contributions of this chapter are as follows. First, we explore novel semantic relatedness measures which we show to be more appropriate for attribute-based object class recognition than the ones used before (Section 3.5). Second, we suggest to combine individual semantic relatedness measures to yield more robust composite measures explicitly combining different language sources (Section 3.6). Third, we show how to expand a given attribute inventory with the help of semantic relatedness and demonstrate superior performance of the expanded inventory over the original one (Section 3.7). Fourth, we show that classifier level fusion further improves performance thereby attaining performance of human supervision (Section 3.8).

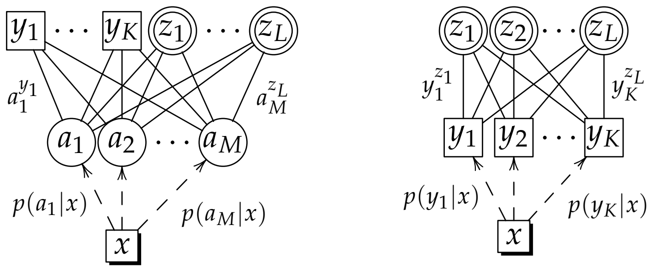
3.2 RELATED WORK

A prerequisite for knowledge transfer is an appropriate representation of transferable knowledge. Different representations have been proposed, ranging from discriminating aspects (Marszalek and Schmid, 2007; Zweig and Weinshall, 2007) to distance metrics (Bart and Ullman, 2005a; Fink, 2004; Thrun, 1996) and class priors (Li *et al.*, 2006; Stark *et al.*, 2009). Descriptive attributes offer an intuitive characterization of transferable knowledge (Ferrari and Zisserman, 2007; Wang and Forsyth, 2009; Kumar *et al.*, 2009; Farhadi *et al.*, 2010a). Lampert *et al.* (2009) introduced an attribute-based object class model for zero-shot recognition, based on human-provided associations between object classes and attributes.

In our earlier work, we demonstrated the successful combination of this object class model and semantic relatedness, replacing human supervision by information automatically mined from linguistic knowledge bases (Rohrbach *et al.*, 2010). In a similar zero-shot setting, Palatucci *et al.* (2009) compare the performance of a linguistic knowledge base (Google Trillion-Word-Corpus) to manual labels. However, the model is applied in the context of a completely different domain, namely, neural decoding of novel thoughts. Wang *et al.* (2009b) classify unseen butterfly categories according to text descriptions. While encouraging results using standard linguistic knowledge bases and semantic relatedness measures have been reported (Palatucci *et al.*, 2009), we believe there is significant room for improvement in the design of these measures towards their use in object class recognition. E.g. we found important differences among individual knowledge bases and semantic relatedness measures that one should exploit to improve robustness of the approach. The first goal of our work is therefore to build upon our previous work and to carefully design a customized inventory of semantic relatedness measures for zero-shot object class recognition.

We also investigate a second object class model for knowledge transfer, the so called direct similarity model (Rohrbach *et al.*, 2010). This model is also based on representing previously unseen object classes relative to known ones, characterizing unseen classes by their semantic relatedness to known classes (Fink, 2004; Bart and Ullman, 2005b). Interestingly, both models exhibit quite different behavior. While at first glance, direct similarity shows better absolute performance in zero-shot recognition, the attribute-based model seemingly generalizes better when leaving the rather artificial experimental setup of the Animals with Attributes data set (Lampert *et al.*, 2009), which assumes disjoint sets of object classes appearing in training and test. The second main goal of this chapter is therefore to leverage this essential advantage of the attribute-based model and push its performance to match that of direct similarity and human supervision.

As concerns linguistic knowledge bases and individual semantic relatedness measures, we go beyond the ones considered in (Rohrbach *et al.*, 2010), e.g. by adding Yahoo Snippets (Chen *et al.*, 2006) and Yahoo Near (Delezoide *et al.*, 2008) (see Section 3.5).



(a) Attribute-based (b) Direct similarity-based

Figure 3.1: Two models for zero-shot object classification. Known classes $y_1, \ldots, y_K$, unseen classes $z_1, \ldots, z_L$, attributes $a_1, \ldots, a_M$ and images x. See Section 3.3 for discussions.

3.3 OBJECT CLASS MODELS FOR KNOWLEDGE TRANSFER

As it is at the core of our approach, we briefly review the attribute-based models (see Figure 3.1(a)) for knowledge transfer at the core of our approach, as introduced by Lampert *et al.* (2009) (called direct attribute prediction model, DAP). Additionally we shortly introduce the direct-similarity based model from Rohrbach *et al.* (2010) (see Figure 3.1(b)) which we compare to. For a more detailed derivation, we refer the reader to Lampert *et al.* (2009) and Rohrbach *et al.* (2010), respectively.

3.3.1 Attribute-based classification

In the attribute-based model, the relation between known classes $y_1, \ldots, y_K$, unseen classes $z_1, \ldots, z_L$, and descriptive attributes $a_1, \ldots, a_M$ is given by a matrix of binary associations values a_m^y respective a_m^z (see Figure 3.1(a)) which encodes whether an attribute is active or inactive for a given class. While this association matrix is provided by human supervision in (Lampert *et al.*, 2009), it is derived from semantic relatedness measured between class and attribute concepts in (Rohrbach *et al.*, 2010). At training time, attribute classifiers are trained using the known classes $y_1, \ldots, y_K$. At test time, the activation of an individual attribute a_m in an image x is measured by its posterior probability $p(a_m|x)$, estimated from its classifier output. Multiple attribute activations are then combined to yield the posterior probability of the (unseen) object class z being present in the image

$$p(z|x) = \sum_{a \in \{0,1\}^M} p(z|a)p(a|x) = \frac{p(z)}{p(a^z)} \prod_{m=1}^{M} p(a_m|x)^{a_m^z}. \tag{3.1}$$

3.3.2 Direct similarity-based classification

The direct similarity model is structurally similar to the attribute-based model. It can be interpreted as a DAP with $M = K$ attributes, where attributes correspond to the known classes $y_1, \ldots, y_K$. The posterior probability of the (unseen) object class z being present in image x is then

$$p(z|x) \propto \prod_{k=1}^{K} \left(\frac{p(y_k|x)}{p(y_k)} \right)^{y_k^z}, \tag{3.2}$$

where y_k^z represents the semantic relatedness between known class y_k and unseen class z, see Figure 3.1(b).

3.4 EXPERIMENTAL SETUP

In the following sections we apply the attribute- and direct similarity-based object class models to the zero-shot classification task defined by the publicly available Animals with Attributes (AwA) data set (Lampert *et al.*, 2009). It consists of 50 mammal classes, each containing at least 92 images, together with a human-provided inventory of 85 attributes and corresponding object class-attribute associations (Kemp *et al.*, 2006; Osherson *et al.*, 1991). We follow the experimental protocol of Rohrbach *et al.* (2010) based on Lampert *et al.* (2009). We use the provided split into 40 training and 10 test classes (24,295 training, 6,180 test images) and the provided pre-computed feature descriptors, namely, RGB color histograms, SIFT, rgSIFT, PHOG, SURF, and local self-similarity histograms. We concatenate all features to a single vector and train histogram intersection kernel SVMs for classification, down-sampling all training images to the minimum number of 92 images available per class. We use libSVM with the built-in probability estimates (following Wu *et al.* (2004)) and a fixed cost parameter C=10.

3.5 INDIVIDUAL SEMANTIC RELATEDNESS (SR) MEASURES

We commence by determining the strength of object class-attribute associations (in the case of the attribute-based model) or object class-object class similarity (for the direct similarity-based model) by individual semantic relatedness measures.

3.5.1 Summary of semantic relatedness measures

We recapitulate briefly the linguistic knowledge bases and semantic relatedness measures we used in (Rohrbach *et al.*, 2010), since these constitute the starting point of our extensions. We put more emphasis on the description of those measures which we newly introduce, namely, Yahoo Snippets and Yahoo Near.

WordNet (Path). WordNet (Fellbaum, 1998) is the largest machine readable expert-created language ontology. Similarity of concepts is usually defined on its hierarchical graph structure, as, e.g., in the Lin measure (Lin, 1998).

Wikipedia (Vector) is the largest community built online encyclopedia. The Explicit Semantic Analysis (ESA) measure (Gabrilovich and Markovitch, 2007) is considered state-of-the-art (Zesch and Gurevych, 2010), representing each term as a vector of frequencies over all articles. Similarity of two terms is computed by the cosine between the two respective vectors.

Yahoo Web (HC). The web itself is apparently the largest collection of textual content. For semantic relatedness computation, actual content is usually summarized in the form of search engine (Yahoo) hit counts (HC). The Dice coefficient then measures similarity of two terms by the relative number of co-occurrences, inferred from hit counts

$$sim_{DICE}(t_1, t_2) = \frac{HC(t_1, t_2)}{HC(t_1) + HC(t_2)}. \tag{3.3}$$

Yahoo Img / Flickr Img (HC). In order to compensate for noise of full web page content, we restrict general web search to image search (Yahoo Img), or to a proper subset of the web devoted to collaborative photo sharing (Flickr Img).

3.5.2 Novel semantic relatedness measures

Yahoo Near (HC). Restricting search engine queries to holonym patterns as proposed by Berland and Charniak (1999) significantly improves the performance of Yahoo Web (HC) but is limited to part attributes. Similar in spirit, we suggest to impose proximity constraints on the occurrences of queried terms. The intuition is that requiring two terms to occur in proximity of one another in a document increases the likelihood of the co-occurrence being non-incidental and possibly even referring to the same physical entity. While Exalead (Delezoide *et al.*, 2008) offers a built-in Near operator providing this functionality, we implemented these constraints for the Yahoo search engine, using its wildcard operator ("*"). The above defined Dice coefficient can then be applied by letting $HC(t_1, t_2) \equiv HC(t_1 \; NEAR_k \; t_2)$, where $t_1 \; NEAR_k \; t_2$ limits the number of words occurring between t_1 and t_2 to at most k. We found $2 \leq k \leq 4$ to work best and thus consistently report results for $k = 4$ in all experiments.

Yahoo Snippets. A robust variation of hit count-based measures has been proposed by Chen *et al.* (2006), relying on short summary texts (snippets) accompanying the actual links returned by search engine (Yahoo) queries. In order to determine the relatedness of terms t_1 and t_2, the search engine is queried for t_1, measuring the frequency of occurrences of t_2 in the returned snippets, which we denote $f(t_2@t_1)$ and vice versa $f(t_1@t_2)$, explaining its common name "Web Search with Double

Checking". The snippet-based approach has two intuitive advantages. First, a term has to qualify for its appearance in a snippet according to some notion of importance, implemented by the search engine. Second, the ranking of search results can be taken into account when crawling snippets, which we do by restricting them to the $1,000$ highest ranked pages. The resulting semantic relatedness measure is computed in analogy to the Dice coefficient:

$$sim_{Snippets}(t_1, t_2) = \frac{f(t_1@t_2) + f(t_2@t_1)}{f(t_1@t_1) + f(t_2@t_2)} \tag{3.4}$$

We note that Chen *et al.* (2006) found CODC (Co-Occurrence Double Check) to outperform $sim_{Snippets}$. However, we found that CODC is not appropriate for the specific case of determining object class-attributes associations. It assumes a symmetric relation between two terms (by requiring $f(t_1@t_2)$ and $f(t_2@t_1)$ to simultaneously be greater than zero), which clearly does not hold for object class-attribute associations.

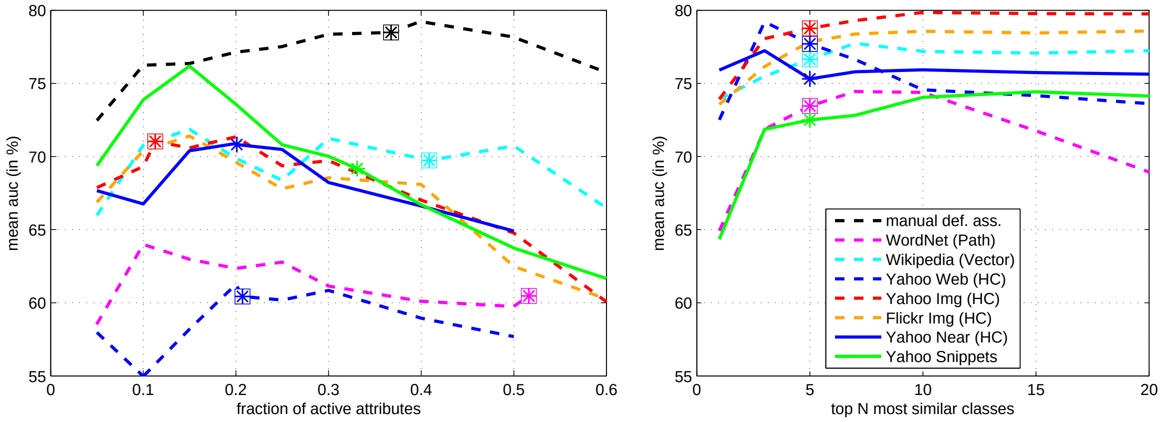
3.5.3 Discretizing semantic relatedness

Both attribute-based and direct similarity-based models for knowledge transfer require the discretization of semantic relatedness values. For the attribute-based model, semantic relatedness values have to be binarized to form an object class-attribute association matrix. This is typically done by applying a threshold t (Lampert *et al.*, 2009). For the direct similarity-based model, discretization is achieved through ranking: determining whether a test image contains an instance of test class z involves combining the classifier outputs corresponding to the N most similar training classes. In both cases, the choice of t or N can have a direct impact on performance (see below).

While Lampert *et al.* (2009) use the mean over all continuous-valued object class-attribute association matrix entries as the threshold t, we suggest to sample different points from the space of meaningful thresholds, according to the fraction of matrix entries becoming 1 after binarization. Likewise, we suggest to vary N for the direct similarity-based model instead of fixing it to $N = 5$ as done in (Rohrbach *et al.*, 2010).

3.5.4 Experimental results for individual semantic relatedness measures

We start with the discussion of zero-shot classification results on the AwA data set (Lampert *et al.*, 2009) using individual semantic relatedness measures, for both attribute-based and direct similarity-based models. Figure 3.2 plots the average classification performance over all 10 test object classes, measured as the mean area under the ROC curve (AUC), for attribute-based (Figure 3.2 (a)) and direct similarity-based models (Figure 3.2 (b)). Each curve corresponds to a distinct experiment using an individual semantic relatedness measure, varying either the applied binarization threshold t (Figure 3.2 (a)) or the number of considered most similar classes N (Figure 3.2 (b)). Additionally, we mark with an asterisk (*) the curve points for



(a) Attribute-based

(b) Direct similarity-based

Figure 3.2: Zero-shot classification results for individual semantic relatedness measures.

choices of t and N according to (Rohrbach *et al.*, 2010) and with a box (□) the curve points actually reported in (Rohrbach *et al.*, 2010). We give results for the measures of (Rohrbach *et al.*, 2010) (dashed curves) and the two novel measures that we propose in this thesis (solid curves). We also give the performance of the human-provided attribute association matrix (black dashed curve).

We begin with the general observation that varying the threshold t has a non-negligible impact on performance (Figure 3.2 (a)). E.g., for Yahoo Snippets (green solid curve), the performance difference is 14.5% between minimum (at a fraction of 0.6 active attributes) and maximum (at 0.15). The second general observation is that we can improve the results reported in (Rohrbach *et al.*, 2010) for all measures by varying t. The third general observations is that performance peaks are mostly located between 0.1 to 0.2 of active attributes, while performance drops beyond 0.2. This is contrary to human-provided associations and can be explained by the observation that the top-ranked associations are more reliable than lower ranks for semantic relatedness. As concerns the relative performance of the different measures, we note that the newly introduced Yahoo snippets (HC) (solid green curve) performs overall best (76.2%), outperforming all other measures by a large margin. The newly introduced Yahoo Near (HC) measure (solid blue curve) improves significantly (9.6% measured between the maxima of both curves) over its natural base line, Yahoo Web (HC) (dashed blue curve). We conclude that we can improve the results of the attribute-based model significantly already at the level of individual semantic relatedness measures.

For the direct similarity-based model (Figure 3.2 (b)), we observe similar general tendencies as for the attribute-based model. Choosing N different from its default value $N = 5$ always improves performance. Performance increases but saturates for higher values of N. As concerns the performance of the newly proposed measures, they tend to perform worse (Yahoo Snippets, solid green curve) or equal (Yahoo Near

(HC), solid blue curve) to the ones used in (Rohrbach *et al.*, 2010). The reasons for the limited improvements of the new measures for the direct similarity-based model are two-fold. First, the room for improvement is limited as, apart from WordNet, the previously used measures (dashed lines) provided already very reliable ranking for the most similar classes. Second, in contrast to attribute-based classification, Yahoo Snippets and Yahoo Near are now required to estimate relatedness of object classes instead of objects and their attributes. Both measures place a proximity requirement between the compared terms and this might reduce coverage for object-object association while objects and their attributes typically are close by, e.g. in phrases such as "*white sheep*" (color attributes), "*elephant*'s *tusks*" (part attributes), or "*swim* with *dolphins* in the *ocean*" (activity and context attributes).

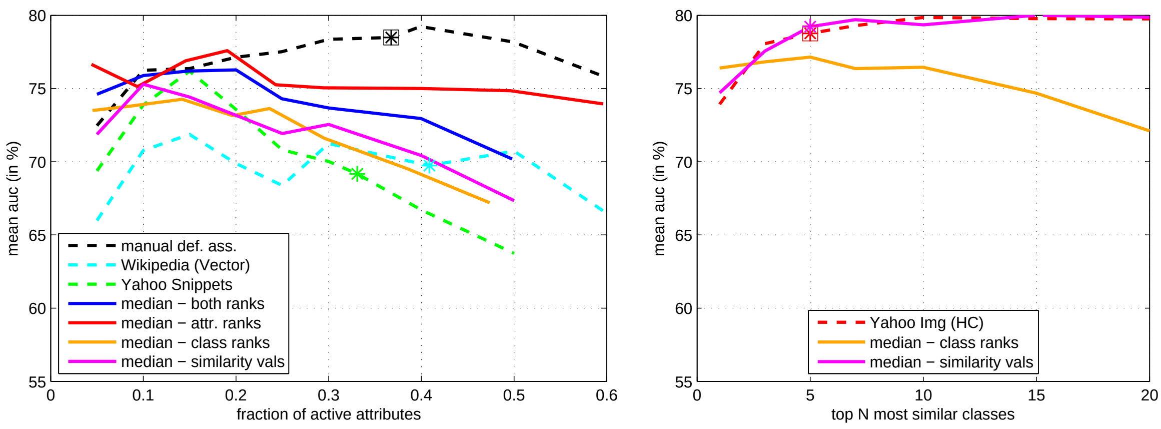
3.6 COMBINED SEMANTIC RELATEDNESS MEASURES

While we showed improved performance for two newly proposed individual semantic relatedness measures in Section 3.5, we observe there is still room for improvement. In particular, we hope to benefit from the complementary nature of different knowledge bases and semantic relatedness measures by combining individual measures to yield composite measures. As an example, consider the false positive associations between the attribute *big* and various object classes. While Wikipedia (Vector) and Yahoo Snippets list *chihuahua* among the 10 most strongly related classes, Yahoo Web (HC) lists *mouse*, Yahoo Img (HC) *mole*, Flickr Img (HC) *rat*, and Yahoo Near (HC) *beaver*. This diverse set of true positive associations is a clear hint towards complementary. In this section, we propose a strategy for exploiting these complementarities by combining measures, namely using median ranks.

Since semantic relatedness values computed by means of different measures are not per se comparable, an obvious pre-processing step for combination is to replace those values by a corresponding integer rank. For a given continuous-valued object class-attribute association matrix, this can be done either row-wise (producing an attribute ranking for each class) or column-wise (producing a class ranking for each attribute). Additionally, we can join both by first computing both attribute and class ranks, scaling them to the range $[0, 1]$, and multiplying the resulting values, yielding three different meaningful alternatives for rank computation. Having computed corresponding ranks for a number of individual semantic relatedness measures, a robust combination is the median over these ranks (i.e., the median over all corresponding entries in the object class-attribute association rank matrices of all measures).

3.6.1 Experimental results for median rank combined measures

Figure 3.3 gives results for the different variants of combining measures described above (solid curves), replicating the best curves of Section 3.5 as a reference (dashed curves). Again, each curve denotes a single experiment, varying the threshold t used



(a) Attribute-based

(b) Direct similarity-based

Figure 3.3: Zero-shot classification results for combined semantic relatedness measures.

for binarization of the object class-attribute association matrix. For the combinations, we consistently combine the five measures Wikipedia (Vector), Yahoo Img (HC), Flickr Img (HC), Yahoo Near (HC), and Yahoo Snippets.

As can be seen in Figure 3.3 (a), median attribute ranks (solid red curve) perform best, outperforming the best individual measures Wikipedia (Vector) (dashed cyan curve) and Yahoo Snippets (dashed green curve) consistently for all thresholds. The maximum performance is reached at a threshold of 0.19 with 77.6% mean AUC, which is close to human-provided associations (dashed black curve, attaining a maximum of 79.2%). At the same time, and in contrast to all other measures, median attribute ranks achieve stable performance beyond 0.3 active attributes. The median of both ranks (solid blue curve) is second best. The third best combination is an unranked version (solid magenta curve), where we directly compute the median over the original semantic relatedness values. It shows clearly inferior performance to the median attribute ranks and median of both ranks, and is even inferior to the individual measure Yahoo Snippets (dashed green curve). Median class ranks performs worst (solid orange curve). We attribute this drop in performance to the fact that using class ranks as object class-attribute associations results in all classes having the same number of active attributes. This is in stark contrast to the typically imbalanced number of active attributes which we observed in our experiments.

As an example of successful recovery from errors in individual measures by median attribute rank combination, consider the attribute *long leg*: while all individual measures wrongfully assign high ranks to classes such as *mole* (Yahoo Img (HC), rank 3), *seal* (Yahoo Near (HC), rank 3), *rat* (Yahoo Snippets, rank 5), *hippopotamus* (Flicker Img (HC), rank 3), and *bat* (Wikipedia (Vector), rank 2), the first erroneous rank for median attribute ranks is *bat* at rank 9.

For the direct similarity-based model results are shown in Figure 3.3(b). Median class ranks (solid orange curve) are inferior to the unranked version (solid magenta curve), whose performance is very close to the best individual measure Yahoo Img (dashed red curve).

3.7 EXPANDED ATTRIBUTE INVENTORY

Combining different linguistic knowledge bases and semantic relatedness measures by ranking enables us to achieve higher performance than using individual measures alone and can almost match human performance. This section takes a very different route compared to previous sections, by expanding the inventory of descriptive attributes provided as part of the AwA data set (Lampert *et al.*, 2009). While this inventory apparently provides a valid encoding of common and discriminating aspects between the various animal classes, intuition suggests two potential ways of increasing the overall robustness of the attribute-based model. The first way is obviously to improve robustness of individual attribute classifiers. The second way is to expand the inventory of attributes, similar in spirit to building strong ensemble classifiers from a plethora of weak ones, in order to solidify the basis on which class level decisions are taken. In the following, we pursue both directions, by explicitly expanding the given inventory of attributes by new ones, which we generate on the basis of the existing ones. In this way, we hope to benefit from increased robustness while preserving the valuable knowledge encoded in the original attribute inventory.

We start from the observation that each attribute in the attribute-based model induces a 2-partitioning of object classes and vice versa: one partition of classes where the attribute is active and another partition of classes where it is inactive. Based on this observation, we suggest to form new partitions (i.e., generate new attributes) by clustering object classes in some feature space. Each cluster then induces a partitioning: the cluster itself constitutes one partition, its complement the other partition. As features, we choose the semantic relatedness values computed between object classes and the original attribute inventory, thus preserving the inherent information encoded in the original attributes. By clustering, we effectively replace individual measurements of semantic relatedness by multiple measurements, which we hope will improve the robustness of the resulting attribute classifiers. Likewise, we vary the parameters of the clustering such that it produces varying numbers of induced attributes, thereby expanding the original attribute inventory also quantitatively.

Prior to clustering, we split the original attributes into a set of distinct categories, namely colors (8: red, green, . . .), texture (3: patches, spotted, stripe), skintype (3: furry, hairless, tough skin), stature (4: big, bulbous,. . .), parts (17: flipper, horn), locomotion (7: fly, hop,. . .), strength (3: strong, weak,. . .), moving behavior (5: active, agile,. . .), nutrition (5: meat, plankton,. . .), hunting style (6: grazer, scavenger,. . .), context (17: arctic, coastal,. . .), behavior (7: fierce, timid,. . .). k-means is then performed on a per-category basis to form aggregate attributes from semantically similar ones (e.g. black&white for the giant panda bear class).

In order to measure the qualitative differences to the original inventory of 85 attributes, we first generate an expanded inventory of size 85. We then further expand

	mean AUC in %		
	original attributes	clustered attributes	
	average	85	164
Individual semantic relatedness measures			
Wikipedia (Vector)	70.4	69.1 (-1.3)	72.4 (+2.0)
Yahoo Img (HC)	70.1	**75.2** (+5.1)	**77.2** (+7.1)
Flickr Img (HC)	69.5	73.7 (+4.2)	74.0 (+4.5)
Yahoo Near (HC)	69.4	70.6 (+1.2)	74.1 (+4.7)
Yahoo Snippets	72.9	72.0 (-0.9)	73.8 (+0.9)
Combined semantic relatedness measures			
median attribute ranks	76.0	74.8 (-1.2)	76.6 (+0.6)
median both ranks	75.3	73.5 (-1.8)	76.8 (+1.5)

Table 3.1: Zero-shot classification results for expanded attribute inventories in comparison to average performance over thresholds [0.1 0.3] from Sections 3.5 and 3.6; discussion in Section 3.7

this inventory by merging it with additional clusterings of varying k, resulting in an expanded inventory of 164 attributes. Please note that our clustering result is a hard assignment, corresponding to a single binarization threshold (0.14 for 85 and 0.22 for 164 attributes).

3.7.1 Experimental results for expanded attribute inventories

In Table 3.1, we give the results for two different clustering variants generating 85 (second rightmost column) and 164 attributes, respectively (rightmost column). The leftmost column lists the average performance of individual measures over varying thresholds between 0.1 and 0.3 as a reference (for the complete results refer to Figures 3.2 and 3.3(a)). Examining Table 3.1 we make two important observations. First, 164 clusters consistently outperform 85 clusters. Second, 164 clusters perform always better than the corresponding original attributes. For hit count-based measures (Yahoo Img (HC), Flicker Img (HC), Yahoo Near (HC)) and Wikipedia, this improvement is particularly pronounced. Notably, Yahoo Img (HC) improves to 77.2% mean AUC, which is very close to the performance of human-provided associations (79.2%). In summary, our results confirm the intuition given in the beginning of this section.

3.8 CLASSIFIER LEVEL FUSION

In Section 3.6 we showed the success of combining different measures on the level of semantic relatedness values. As a final step, this section explores fusing the different measures on classifier level. We achieve this by combining the class probabilities

		respective best without fusion			*fused*	
				mean		mean
#	Setting	reference	thresh	auc (%)	thresh	auc (%)
1	AwA attributes	Sec. 3.5, Fig. 3.2(a)	0.15	76.2	0.15	75.9 (-0.3)
2	85 clustered attributes	Sec. 3.7, Table 3.1	0.14	75.2	0.22	79.0 (+3.8)
3	164 clustered attributes	Sec. 3.7, Table 3.1	0.22	77.2	0.22	**79.5** (+2.3)
4	Direct similarity	Sec. 3.5, Fig. 3.2(b)	10	79.9	10	75.9 (-4.0)

Table 3.2: Classifier level fusion. Details in Section 3.8.

(i.e. the $p(z|x)$ values of Equation (3.1)) returned by different models. We use the product of the class probabilities for combination. We fuse the top 5 measures already combined in Section 3.6 for the attribute-based and direct similarity-based model (Section 3.5), as well as for expanded attribute inventories (Section 3.7).

3.8.1 Experimental results

Table 3.2 shows the results of fusion (rightmost columns) in comparison to the best results achieved without fusion for the respective settings (middle columns). As can be seen in lines 2 and 3 of Table 3.2, a significant improvement is achieved when fusing the classifier probabilities of the expanded attribute inventories (85 and 164 clustered). The combined model achieves a mean AUC of 79.0% and 79.5%, respectively, which is on the level of human-provided associations (79.2%). For the direct similarity-based model, fusion does not improve performance (line 4).

Fusing the predictions of models based on the original AwA attribute set (75.9% mean AUC) cannot exceed the best performing single measure Yahoo Snippets with 76.2% mean AUC (Table 3.2, line 1). However, we note that the fused measure provides consistently higher performance than the individual measures for the non-peak locations on the respective curves (not shown in the table). The fused measure is apparently not as sensitive to the selection of the binarization threshold, which is a valuable characteristic on its own. We consider this a highly promising result, as we managed to reach a performance level on par with using human-provided associations. As this is achieved for an attribute-based model, we expect better generalization than for direct similarity-based models, which we will explore in the next section.

3.9 EXTENDING TEST SET WITH IMAGES FROM KNOWN CLASSES

In all previous experiments, following the experimental protocol of Lampert *et al.* (2009), the set of object *classes* used for training and test were disjoint. This setting assumes that no images belonging to the known (training) classes are present at testing time. This setting is less challenging, as it does not require the zero-shot classifier to reject images from classes it already knows (i.e. the training classes).

				mean auc in %	
#	Setting / measure	Sec.	threshold	imgs: test	+ train cls
Object - Attribute Associations					
1	manually defined associations	3.5	0.40	79.2	79.4 (+0.2)
2	Yahoo Img (HC)	3.5	0.11	71.0	73.2 (+2.2)
3	median: attribute ranks	3.6	0.19	77.6	**79.2** (+2.4)
4	164 clustered: Yahoo Img (HC)	3.7	0.22	77.2	76.9 (-0.3)
5	classifier fusion: 164 clustered	3.8	0.22	**79.5**	78.9 (-0.6)
Direct Similarity					
6	Yahoo Img (HC)	3.5	5	78.8	76.0 (-2.8)
7	Yahoo Img (HC)	3.5	10	**79.9**	**76.4** (-2.5)
8	classifier fusion	3.8	10	75.9	72.3 (-3.6)

Table 3.3: Effect of images from known classes in the test set. Selection of respective best from Sections 3.6-3.8. Discussion in Section 3.9.

Using images from the training classes (that were not used for training) as additional negative examples for testing is an especially difficult (adversary) setting, as it requires the classifier to generalize over the known classes. We argue that this more difficult setting is also more realistic and allows us to draw conclusions that are more appropriate to a real-life object recognition setting. Thus, following (Rohrbach *et al.*, 2010), we report results using all images from the test classes not used for training as additional negatives in the test set.

3.9.1 Experimental results

Table 3.3 lists the best results from (Rohrbach *et al.*, 2010) as well as the best measures and combinations of the previous sections. The second last column gives results when including training class images as negatives in comparison to the performance reported in the previous sections (third last column).

The most important observations based on the results in Table 3.3 are: First, while human-provided associations show stable results (line 1), performance of direct similarity significantly drops when including training class images (line 6). We could slightly increase overall performance by varying thresholds (line 7), but direct similarity does not level with human-provided associations for the more difficult adversary setting, even when fusing on classifier level (line 8). Second, in contrast to direct similarity, we found attribute-based measures, e.g. Yahoo Img (line 2), to slightly improve in most cases, i.e. generalize well. Third, the best combined models, median attribute ranks (line 3) and classifier fusion with the 164 clustered attribute inventory (line 5) are not only very competitive in terms of performance, but also perform well in this adversary setting (79.2%, 78.9%), on par with the model using human-provided associations (79.4%). This property makes these measures favorable to those based on direct similarities that are less suited to recognize (and reject) training classes at testing stage.

3.10 CONCLUSIONS

In this chapter we propose several tools to increase the robustness of semantic relatedness for use in attribute-based zero-shot object class recognition, leading to performance on par with human supervision. First, on the level of individual measures we find Yahoo Snippets to provide significantly higher performance than the measures used in our previous work (Rohrbach *et al.*, 2010). Second, combining individual measures on the level of semantic relatedness values achieves performance close to human-provided associations using attribute ranks. Third, expanding the attribute inventory using clustering also reaches performance close to human supervision for the Yahoo Image (HC) measure. Finally, fusing measures on classifier level achieves performance on par with human supervision for expanded attribute inventories. This is particularly valuable, since the attribute-based model generalizes well even for the difficult setting when images from known classes are added to the test set.

In the next chapter, we evaluate how these measures and models scale to 1,000 object classes for traditional supervised object recognition and to 200 unseen object classes.

4 KNOWLEDGE TRANSFER AND ZERO-SHOT LEARNING IN A LARGE-SCALE SETTING

Contents

In the previous chapter we explored how to increase robustness of semantic relatedness mined from language resources for visual knowledge transfer. However, our experiments were still limited in the number of object classes considered. To support claims of knowledge transfer w.r.t. scalability we evaluate knowledge transfer in a large-scale setting in this chapter. To this end, we provide an extensive evaluation of three popular approaches to knowledge transfer on a recently proposed large-scale data set, the ImageNet Large Scale Visual Recognition Competition 2010 data set. In a first setting they are directly compared to one-vs-all classification often neglected in knowledge transfer papers and in a second setting we evaluate their ability to enable zero-shot learning. While none of the knowledge transfer methods can improve over one-vs-all classification they prove valuable for zero-shot learning, especially hierarchical and direct similarity based knowledge transfer. We also propose and describe several extensions of the evaluated approaches that are necessary for this large-scale study.

In Chapter 6 we will pick up the idea of knowledge transfer with attributes again, but this time for composite activities rather than objects and using a script data to compute semantic similarity. In Chapter 7 we present how to exploit unlabeled data of the unseen classes and how to integrate a few sample instances, comparing to the results in this chapter.

4.1 INTRODUCTION

Inspired by the success of recent object class recognition on individual classes, the simultaneous recognition of many classes has become an active research area. Scaling recognition to larger numbers of classes poses challenges with respect to the expressiveness and learnability of object models as well as the need for increasing amounts of training data. Knowledge transfer between object classes has been advertised as a promising route towards scalable recognition, by efficiently re-using acquired knowledge in the context of newly posed, but related recognition tasks. While experimental studies connected to knowledge transfer have shown promising results they are often limited w.r.t. the size of employed data sets.

As a consequence, it remains unclear whether the benefits demonstrated in small-scale experiments considering only a few classes really take effect in large-scale settings. In fact, Deng *et al.* (2010) found that the relative performance of different recognition methods can change when increasing test database size by an order of magnitude. The major contribution of this chapter is therefore to revisit three recently proposed knowledge transfer approaches and to evaluate them in a truly large-scale setting, effectively starting where previous evaluations have left off. We evaluate knowledge transfer on the ImageNet data set (Deng *et al.*, 2009), specifically, on the associated ImageNet Large Scale Visual Recognition Competition 2010 (ILSVRC10) subset (Berg *et al.*, 2010a). It consists of over 1.2 million images of 1,000 object classes, providing a currently unparalleled test bed for vision algorithms in terms of both scale and diversity. Being based on WordNet (Fellbaum, 1998) synonym sets, ImageNet offers the additional advantage of providing a hierarchical organization of object classes according to hypernym/hyponym relations, lending itself to knowledge transfer using object class hierarchies.

Our experimental study follows three prominent directions in knowledge transfer, which have proven effective for comparatively small numbers of object classes. The first direction imposes a hierarchical structure on the space of object classes, according to the general-to-specific ordering defined by the data set (Griffin and Perona, 2008; Marszalek and Schmid, 2007; Zweig and Weinshall, 2007). The second direction is based on representing object classes relative to an inventory of generic visual attributes (Farhadi *et al.*, 2010a; Lampert *et al.*, 2009; Rohrbach *et al.*, 2010), where classes are characterized by distinct patterns of attribute activations. The third direction is based on direct similarities to related classes effectively using the classifiers of most similar classes (Bart and Ullman, 2005b; Fink, 2004; Rohrbach *et al.*, 2010). For all three directions we go far beyond previous studies in terms of data set size, and evaluate knowledge transfer in the context of both traditional multiclass classification and zero-shot recognition.

This chapter makes the following contributions: First, to the best of our knowledge, we are the first to provide an in-depth study of knowledge transfer in a truly large-scale setting. Second, we compare three different approaches to knowledge transfer: one based on an object class hierarchy, one based on attributes, and one based on direct similarity. Third, we contrast knowledge transfer with the traditional approach of one-versus-all classification (Rifkin and Klautau, 2004), which is often neglected in previous knowledge transfer work. Fourth, we challenge fully unsupervised transfer in a zero-shot recognition task aiming to recognize 200 unseen test classes. Fifth, we propose technical modifications to several approaches making them applicable to large-scale data.

The remainder of the chapter is organized as follows. Section 4.2 discusses related work. Section 4.3 introduces the different knowledge transfer approaches. Section 4.4 motivates our setup for the experiments in Sections 4.5 and 4.6.

4.2 RELATED WORK

Knowledge transfer for object class recognition comes in different flavors, such as joint learning of multiple classes (Torralba *et al.*, 2004) or transferring object class priors (Li *et al.*, 2006). Recently, three lines of research have gained particular popularity due to their potential scalability.

A first line of research exploits the hierarchical structure of the object class space imposed by a general-to-specific ordering, either based on an existing hierarchy (Marszalek and Schmid, 2007; Zweig and Weinshall, 2007) or learned from visual features (Griffin and Perona, 2008). Scalability is achieved by associating classifiers to each hierarchy node, allowing for classification in a divide-and-conquer fashion. Our hierarchical classification is closest to Deng *et al.* (2009), combining classifier scores of distinct subgraphs to yield final classification scores. Deng *et al.* (2010) follow a different route by forming a weighted average of all classifiers in a hierarchy for classification. While the latter two approaches report multiclass classification results on (subsets of) the ImageNet data set, our study additionally considers zero-shot recognition.

A second line of research uses an intermediate layer of descriptive attributes to represent object classes (Farhadi *et al.*, 2010a; Lampert *et al.*, 2009; Rohrbach *et al.*, 2010), encoding high-level visual properties that can be shared among object classes, hence promoting scalability. Our attribute-based object class model is inspired by Lampert *et al.* (2009, 2013), and uses linguistic knowledge bases to determine both an attribute inventory and the associations between object classes and attributes fully automatically (Rohrbach *et al.*, 2010).

A third line of research uses direct similarities between object classes. Bart and Ullman (2005b) encode instances of previously unknown classes as collections of "familiar" classifier responses, i.e., similarities to known classes, and applying a

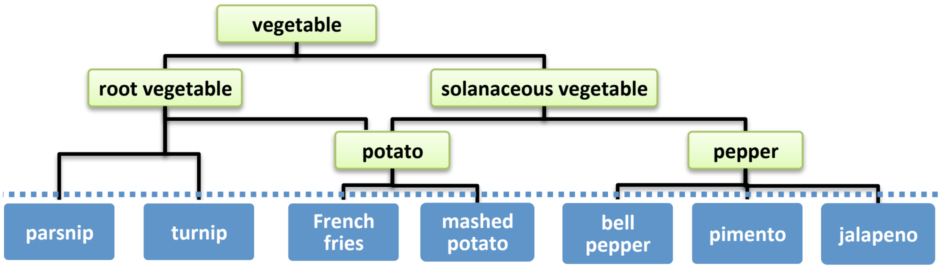


Figure 4.1: ILSVRC10 subgraph. Leaf (blue), inner nodes (green).

nearest-neighbor scheme for classification. While most work based on similarity between classes (Bart and Ullman, 2005b; Fink, 2004) require a few training samples for new classes, we employ our unsupervised approach (Rohrbach *et al.*, 2010) where class similarities are mined automatically using semantic relatedness measures with linguistic knowledge bases like Wikipedia or web search.

4.3 KNOWLEDGE TRANSFER APPROACHES

In this chapter we explore two distinct settings for knowledge transfer. In a first experiment (Section 4.5) we assume that training data is available for all classes. In this setting knowledge can be transferred (or shared) among all classes and thus may lead to better classification performance. This setting is called *knowledge sharing* in the following. In the second experiment we assume that training data is available for a subset of known classes and that no training data is available for the remaining unseen classes. This setting is called *zero-shot recognition* and described in Section 4.6. We have chosen these two distinct settings as they represent two extreme cases for knowledge transfer.

The following gives an overview of the different knowledge transfer approaches explored in our study. Section 4.3.4 then describes how semantic relatedness is used to enable unsupervised attribute- and direct similarity-based knowledge transfer.

4.3.1 Hierarchy-based knowledge transfer

We exploit the hierarchical structure of the ILSVRC10 to train two types of classifiers (see for a small sample subgraph Figure 4.1). We train classifiers for leaf nodes z_l by using training images of that node as positive samples and all other images as negative samples. Additionally we train classifiers for inner nodes y_i using all images associated to hyponyms of y_i as positive and all images outside the subtree rooted at y_i as negative examples. Figure 4.1 shows an example, where a classifier for *solanaceous vegetable* uses *French fries*, *mashed potato*, *bell pepper*, *pimento*, and *jalapeno* images as positives as well as *parsnip* and *turnip* images as negative examples. We exclude the root and any trivial nodes (with only a single hyponym), as they do not

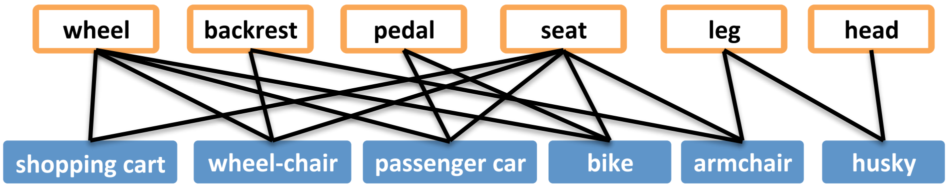


Figure 4.2: Example part attributes (orange), object classes (blue).

provide additional information, resulting in a total of 370 inner node classifiers.

We distinguish three approaches. First, for scoring image x according to a leaf class z_l, we average over all classifier scores $s(y_i|x)$ of hypernyms H_{z_l} of z_l (for a *bell pepper* classifier we thus use the *pepper* and *solanaceous vegetable* classifiers), which we denote the **inner WordNet nodes** model:

$$s^{inn}(z_l|x) = \frac{\sum_{y_i \in H_{z_l}} s(y_i|x)}{|H_{z_l}|} \tag{4.1}$$

Second, since this model is not capable to distinguish among leaf classes z_l that share the same hypernyms, such as *French fries* and *mashed potato*, we also include leaf node classifiers in the **all WordNet nodes** model:

$$s^{all}(z_l|x) = \frac{s(z_l|x) + \sum_{y_i \in H_{z_l}} s(y_i|x)}{1 + |H_{z_l}|} \tag{4.2}$$

The third approach is based on the hierarchical cost sensitive classifier proposed by Deng *et al.* (2010). This formulation tries to optimize for the hierarchical error, defined in Section 4.4.2. To estimate the score of a certain class z_l we use cost-weighted classifier probabilities of all **leaf nodes** Z_l**, cost sensitive** to the cost $c_{z_i}^{z_l}$ between nodes z_i and z_l which is equivalent to the hierarchical error:

$$s^{cost}(z_l|x) = -\sum_{z_i \in Z_l} c_{z_i}^{z_l} p(z_i|x) \tag{4.3}$$

The hierarchy-based model allows for a flexible combination of leaf and inner node classifiers. In the *knowledge sharing* case the inner and leaf node classifiers are trained on training data from all classes. In the *zero-shot* case only those leaf node classifiers can be trained for which training data is available and the inner node classifiers are trained on the known classes only. Figure 4.5 gives an example for transferring knowledge using leaf, inner, and all WordNet nodes models accordingly for the *zero-shot* case.

4.3.2 Attribute-based knowledge transfer

We adopt the probabilistic direct attribute prediction model (DAP) introduced by Lampert *et al.* (2009). The DAP represents object classes z_l relative to an inventory of

Dataset & Approach	Error	Product	Sum
ILSVRC 10, inner nodes	Top 1	93.5	**90.9**
ILSVRC 10, inner nodes	Top 5	80.1	**71.6**

Table 4.1: Evaluation of the probabilistic product model suggested by Lampert *et al.* (2009) vs. our sum model, see Section 4.3.2. Error in %.

descriptive attributes a_m, realized as probabilistic attribute classifiers $p(a_m|x)$. In the *knowledge sharing* case these are trained on all classes whereas in the *zero-shot* case these are trained on known classes only. Once trained, the attribute classifiers can be flexibly combined to recognize previously unseen classes in the *zero-shot* setting or to recognize known classes in the *knowledge sharing* case. The association between object classes z_l and attributes a_m (see Figure 4.2 for an example) is controlled by a matrix of indicator variables $a_m^{z_l}$. Assuming mutual independence of attributes and uniform priors $p(a_m) = 0.5$ yields the following probability estimate of class z_l being present in image x (Rohrbach *et al.*, 2010):

$$p^{attr}(z_l|x) \propto \prod_{m=1}^{M} (2 * p(a_m|x))^{a_m^{z_l}} \tag{4.4}$$

For efficiency reasons, we propose the following non-probabilistic sum formulation, which replaces calibrated attribute probabilities $p(a_m|x)$ by zero-boundary attribute decision scores $s(a_m|x)$:

$$s^{attr}(z_l|x) = \frac{\sum_{m=1}^{M} s(a_m|x)^{a_m^{z_l}}}{\sum_{m=1}^{M} a_m^{z_l}}, \tag{4.5}$$

Although this formulation does not require calibrated probabilities, it does require normalized scores. We found empirically that a simple z-score is sufficient.

In order to validate the sum formulation, we compare its performance to the probabilistic formulation in Table 4.1 for both error measures (see Section 4.4.2 for details). The important observation is that the sum formulation outperforms the probabilistic formulation.We thus use the sum formulation in the following.

4.3.3 Direct similarity-based knowledge transfer

Motivated by its superior classification performance (Rohrbach *et al.*, 2010), we also include a direct similarity based approach. This can be defined as a modification of the attribute-based model that represents object classes relative to a set of K semantically related reference classes z_k, implemented by classifiers $s(z_k|x)$:

$$s^{dir}(z_l|x) = \frac{\sum_{k=1}^{K} s(z_k|x)}{K}, \tag{4.6}$$

Direct similarity is used only in zero-shot experiments as the most related known class in the knowledge sharing setting is always the class itself.

4.3.4 Semantic relatedness for attribute- and direct similarity-based approaches

The attribute–based approach relies on an association matrix between a set of attributes and the object classes. The ILSVRC10, however, is neither provided with a set of attributes nor with manual class-attribute associations. Therefore we rely on part attributes mined from WordNet to generate an inventory of attributes for all classes (Rohrbach *et al.*, 2010). In total we mine 811 part attributes. An alternative to mine attributes would be to use WordNet's synset definitions as proposed by Russakovsky and Fei-Fei (2010).

For these mined attributes we use semantic relatedness measures in connection with linguistic knowledge bases to automatically determine associations between the attributes and object classes. While in Chapter 3 and (Rohrbach *et al.*, 2010) each class and attribute is associated with one term, the classes and attributes in this chapter refer to WordNet concepts, called *synsets*, which are represented by several terms. As the semantic relatedness measures are based on terms rather than semantic concepts we take the median over all possible term combinations for a specific association.

For mining class-attribute associations we choose the best performing measures from Chapter 3 and (Rohrbach *et al.*, 2010) which are applicable to large scale: (1) the explicit semantic analysis (Gabrilovich and Markovitch, 2007) based on *Wikipedia* (Szarvas *et al.*, 2011); (2) *Yahoo Holynyms* which is based on hitcounts and uses specific part queries such as "the wheel of the car"; (3) *Yahoo Image* which is based on image-search hitcounts; and (4) *Yahoo Snippets* which is based on web page summaries returned by the search engine. For the direct similarity based approach we replace Yahoo Holonyms with simple *Yahoo Web* queries as it is not applicable for direct similarity. For improved robustness of the attributes we also compute a class level fusion over *all attributes*.

Robust associations for large scale. In contrast to prior work we have a significantly larger amount of potential classes associated to each attribute. To learn precise attribute classifiers we use only the most likely classes as positives and least likely as negatives, leaving out the potentially noisy middle part. For the attribute *backrest* in Figure 4.2 we would thus use *wheelchair* and *armchair* as positives, *bike* and *husky* as negatives, and not use the classes *shopping cart* and *passenger car* which are uncertain in respect to the attribute *backrest*.

Parameter selection. For attribute- and direct similarity-based knowledge transfer, continuous semantic relatedness measures have to be discretized to yield binary associations between attributes and object classes and in between object classes, respectively, by thresholding. Since we found large performance differences depending on thresholding in Chapter 3, we determine threshold values on the validation set, and fix them for the rest of the experiments. In particular, for attribute-based knowledge transfer, we set the threshold such that, on average, 3% of all attributes are active for a given object class. For the direct similarity based approach, we set the threshold such that the $K = 5$ most related object class models are considered.

4.4 EXPERIMENTAL SETUP

Evaluating and comparing the different knowledge transfer approaches of Section 4.3 in a large scale setting requires careful design of the experimental setup. The following details and argues for our choices concerning data set, image representation, and learning methodology.

4.4.1 Dataset

The number of available datasets containing more than a few hundred object classes with sufficiently many images per class is still limited. Caltech256 (Griffin *et al.*, 2007) is frequently used, however, it consists only of 256 classes and 30k images. NUS-WIDE (Chua *et al.*, 2009) is significantly larger with 270k images and over 5k unique tags but contains ground truth for only 81 categories. The tiny image data set (Torralba *et al.*, 2008) (80 million images, loosely labeled with 75,062 WordNet nouns) provides a significantly larger number of images but is mostly restricted to 32x32 pixel images.

Deng *et al.* (2009) proposed ImageNet (3.2 million images of 5247 WordNet synonym sets) as a resource for truly large-scale experimentation. Based on this dataset the ImageNet Large Scale Visual Recognition Challenge 2010 (ILSVRC10, Berg *et al.*, 2010a) has been introduced. We have chosen this subset for large-scale experiments as it is a well-defined subset of 1,000 object classes (1.2 million images, divided into distinct portions for training, validation, and test) for classification experiments, suggesting this benchmark to be the de-facto choice for large-scale experiments in the near future.

4.4.2 Performance measures

ILSVRC10 (Berg *et al.*, 2010a) introduced and defined the following performance measures used in this chapter and in Chapter 7. Performance is measured as the top-n error rate (the n most confident classification hypotheses are considered as potentially correct) and distinguishes two error measures. The first is a *flat* measure which equals 0 if the test class is predicted correctly within the n most confident hypotheses, and 1 otherwise. The second is a *hierarchical* measure, which equals the minimum height of the lowest common ancestors between true and hypothesized classes. As suggested by Berg *et al.* (2010a) we report top-n errors for $n = 5$ and $n = 1$, which corresponds to $1-$accuracy. In order to avoid fitting the test data, we use the provided validation set for preliminary experimentation and parameter selection (Figure 4.3 and 4.4(a), Table 4.1 and 4.2). The final results (Section 4.5 and 4.6, Figure 4.4(b), Table 4.3 and 4.4) are obtained on the test set.

Model	Descriptor	Learning method	Total dim.	Err. top 5	1
BoW (Berg *et al.*, 2010a)	Sift	LibLinear	1,000	80	91
BoW	Sift	MeanSGD	1,000	72	86
BoW + SPM	rgSift	MeanSGD	8,000	59	76
LLC + SPM	rgSift	MeanSGD	21,000	50	69
Fisher vector	rgSift	MeanSGD	32,768	43	61
LLC+SPM, Fisher	**rgSift**	**MeanSGD**	**53,768**	**38**	**57**
Fisher+SPM (Perronnin *et al.*, 2010)	Sift, Color	SGD	262,144	34	–
LLC,SVC+SPM (Lin *et al.*, 2011)	Hog, Lbp	ASGD	1,179,648	28	47

Table 4.2: One-vs-all performance of different methods on ILSVRC10. BoW: bag of visual words, SPM: spatial pyramid matching (Lazebnik *et al.*, 2006), LLC: locality-constrained linear coding (Wang *et al.*, 2010), Fisher vector (Perronnin *et al.*, 2010), SVC: Super-Vector Coding (Xi Zhou and Huang, 2010), Lbp: local binary patterns, SGD: stochastic gradient decent (Bordes *et al.*, 2009), ASGD: averaging SGD (Lin *et al.*, 2011).

4.4.3 Image representation

In order to allow for a sufficient range of experiments on the ILSVRC10 dataset, we require an image representation that is both powerful enough to achieve good performance and reasonably sized to support efficient learning. We thus base our choice on the outcome of the ILSVRC10 competition, which we recapitulate in part in Table 4.2, and seek to find a compromise between performance and manageable runtimes.

We observe that the performance ranges from 80% top-5 error rate for a BoW Sift baseline (Table 4.2, first row) to an impressive performance of as low as 34% and 28% top-5 error of the best performing approaches (Table 4.2, last two rows).

We note that in the ILSVRC 2012 and 2013 deep learning approaches using convolutional neural networks (Krizhevsky *et al.*, 2012) have overtaken these approaches. Zeiler and Fergus (2013) achieve top-5 error as low as 11%[2].

In an attempt to regulate the performance-runtime tradeoff, we explore different combinations of techniques used by the best performing approaches (Perronnin *et al.*, 2010; Lin *et al.*, 2011) such as spatial pyramid matching (SPM, Lazebnik *et al.*, 2006)), locality-constrained linear coding (LCC, Wang *et al.*, 2010), and the Fisher vector (Perronnin *et al.*, 2010) (we adapted the implementation of Jégou *et al.* (2010)), in connection with the color sift variant rgSift (van de Sande *et al.*, 2010) (Table 4.2, rows 2 to 6).

As can be seen from Table 4.2 and Figure 4.4(a) (blue dots) the performance increases monotonically with descriptor dimensionality. While the last two approaches perform best they use feature vectors of several 100k and over one million dimen-

[2] http://www.image-net.org/challenges/LSVRC/2013/results.php

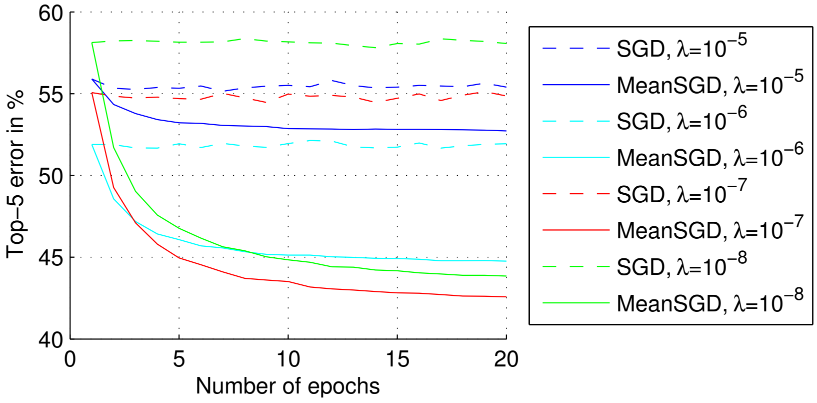


Figure 4.3: Convergence of SGD and MeanSGD for different step sizes λ on ILSVRC10 (setting: one-vs-all, Fisher vector, rgSift).

sions, resulting in prohibitive runtimes for our purposes. For this thesis we opt for the Fisher vector and LLC+SPM representation as a sensible compromise between performance (38% top-5 error rate, Table 4.2, row 6) and runtime. For combining the two representations we simply average their scores. We fix this representation for all remaining experiments.

4.4.4 Learning method

Motivated by the potential of stochastic gradient-based optimization for rapid convergence, and in line with the two best performing ILSVRC10 approaches, we use linear SVM classifiers, trained using stochastic gradient descent (SGD) (Bordes *et al.*, 2009). Similar in spirit to averaging SGD (ASGD) (Polyak and Juditsky, 1992; Lin *et al.*, 2011), we average the SVM's weight and bias. However, in contrast to Lin *et al.* (2011) we do not average after each step, but take the mean of the results after each epoch (one pass over the data). More specifically, we save the weight vector w_i and bias b_i after each epoch i (the data is randomly reordered before each epoch). While the score of the normal SGD after n epochs only depends on the weights and bias after the final epoch

$$f_{SGD}(x) = \langle w_n, x \rangle + b_n, \tag{4.7}$$

we compute the mean over all epochs in MeanSGD:

$$f_{MeanSGD}(x) = \frac{\sum_{i=1}^{n} \langle w_i, x \rangle + b_i}{n} \tag{4.8}$$

(where $\langle w, x \rangle$ is the scaler product of w an x).

As can be seen in Figure 4.3, using MeanSGD (solid lines) instead of SGD (dashed lines) significantly speeds up convergence and improves performance. We use hinge loss and fix, according to Figure 4.3, the step size λ to 10^{-7} and the number of epochs n to 20 epochs.

In order to benefit from modern multi-core hardware, we further implemented a parallelized version of MeanSGD based on Bouttou's SGD (Bottou, 2010), exploiting data parallelism. It requires about 20 hours (including file and network I/O) for training all 1,000 one-vs-all classifiers with 20 epochs using the 53,768 dimensional Fisher vector on a 32-core machine. The code including a Matlab wrapper is available on our webpage.

4.5 LARGE SCALE KNOWLEDGE SHARING

As motivated in Section 4.3, in a first set of experiments we consider the *knowledge sharing* case where we assume to have training samples for all classes.

Table 4.3 gives results for classifying all test images of the ILSVRC10 data set into 1,000 classes, using the provided training set for training. Performance is measured in terms of the corresponding flat and hierarchical (in brackets) variants of top-5 and top-1 error (see Section 4.4.2). The table compares the performance of standard one-vs-all classification (part 1 of Table 4.3, using leaf node z_l classifiers only), hierarchical models (part 2), and attribute-based models (part 3).

We proceed by examining Table 4.3 from top to bottom. First, we observe that the standard one-vs-all approach (Table 4.3 part 1) achieves a remarkable top-5 error rate of 37.6% with a hierarchical error rate of 2.91.

In contrast, the hierarchical model using only inner nodes (Table 4.3 part 2) performs relatively poorly (top-5 error of 71.3%, hierarchical error 7.31). This drop is understandable, considering the much smaller number of available inner node classifiers (370 compared to 1,000 leaf node classifiers). Adding the leaf nodes boosts the performance of the hierarchical model by more than 20% w.r.t. the flat top-5 error rate (50.4%, hierarchical error rate 5.49). Surprisingly, the resulting performance is still slightly worse than one-vs-all – the effect of the added confusion by more uniformly weighted classifiers is apparently more pronounced than the added discriminative power. When examining the results more closely we find that the performance of the inner leaf node classifier does not correlate with the level of abstraction in the hierarchy. However, we find that it strongly depends on the semantic grouping, e.g. the category flower which is associated with 87 leaf nodes can be very well separated from other nodes in contrast to the class node described with the synset {fastener, fastening, holdfast, fixing}, which has 10 visually diverse and difficult child nodes such as *button, hair slide, knot,* and *screw.*

The hierarchical approach based on the approach proposed by Deng *et al.* (2010) uses one-vs-all leaf nodes, but makes them sensitive to the hierarchical cost (see Section 4.3.1). With 48.6% top-5 error (Table 4.3 part 2) it clearly outperforms the hierarchical approach using only inner WordNet nodes (by 23%) and slightly all WordNet nodes (by 2%). However, compared to plain one-vs-all the flat top-5 error increases by 11% and even the hierarchical error by 1.8. The main reason for this less discriminant hierarchical classifier seems to be that this approach uses all classifiers but the one trained for the specific class to be detected.

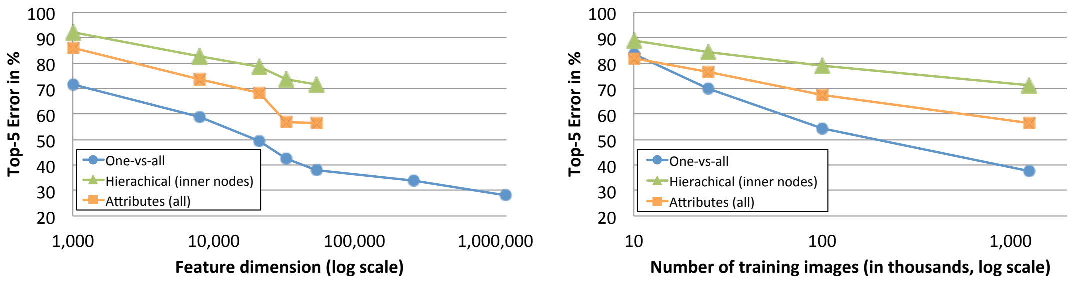
Approach	Top 5 Error	Top 1 Error
1. One-vs-all		
(=leaf WordNet nodes)	37.6 (2.91)	57.2 (5.77)
2. Hierarchical		
inner WordNet nodes	71.3 (7.31)	90.7 (8.69)
all WordNet nodes	50.4 (5.49)	67.9 (7.54)
leaf nodes, cost sensitive	48.6 (4.71)	60.2 (5.66)
SVM stacking, all nodes	36.8 (2.84)	56.3 (5.59)
3. Attributes		
Wikipedia	63.7 (5.21)	81.5 (8.52)
Yahoo Holonyms	68.7 (5.61)	87.1 (9.24)
Yahoo Image	74.0 (5.80)	90.6 (10.28)
Yahoo Snippets	67.2 (5.33)	84.6 (8.55)
all attributes	56.4 (4.63)	75.9 (7.32)
SVM stacking, all attributes	43.8 (3.38)	63.5 (6.34)

Table 4.3: Large scale knowledge sharing results. Shown is flat error in % and hierarchical error in brackets.

The last line of Table 4.3 part 2 gives the results for a stacking-based combination of inner and leaf node classifiers. We use a SVM (MeanSGD) stacked on top of the scores of all nodes and both features to learn the relative importance of the nodes, i.e. we learn one-vs-all classifiers which use the classifier scores as feature vectors. In contrast to the previous hierarchical approaches the trained SVM now correctly attenuates the influence of weak (inner) nodes and achieves a top-5 error of 36.8% which is even slightly better than one-vs-all.

Table 4.3 part 3 gives results for attribute-based models using different semantic relatedness measures for determining object class-attribute associations. On average, using single measures (Wikipedia, Yahoo Holonyms, Image, or Snippets) performs in the same order of magnitude as inner WordNet nodes. When combining all attribute-classifiers from the different measures we improve performance by more than 10% to 56.4% top-5 error (15% lower than inner WordNet nodes). However, this cannot compete with the hierarchical approaches including the discriminative leaf nodes.

In the same fashion as for all WordNet nodes we can also stack a SVM on top of the different attribute classifiers to learn an optimal weighting between them. This results in a significant reduction in error by 13% to 43.8% top-5 error, which is, however, still 6% higher than one-vs-all or 7% higher than the stacked hierarchical approach.



(a) Error vs. number of feature dimensions (for details see Table 4.2)

(b) Error vs. number of training images.

Figure 4.4: Error vs. (a) feature dimensionality and (b) number of training images.

4.5.1 Influence of feature representation and amount of training data

In this experiment we further analyze the dependency with respect to the number of feature dimensions and the amount of available training data. In addition to one-vs-all we pick the best approach for both knowledge transfer settings which is not based on one-vs-all leaf nodes: *inner WordNet nodes* for hierarchical setting and *all attributes*.

In Figure 4.4(a) we plot the error versus the feature dimensionality of the approaches listed in Table 4.2. We observe that for all approaches the performance increases logarithmically with increased feature dimension. From the SIFT representation (1,000 dimensional) to the combined LLC and Fisher vector (53,768 dimensional) the error decreases the most for one-vs-all by 34%, but still strongly by 29% for attributes and 21% for inner WordNet nodes. The relative performance difference between the approaches remains mainly stable across the different features representations which indicates that relative results of the approaches are independent of a specific feature representation.

In Figure 4.4(b) we show results for a reduced amount of training data per class to 10, 25, and 100 samples. The first observation is that the hierarchical and the attribute-based knowledge transfer schemes degrade less (17% and 25%, respectively) than the one-vs-all (46%) scheme. However, the relative ordering remains the same for 100 and 25 samples per class. Only for the rather extreme case of only 10 training samples the attribute-based approach slighly outperforms one-vs-all classification by 1.7%.

4.5.2 Summary

We conclude that the benefit of knowledge transfer is in fact limited for this knowledge sharing and standard multiclass classification setting and becomes apparent only in the stacking-based approaches. In case of limited feature representation or reduced training data the absolute performance differences between the approaches

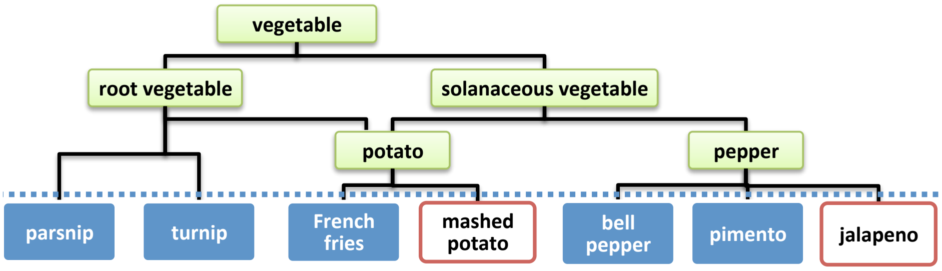


Figure 4.5: Zero-shot recognition using hierarchies. Unseen object classes (red) *mashed potato* / *jalapeno* can be recognized using neighboring leaf node (*French fries* / *bell pepper*, *pimento*), inner node (*potato* / *pepper*), or all (the respective unions) classifiers.

decrease, but one-vs-all remains among the best. The hierarchical based approaches only show reasonable performance when leaf nodes are included. As concerns attribute-based approaches, we observe that using all attribute-classifiers based on multiple semantic relatedness measures significantly improves performance.

4.6 LARGE-SCALE ZERO-SHOT RECOGNITION

In this section, we apply the knowledge transfer approaches of Section 4.3 to a zero-shot recognition setting, in which the sets of object classes of training and test are *disjoint*. We hence denote training object classes as *known*, and test classes as *unseen*. In order to solve the zero-shot recognition task, knowledge obviously has to be transferred between training and test classes. Lampert *et al.* (2009) provided a first benchmark for zero-shot recognition in the form of the Animals-with-Attributes (AwA) data set, consisting of approximately 30,000 images, divided into 40 known animal classes for training and 10 unseen animal classes for testing. In the present experimental study, we lift zero-shot recognition to another level both in terms of data set scale and diversity, by applying it to almost two orders of magnitude more images. In particular, we divide the ILSVRC10 data set randomly into two disjoint sets of object classes, one assumed known (800 classes), and one assumed unseen (200 classes). In all experiments, we further maintain the original split into training and test data defined by the ILSVRC10 data set, meaning that we train on the known (800 class) fraction of the original training set (1,005,761 images), and test on the unseen (200 class) fraction of the original test set (30,000 images).

4.6.1 Results

Table 4.4 gives results for zero-shot recognition, comparing hierarchical (part 1), attribute-based (part 2), and direct similarity-based (part 3) models. In analogy to Table 4.3, the table further distinguishes among hierarchical models using leaf, inner,

	On 200 unseen classes			
Approach	Top-5 Error		Top-1 Error	
1. Hierarchical				
leaf WordNet nodes	72.8	(4.72)	91.3	(11.73)
inner WordNet nodes	66.7	(4.20)	88.7	(11.16)
all WordNet nodes	65.2	(4.10)	88.4	(11.24)
2. Attributes				
Wikipedia	80.9	(5.17)	94.5	(11.69)
Yahoo Holonyms	77.3	(4.91)	94.0	(12.56)
Yahoo Image	81.4	(5.19)	95.5	(12.53)
Yahoo Snippets	76.2	(4.87)	93.3	(11.53)
all attributes	70.3	(4.57)	90.4	(11.62)
3. Direct Similarity				
Wikipedia	75.6	(5.20)	91.8	(11.28)
Yahoo Web	69.3	(4.49)	89.7	(11.10)
Yahoo Image	72.0	(4.60)	90.7	(11.26)
Yahoo Snippets	75.5	(4.89)	91.6	(11.27)
all measures	66.6	(4.41)	88.4	(10.65)

Table 4.4: Large scale zero-shot recognition results. Flat error in % and hierarchical error in brackets.

and all hierarchy nodes, as well as among different semantic relatedness measures for attribute-based and direct similarity-based models. As the relative ranking of the methods is nearly identical between the different error measures (top-5, top-1, flat and hierarchical error) we use the flat top-5 error as the basis for our discussion.

On average, we observe a significant amount of error across the compared approaches. We stress that this can be expected, since the zero-shot recognition task is of considerable difficulty, and cannot be solved without transferring knowledge between potentially unrelated object classes.

Examining the performance of the hierarchical methods (Table 4.4 part 1) we observe a top-5 error of 72.8% using leaf WordNet nodes only. This is the closest setting examined here to one-vs-all classification. It uses the WordNet hierarchy to identify the most similar known leaf node classes for an unseen test class (see Figure 4.5). Using the inner WordNet nodes only, the performance improves to a top-5 error of 66.7%. This is remarkable, since, in comparison to leaf node classifiers, only far fewer and less specific inner node classifiers are used. Furthermore it is in contrast to results in the knowledge sharing experiment (using all classes for training) where performance drops for inner nodes (see Table 4.3): while we benefit from knowledge transfer through the inner nodes for zero-shot recognition, we are loosing precision compared to one-vs-all when sharing knowledge in the inner nodes. The error can slightly be reduced to 65.2% using all WordNet nodes, effectively combining the two previous settings.

Part 2 of Table 4.4 shows the results for attributed-based models using the fully unsupervised mining of both attribute inventory and object class-attribute associations. Overall the obtained error rates for the individual relatedness measures are not competitive to the ones obtained by the hierarchical models. Yahoo Snippets performs best with 76.2% top-5 error. However, when combining all attribute measures we achieve a top-5 error of 70.3% which lies between the performance of leaf and inner WordNet nodes.

On the other hand, the direct similarity-based models reported in part 3 of Table 4.4 obtain as low as 69.3% top-5 error for Yahoo Web and competitive 66.6% when combining the classifiers of all measures, which is only slightly worse than the best performance obtained by a hierarchical method (all WordNet nodes with 65.2%).

The slightly favorable role of direct similarity compared to attribute-based models is consistent with our previous findings (Rohrbach *et al.*, 2010). It can be explained by both the limited quality of the automatically mined part attribute inventory and by having one vs. two potential sources of introducing label noise into the system by means of semantic relatedness (mined object class-attribute associations).

The strong performance of hierarchical models can be attributed to the increased amount of supervision given by the hierarchy, while the attribute- and direct similarity-based models are fully unsupervised.

4.7 CONCLUSION

This chapter explored knowledge transfer in a truly large-scale setting, going far beyond experimental studies of prior work in knowledge transfer w.r.t. data set scale, diversity, and range of tested methods. Our evaluation is based on the large ImageNet challenge (ILSVRC10, Berg *et al.*, 2010a) and includes three prominent approaches to knowledge transfer.

For the fully supervised knowledge sharing experiment, the hierarchical approach using the inner or all node classifiers obtained inferior performance to the leaf nodes only, corresponding to the one-vs-all classifiers. Only when learning a stacked one-vs-all SVM on top, the hierarchical approach could slightly surpass performance of the one-vs-all classifiers. In the zero-shot recognition setting however, the hierarchical approaches obtained overall best performance of the explored knowledge transfer methods.

The attribute based knowledge transfer methods, in their fully unsupervised incarnation as explored in this chapter, consistently produced higher error rates than the hierarchy and direct similarity-based knowledge transfer methods. As pointed out before this reduced performance can be – at least partly – explained by the limited nature of attributes used here that were restricted to automatically mined part attributes. It remains an open research question how to obtain an inventory of representative and descriptive attributes for this kind of approach.

The direct similarity based knowledge transfer method performed on a similar level as the hierarchical methods. This is remarkable as this approach is fully unsupervised using semantic relatedness to automatically find the most related known classes. This is in contrast to the hierarchical methods that require additional information given as a hierarchy.

A DATABASE FOR FINE GRAINED ACTIVITY DETECTION OF COOKING ACTIVITIES 5

Contents

After we focused on visual object recognition in images in the previous two chapters, we switch to activity recognition in videos with this chapter. This chapter sets the basis for the following chapters by introducing a novel database of cooking activities and evaluating different activity approaches on it. Based on the techniques introduced in this chapter we use script data to improve recognition of composite cooking activities in Chapter 6 and look at grounding activity descriptions in videos in Chapter 8. Finally, in Chapter 9, we show how to automatically generate natural language sentences for cooking activities.

Moving to the challenging problem of fine-grained activity recognition, we propose a novel database of 65 cooking activities, continuously recorded in a realistic setting. Activities are distinguished by fine-grained body motions that have low inter-class variability and high intra-class variability due to diverse subjects and ingredients. We benchmark two approaches on our dataset, one based on articulated pose tracks and the second using holistic video features. While the holistic approach outperforms the pose-based approach, our evaluation suggests that fine-grained activities are more difficult to detect and the body model can help in those cases. Providing high-resolution videos as well as an intermediate pose representation we hope to foster research in fine-grained activity recognition.

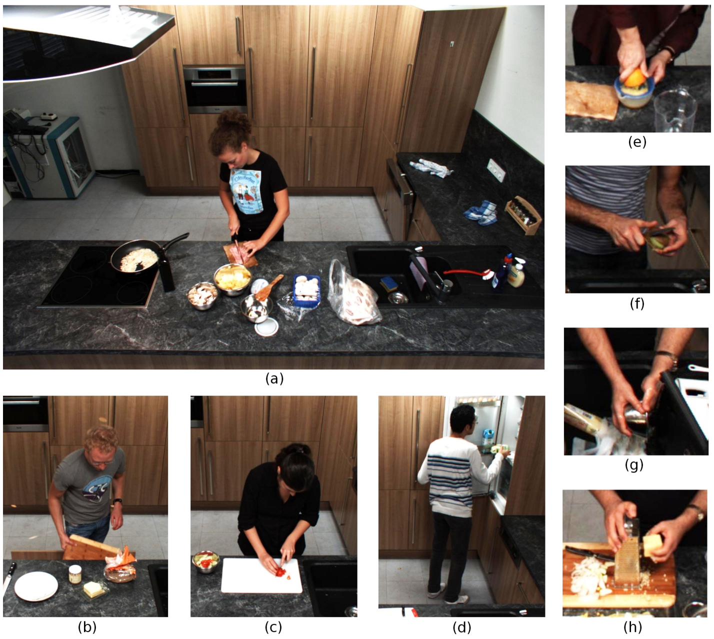


Figure 5.1: Fine grained cooking activities. (a) Full scene of *cut slices*, and crops of (b) *take out from drawer*, (c) *cut dice*, (d) *take out from fridge*, (e) *squeeze*, (f) *peel*, (g) *wash object*, (h) *grate*

5.1 INTRODUCTION

Human activity recognition has gained a lot of interest due to its potential in a wide range of applications such as human-computer interaction, smart homes, elderly/child care, or surveillance. At the same time, activity recognition still is in its infancy due to the many challenges involved: large variety of activities, limited observability, complex human motions and interactions, large intra-class variability vs. small inter-class variability, etc. Many approaches have been researched ranging from low level image and video features (Chakraborty *et al.*, 2011; Laptev, 2005; Wang *et al.*, 2011), over semantic human pose detection (Singh and Nevatia, 2011), to temporal activity models (Gehrig *et al.*, 2009; Niebles *et al.*, 2010; Sia *et al.*, 2011).

While impressive progress has been made, we argue that the community is still addressing only part of the overall activity recognition problem. When analyzing current benchmark databases, we identified three main limiting factors. First, many activities considered so far are rather coarse-grained, i.e. mostly full-body activities, e.g. *jumping* or *waving*. This appears rather untypical for many application domains where we want to differentiate between more fine-grained activities, e.g. *cut* (Figure 5.1a) and *peel* (Figure 5.1f). Second, while datasets with large numbers of activities exist, the typical inter-class variability is high. This seems rather unrealistic for many applications such as surveillance or elderly care where we need to differentiate between highly similar activities. And third, many databases address the problem of activity *classification* only without looking into the more challenging and clearly more realistic problem of activity detection in a continuous data stream. Notable exceptions exist (see Section 5.2) even though these have other limitations such as small number of classes.

This chapter therefore proposes a new activity dataset that aims to address the above three shortcomings. More specifically we propose a dataset that contains 65 activities that are for the most part fine-grained, where the inter-class variability is low, and that are recorded continuously so that we can evaluate both classification and detection performance. More specifically, we consider the domain of recognizing cooking activities where it is important to recognize small differences in activities as shown in Figure 5.1, e.g. between *cut* (Figure 5.1a) and *peel* (Figure 5.1f), or at an even finer scale between *cut slices* (5.1a) and *cut dice* (5.1c).

The contribution in this chapter is twofold: First, we introduce a novel dataset which distinguishes 65 fine-grained activities. We propose a classification and detection challenge together with appropriate evaluation criteria. The dataset includes high resolution image and video sequences (jpg/avi), activity class and time interval annotations, and precomputed mid level representations in the form of precomputed pose estimates and video features. We also provide an annotated body pose training and test set. This allows to work on the raw data but also on higher level modeling of activities. Second, we evaluate several video descriptor and activity recognition approaches. On the one hand we employ a state-of-the-art holistic activity descriptor based on dense trajectories proposed by Wang *et al.* (2011, 2013) using a trajectory description, HOG and HOF (Laptev *et al.*, 2008), and MBH (Dalal *et al.*, 2006). On the other hand we propose two approaches based on body pose tracks, motivated from work in the sensor-based activity recognition community (Zinnen *et al.*, 2009). From the experimental results we can conclude that fine grained activity recognition is clearly beyond the current state-of-the-art and that further research is required to address this more realistic and challenging setting.

5.2 RELATED WORK

We first discuss related datasets for activity recognition, and then related approaches to the ones benchmarked on our dataset. Aggarwal and Ryoo (2011) give an extensive survey of the field.

5.2.1 Activity Datasets

Even when excluding single image action datasets such as the Stanford-40 Action Dataset (Yao *et al.*, 2011) or the Pascal Action Classification Challenge (Everingham *et al.*, 2011), the number of proposed activity datasets is quite large (Ahad *et al.* (2011) list over 30 datasets). Here, we focus on the most important ones with respect to database size, usage, and similarity to our proposed dataset (see Table 5.1). We distinguish four broad categories of datasets: full body pose, movie, surveillance, and assisted daily living datasets – our dataset falls in the last category.

	cls,	classes	clips:	sub-	#	reso-
Dataset	det		videos	jects	frames	lution
Full body pose datasets						
KTH (Schuldt *et al.*, 2004)	cls	6	2,391	25	≈200,000	160x120
USC gestures (Natarajan *et al.*, 2008)	cls	6	400	4		740x480
MSR action (Yuan *et al.*, 2009)	cls,det	3	63	10		320x240
Movie datasets						
Hollywood2 (Marszalek *et al.*, 2009)	cls	12	1,707:69			
UCF50[3]	cls	50	>5,000			
HMDB51 (Kuehne *et al.*, 2011)	cls	51	6,766			height:240
ASLAN (Kliper-Gross *et al.*, 2012)	cls	432	3,631:1,571			
Coffee and Cigarettes (Laptev'07)	det	2	264:11			
High Five (Patron-Perez *et al.*, 2010)	cls,det	4	300:23			
Surveillance datasets						
PETS 2007 (Ferryman, 2007)	det	3	10		32,107	768x576
UT interaction (Ryoo *et al.*, 2009)	cls,det	6	120	6		
VIRAT (Oh *et al.*, 2011)	det	23	17			1920x1080
Assisted daily living datasets						
TUM Kitchen (Tenorth *et al.*, 2009)	det	10	20	4	36,666	384x288
CMU-MMAC (de la Torre *et al.*, 2009)	cls,det	>130		26		1024x768
URADL (Messing *et al.*, 2009)	cls	17	150:30	5	≤ 50,000	1280x720
Our database	cls,det	65	5,609:44	12	881,755	1624x1224

Table 5.1: Overview of activity recognition datasets: We list if datasets allow for classification (cls), detection (det); number of activity classes; number of clips extracted from full videos (only one listed if identical), number of subjects, total number of frames, and resolution of videos. We leave fields blank if unknown or not applicable.

The full body pose datasets are defined by actors performing full body actions. KTH (Schuldt *et al.*, 2004), USC gestures (Natarajan and Nevatia, 2008), and similar datasets (Singh and Nevatia, 2011) require classifying simple full body and mainly repetitive activities. The MSR actions (Yuan *et al.*, 2009) pose a detection challenge limited to three classes. In contrast to these full body pose datasets, our dataset contains more and in particular fine-grained activities.

The second category consists of movie clips or web videos with challenges such as partial occlusions, camera motion, and diverse subjects. UCF50[3] and similar (Liu *et al.*, 2009; Niebles *et al.*, 2010; Rodriguez *et al.*, 2008) datasets focus on sport activities. Kuehne *et al.*'s evaluation suggests that these activities can already be discriminated by static joint locations alone (Kuehne *et al.*, 2011). Hollywood2 (Marszalek *et al.*, 2009), HMDB51 (Kuehne *et al.*, 2011), and ASLAN (Kliper-Gross *et al.*, 2012) have very diverse activities. Especially HMDB51 (Kuehne *et al.*, 2011) is an effort to provide a large scale database of 51 activities while reducing database bias. Although it includes similar, fine-grained activities, such as *shoot bow* and *shoot gun* or *smile* and *laugh*, most classes have a large inter-class variability and the videos are low-resolution. Kliper-Gross *et al.* (2012) focus on a larger number of activities but with little training data per category. Here the focus is to identify similar videos rather than categorising them. A significantly larger video collection is evaluated during the TRECVID challenge (Over *et al.*, 2012). In the 2012 challenge consisted of 291h of short videos from the Internet Archive (archive.org) and more than 4,000h of multi-media (audio and video) data. The challenge consists of different tasks including semantic indexing and multi-media event recognition of 20 different event categories such as *making a sandwich* and *renovating a home*. Large parts of the data are, however, only available to the participants during the challenge. Although our dataset is easier in respect to camera motion and background, it is challenging with respect to a smaller inter-class variability.

The datasets Coffee and Cigarettes (Laptev and Pérez, 2007) and High Five (Patron-Perez *et al.*, 2010) are different to the other movie datasets by promoting activity detection rather than classification. This is clearly a more challenging problem as one not only has to classify a pre-segmented video but also to detect (or localize) an activity in a continuous video. As these datasets have a maximum of four classes, our dataset goes beyond these by distinguishing a large number of classes.

The third category of datasets is targeted towards surveillance. The PETS (Ferryman, 2007) or SDHA2010[4] workshop datasets contain real world situations form surveillance cameras in shops, subway stations, or airports. They are challenging as they contain multiple people with high partial occlusion. The UT interaction (Ryoo and Aggarwal, 2009) requires to distinguish 6 different two-people interaction activities, such as *punch* or *shake hands*. The VIRAT (Oh *et al.*, 2011) dataset is a recent attempt to provide a large scale dataset with 23 activities on nearly 30 hours of video. Although the video is high-resolution people are only of 20 to 180 pixel height.

[3] http://vision.eecs.ucf.edu/data.html

[4] http://cvrc.ece.utexas.edu/SDHA2010/

Overall the surveillance activities are very different to ours which are challenging with respect to fine-grained body-pose motion.

For the domain of *Assisted daily living (ADL) datasets*, which also includes our dataset, only recently datasets have been proposed in the vision community. The University of Rochester Activities of Daily Living Dataset (URADL) (Messing *et al.*, 2009) provides high-resolution videos of 10 different activities such as *answer phone*, *chop banana*, or *peel banana*. Although some activities are very similar, the videos are produced with a clear script and contain only one activity each. In the TUM Kitchen dataset (Tenorth *et al.*, 2009) all subjects perform the same high level activity (*setting a table*) and rather similar actions with limited variation. Roggen *et al.* (2010) and de la Torre *et al.* (2009) present recent attempts to provide several hours of multi-modal sensor data (e.g. body worn acceleration and object location). But unfortunately people and objects are (visually) instrumented, making the videos visually unrealistic. In the CMU-MMAC dataset (de la Torre *et al.*, 2009) all subjects prepare the identical five dishes with very similar ingredients and tools. In contrast to this our dataset contains 14 diverse dishes, where each subject uses different ingredients and tools in each dish. de la Torre *et al.* also record an egocentric view. Here and similar in (Farhadi *et al.*, 2010a; Fathi *et al.*, 2011; Stein and McKenna, 2013) the camera view mainly shows hands and manipulated cooking ingredients. Also recorded in an egocentric view, Pirsiavash and Ramanan (2012) propose a dataset of 18 diverse daily living activities, not restricted to the cooking domain, recorded in different houses in non-scripted fashion.

Overall our dataset fills the gap of a large database with realistic, fine-grained activities, posing a classification and detection challenge in high resolution video sequences.

5.2.2 Holistic approaches for activity recognition

Most approaches for human activity recognition in video focus on using holistic video features, some use the human body pose as a basis. To create a discriminative feature representation of a video many approaches first detect space-time interest points (Chakraborty *et al.*, 2011; Laptev, 2005) or sample them densely (Wang *et al.*, 2009a) and then extract diverse descriptors in the image-time volume, such as histograms of oriented gradients (HOG) and flow (HOF) (Laptev *et al.*, 2008) or local trinary patterns (Yeffet and Wolf, 2009).

Messing *et al.* (2009) found improved performance by tracking Harris3D interest points (Laptev, 2005). The second of the two benchmark approaches we evaluate (see Section 5.4.2), is based on this idea: Wang *et al.* (2011, 2013) track dense feature points and extract strong video features (HOG, HOF, MBH) around these tracks. They report state-of-the art results on KTH (Schuldt *et al.*, 2004), UCF YouTube (Liu *et al.*, 2009), Hollywood2 (Marszalek *et al.*, 2009), and UCF sports (Rodriguez *et al.*, 2008).

Other directions include template based approaches (Rodriguez *et al.*, 2008) or segmenting the space-temporal data and constructing a graph from this (Brendel and Todorovic, 2011). Another direction is to detect activities with a body-worn camera (Spriggs *et al.*, 2009).

5.2.3 Body pose for activity recognition

Many human activities such as *sitting*, *standing*, and *running* are defined in terms of body poses and their motion. However, compared to the number of holistic approaches there exist still little work on visual activity recognition based on articulated pose estimation, also exceptions exist, including (Ferrari *et al.*, 2008; Singh and Nevatia, 2011; Raptis and Sigal, 2013). Pose-based activity recognition appears to work particularly well for images with little clutter and fully visible people as in the gesture dataset from Singh and Nevatia (2011). Estimates of people poses were also used as auxiliary information for activity recognition in single images (Yang *et al.*, 2010). However, these systems have not shown to be effective in complex dynamic scenes with frequent occlusions, truncation and complex poses. This seems also in line with the recent study of Jhuang *et al.* (2013) who show improved activity recognition using ground truth pose estimates, but when estimating human pose automatically they only show it for fully visible bodies. So far, action recognition in such scenes was addressed only by holistic feature-based methods such as (Laptev *et al.*, 2008) due to the difficulty of reliable pose estimation in the complex real-world conditions.

Sung *et al.* (2011) use depth information from a Kinect to estimate pose (Shotton *et al.*, 2011) and distinguish 12 activities. However, in an initial test we found that the Kinect sensor has difficulties to capture fine grained activities due to limited resolution.

5.3 FINE GRAINED HUMAN ACTIVITY DATABASE

For our dataset of fine grained activities we video recorded participants cooking different dishes. Videos are annotated with activity categories on time intervals and a subset of frames was annotated with human pose.

5.3.1 Database recording

We recorded 12 participants performing 65 different cooking activities, such as *cut slices*, *pour*, or *spice*. To record realistic behavior we did not record activities individually but asked participants to prepare one to six of a total of 14 dishes such as *fruit salad* or *cake* containing several cooking activities. In total we recorded 44 videos with a total length of more than 8 hours or 881,755 frames.

In order to get a variation in activities we always told a participant beforehand to prepare a certain dish (e.g. *salad*), including a set of ingredients (*cucumber, tomatoes, cheese*) and potential tools (*grater*) to use. Instructions were given verbally and frequently participants diverted from the instructions by changing tools, and/or ingredients adding to the variability of the activities. Prior to recording participants were shown our kitchen and places of the required tools and ingredients to feel at home. During the recording participants could ask questions in case of problems and some listened to music. We always start the recording prior to the participant entering the kitchen and end it once the participant declares to be finished, i.e. we do not include the final cleaning process. There was a variety of 14 dishes, namely *sandwich, salad, fried potatoes, potato pancake, omelet, soup, pizza, casserole, mashed potato, snack plate, cake, fruit salad, cold drink*, and *hot drink*. Within these dishes each person used different ingredients resulting in very dissimilar videos, e.g. some participants cooked a packet soup while others prepared it from scratch. Dish preparation time varies from 3 to 41 minutes. For statistics on the activities see Table 5.5. Most participants were university students from different disciplines recruited by e-mail and publicly posted flyers and paid; cooking experience ranging from beginner cookers to amateur chefs.

We recorded in our kitchen (see Figure 5.1(a)) with a 4D View Solutions system using a Point Grey Grashopper camera with 1624x1224 pixel resolution at 29.4fps and global shutter. The camera is attached to the ceiling, recording a person working at the counter from the front. We provide the sequences as single frames (jpg with compression set to 75) and as video streams (compressed weakly with mpeg4v2 at a bitrate of 2500).

5.3.2 Database annotations

Activities were annotated with a two-stage revision phase by 6 people with start and end frame as well as the activity categories (see Table 5.5) using the annotation tool Advene (Aubert and Prié, 2007). The dataset contains a total of 5,609 annotations of 65 activity categories. This includes a background activity for the detection task which is generated automatically for all intervals without any other manual annotation for at least 30 frames (1 second), e.g. because the person is not (yet) in the scene or doing an unusual activity which is not annotated.

A second type of annotation is articulated human pose. A subset of frames has been annotated with shoulder, elbow, wrist, and hand joints as well as head and torso. We have 1,071 frames of 10 subjects for training (5 subjects are from separate recordings). For testing we sample 1,277 frames from all activities with the remaining 7 subjects.

We also provide intermediate representations of holistic video descriptors, human pose detections, tracks, and features defined on the body pose (Section 5.4). We hope this will foster research at different levels of activity recognition.

5.4 APPROACHES

To better understand the state-of-the-art for the challenging task of fine-grained activity recognition we benchmark two approaches on our new dataset. The first

Method	Torso	Head	upper arm r	 l	lower arm r	 l	All
Original models							
CPS (Sapp *et al.*, 2010)	**67.1**	0.0	53.4	48.6	**47.3**	37.0	42.2
FMP (Yang and Ramanan, 2011)	63.9	**72.1**	**60.2**	**59.6**	42.1	**46.7**	**57.4**
PS (Andriluka *et al.*, 2009)	58.0	45.5	50.5	57.2	43.3	38.8	48.9
Trained on our data							
FMP (Yang and Ramanan, 2011)	79.6	67.7	60.7	60.8	50.1	50.3	61.5
PS (Andriluka *et al.*, 2009)	**80.1**	**80.0**	**67.8**	**69.6**	48.9	49.6	66.0
FPS (our model)	78.5	79.4	61.9	64.1	**62.4**	**61.0**	**67.9**

Table 5.2: Comparison of 2D upper body pose estimation methods, percentage of correct parts (PCP).

(Section 5.4.1) uses features derived from an upper body model motivated by the intuition that human body configurations and human body motion should provide strong cues for activity recognition in general but particularly for fine-grained activity recognition. The second (Section 5.4.2) is a state-of-the-art method (Wang *et al.*, 2011, 2013) that has shown promising results on various datasets.

5.4.1 Pose-based approach

The first approach is based on estimates of human body configurations. The purpose of this approach is to investigate the complexity of the pose estimation task on our dataset and to evaluate the applicability of state-of-the-art pose estimation methods in the context of activity recognition.

Although pose-based activity recognition approaches were shown to be effective using inertial sensors (Zinnen *et al.*, 2009), they have not been evaluated when the poses are estimated from monocular images. Inspired by Zinnen *et al.* (2009) we build on a similar feature set, computing it from the temporal sequence of 2D body configurations. In the following we first evaluate the state-of-the-art in 2D pose estimation in the context of our dataset. We then introduce our pose-based activity recognition approach that builds on the best performing method.

5.4.1.1 *2D human pose estimation*

In order to identify the best 2D pose estimation approach we use our 2D body joint annotations (see Section 5.3.2). We compare the performance of three recently proposed methods: the cascaded pictorial structures (CPS) (Sapp *et al.*, 2010), the flexible mixture of parts model (FMP) (Yang and Ramanan, 2011) and the implementation of pictorial structures model (PS) of Andriluka *et al.* (2009). Notice that these methods are designed for generic 2D pose estimation. In particular they do not rely on background subtraction or strong assumptions on the appearance of body limbs (e.g. skin color).

For evaluating these methods we adopt the PCP measure (percentage of correct parts) proposed by Ferrari *et al.* (2008) that computes the percentage of body parts correctly localized by the pose estimation method. A body part is considered to be localized correctly if the predicted endpoints of the part are within half of the part length from their ground-truth positions. We first compare the implementations and pre-trained models made publicly available by the authors. Results are shown in the upper part of Table 5.2. The FMP model performs best, likely due to its ability to handle foreshortening of the body parts that occurs frequently in our data.

To push the performance further we retrain the two best performing models (FPM and PS) on our training set, which results in improvements from 57.4 to 61.5 PCP for the FMP model and from 48.9 to 66 PCP for the PS model (Table 5.2, last column). While demonstrating best results, the PS model is still defined in terms of rigid body parts, which is suboptimal for our task. In order to address that we define a flexible variant of the PS model (FPS) that instead of 6 parts used in the original model, consists of 10 parts corresponding to head, torso, as well as left and right shoulder, elbow, wrist and hand. While overall the extended FPS model improves over PS model only by 1.9 PCP (66.0 PCP for PS models vs. 67.9 PCP for FPS), it improves the detection of lower arms by more than 11 PCP which are most important for fined-grained activity recognition. Based on this comparison we rely on the FPS model in the subsequent steps of our pose-based activity recognition pipeline. Figure 5.2 visualizes several examples of the estimated poses for FPS. Notice that we can correctly estimate poses for a variety of activities and body configurations while maintaining precise localization of the body joints.

To extract the trajectories of body joints, an option is to extend our pose estimation to the temporal domain. However, temporal coupling of joint positions significantly complicates inference and approaches of this kind have only recently begun to be explored in the literature (Sapp *et al.*, 2011b). Moreover, our dataset consists of over 800,000 frames and to deal with this sheer complexity of estimating human poses for this dataset we choose a different avenue which relies on search space reduction (Ferrari *et al.*, 2008) and tracking. To that end we first estimate poses over a sparse set of frames (every 10-th frame in our evaluation) and then track over a fixed temporal neighborhood of 50 frames forward and backward. For tracking we match SIFT features for each joint separately across consecutive frames. To discard outliers we find the largest group of features with coherent motion and update the joint position based on the motion of this group. In order to reduce the search space further we use a person detector (Felzenszwalb *et al.*, 2010) and estimate the pose of the person within the detected region with 50% border around.

This approach combines the generic appearance model learned at training time with the specific appearance (SIFT) features computed at test time. When initialized with successful pose estimates it provides reliable tracks of joints in the local temporal neighborhood (see Figure 5.3).

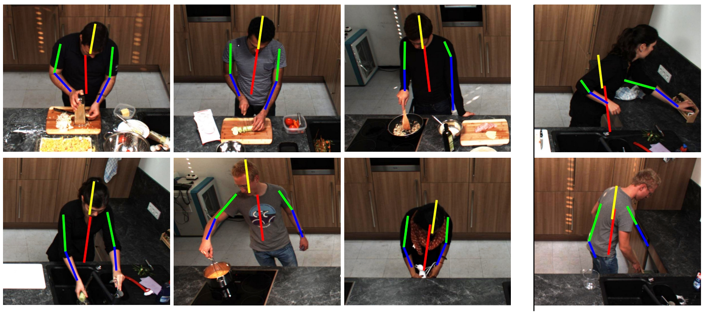


Figure 5.2: Examples of correctly estimated 2D upper body poses (left) and typical failure cases (right).

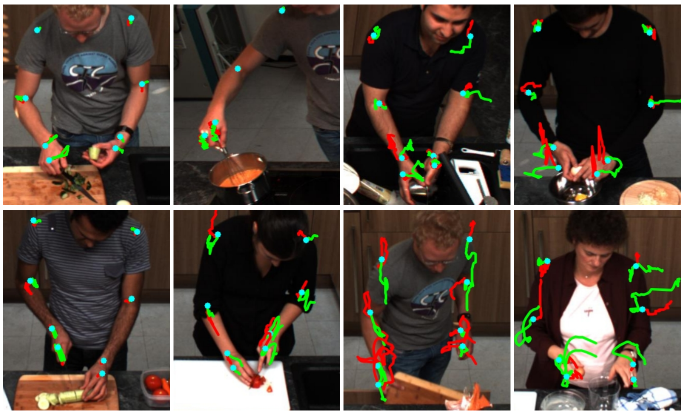


Figure 5.3: Sample tracks for different activities. *Backward tracks in green*, *forward tracks in red* and *initial pose in cyan*. First row, (left to right): *peel*, *stir*, *wash objects*, *open egg*, Second row (left to right): *cut slices*, *cut dice*, *take out from drawer*, *open egg*.

5.4.1.2 *Body model and FFT features*

Given the body joint trajectories we compute two different feature representations: Manually defined statistics over the body model trajectories, which we refer to as *body model features* (BM) and Fourier transform features (FFT) from Zinnen *et al.* (2009) which have shown effective for recognizing activities from body worn wearable sensors.

For the BM features we compute the *velocity* of all joints (similar to gradient calculation in the image domain) which we bin in a 8-bin histogram according to its direction, weighted by the speed (in pixels/frame). This is similar to the approach by Messing *et al.* (2009) which additionally bins the velocity's magnitude. We repeat this by computing *acceleration* of each joint. Additionally we compute *distances* between the right and corresponding left joints as well as between all 4 joints on each body half. For each distance trajectory we compute statistics (mean, median, standard deviation, minimum, and maximum) as well as a rate of change histogram, similar to velocity. Last, we compute the angle trajectories at all inner joints (wrists, elbows, shoulders) and use the statistics (mean etc.) of the angle and angle speed trajectories. This totals to 556 dimensions.

The FFT feature contains 4 exponential bands, 10 cepstral coefficients, and the spectral entropy and energy for each x and y coordinate trajectory of all joints, giving a total of 256 dimensions.

For both features (BM and FFT) we compute a separate codebook for each distinct sub-feature (i.e. velocity, acceleration, exponential bands etc.) which we found to be more robust than a single codebook. We set the codebook size to twice the respective feature dimension, which is created by computing k-means from all features (over 80,000). We compute separately both features for trajectories of length 20, 50, and 100 (centered at the frame where pose was detected) to allow for different motion lengths. The resulting features for different trajectory lengths are combined by stacking and give a total feature dimension of 3,336 for BM and 1,536 for FFT.

5.4.2 Holistic approach

Most approaches for activity recognition are based on a bag-of-words representations. We pick the state-of-the-art dense trajectories approach (Wang *et al.*, 2011, 2013) which extracts histograms of oriented gradients (HOG), flow (HOF Laptev *et al.*, 2008), and motion boundary histograms (MBH Dalal *et al.*, 2006) around densely sampled points, which are tracked for 15 frames by median filtering in a dense optical flow field. The x and y *trajectory* speed is used as a fourth feature. Using their code and parameters which showed state-of-the-art performance on several datasets we extract these features on our data. Following Wang *et al.* (2011) we generate a codebook for each of the four features of 4,000 words using k-means from over a million sampled features.

5.4.3 Activity classification and detection

We train classifiers on the feature representation described in the previous section given the ground truth intervals and labels. We train one-vs-all SVMs using mean SGD as introduced in Section 4.4.4 with a χ^2 kernel approximation (Vedaldi and Zisserman, 2010). While we use ground truth intervals for computing classification results we use a sliding window approach to find the correct interval of a detection. To efficiently compute features of a sliding window we build an integral histogram over the histogram of the codebook features. We use non maximum suppression over different window lengths and start with the maximum score and remove all overlapping windows. In the detection experiments we use a minimum window size of 30 with a step size of 6 frames; we increase window and step size by a factor of $\sqrt{2}$ until we reach a window size of 1800 frames (about 1 minute). Although this will still not cover all possible frame configurations, we found it to be a good trade-off between performance and computational costs.

	Multi-class		per class
Approach	Precision	Recall	AP
Pose-based approaches			
BM	22.1	21.8	27.4
FFT	23.4	22.4	30.4
Combined	**28.6**	**28.7**	**34.6**
Holistic approaches			
Trajectory	35.4	33.3	42.0
HOG	39.9	34.0	52.9
HOF	43.3	38.1	53.4
MBH	44.6	40.5	52.0
Combined	**49.4**	**44.8**	**59.2**
Pose + Holistic	**50.4**	**45.1**	57.9

Table 5.3: Classification results on MPII Cooking Activities, in % (see Section 5.5.1)

5.5 EVALUATION

We propose the following experimental setup for our dataset and include evaluation scripts with the dataset. We have a total of 12 subjects, of which 5 subjects are used to train the body pose model. The remaining 7 subjects are used to perform leave-one-person-out cross-validation. That means that for the 7 cross-validation rounds, training of the activity recognition approaches can use the data from the other 11 subjects.

We report multi-class precision (Pr) and recall (Rc), as well as single class average precision (AP), taking the mean over all test runs. If there is no ground truth label for a certain activity for a given test run (=subject), we ignore this subject when computing mean AP for that particular activity. For detection we use the midpoint hit criterion to decide on the correctness of a detection, i.e. the midpoint of the detection has to be within the groundtruth. If a second detection fires for one groundtruth label, it is counted as false positive. We provide evaluation scripts for comparable results.

5.5.1 Classification results

Table 5.3 summarizes the classification results. The first section of the table shows results for the approaches based on the articulated pose model (see Section 5.4.1), while the second section shows results of the state-of-the-art holistic dense trajectories (Wang *et al.*, 2011) feature representation (see Section 5.4.2). Overall we achieve a mean multi-class recall or accuracy (Table 5.3, second last column), between 21.8% and 45.1% which should be compared to chance level of 1.6% for the 64 classes (we exclude the background class for classification).

We first examine the pose-based approaches. The body model features on the joint tracks (BM) achieve a multi-class precision (Pr) of 22.1%, a recall (Rc) of 21.8% and a mean average precision (AP) of 27.4%. When comparing this to the FFT features, we observe that FFT performs slightly better, improving over BM regarding Pr, Rc, and AP by 1.3%, 0.6%, and 3.0%, respectively. Combining BM and FFT features using stacking (Table 5.3, line 3) yields a significant improvement, reaching 28.6% Pr, 28.7% Rc, and 34.6% AP. We attribute this to the complementary information encoded in the features: While BM encode among others velocity-histograms of the joint-tracks and statistics between tracks of different joints, FFT features encode FFT coefficients of individual joints.

Next we compare the results of the holistic approaches (Sec. 2, Table 5.3) based on dense trajectories (Wang *et al.*, 2011). Trajectory has the lowest performance with 35.4% Pr, 33.3% Rc, and 42.0% AP. In line with results reported by Wang *et al.* (2011) for other datasets HOG, HOF, and motion boundary histograms (MBH) improve over this performance. MBH achieves 44.6% Pr, 40.5% Rc, and 52.0% AP. Combining all holistic approaches again significantly improves performance by more than 4% to 49.4% Pr, 44.8% Rc and 59.2% AP.

It is interesting to note that the pose-based approaches achieve significantly lower performance than the holistic approaches. This may be attributed to the rather sparse joint trajectories of the pose-based approach, while the holistic approach benefits from HOF, HOG, and MBH features around the dense tracks. Additionally we found that pose-estimation does not always give correct results, especially for non-frontal poses or self-occlusion, making the resulting tracks and features fail.

A low-level combination of pose and holistic approaches (Table 5.3, last line) shows slight improvement over the holistic approach (Table 5.3, second last line). We achieve 50.4% multi-class precision, 45.1% multi-class recall (or accuracy), and 57.9% AP (slightly dropped). Although we believe this is an encouraging first result, it shows that fine-grained activity recognition is indeed difficult.

A more detailed class level evaluation based on the confusion matrix (not shown) reveals that fine-grained activities with low inter-class variability are highly confused (e.g. different *cut* activities) while less fine-grained activities such as *wash objects* or *take out from drawer* are hardly confused. This underlines the difficulty of fine-grained activity recognition vs. full- or upper-body activities.

Examining the intermediate representation of 2D tracks we found that the tracks for fine-grained activities *peel* vs. *cut slices* (Figure 5.3, first column) can distinguish fine-grained movements (sideways hand movement vs. vertical movement) highlighting the potential benefit of using body-pose features.

5.5.2 Detection results

Table 5.4 shows detection results and Table 5.5 results per class of the respective combined approaches. Overall performance ranges from the combined pose-based approaches of 17.7% AP (8.6% Pr, 21.3% Rc) over 44.2% AP (17.7% Pr, 40.3% Rc)

Approach	Multi-class Precision	 Recall	per class AP
Pose-based approaches			
BM	6.7	16.1	13.0
FFT	6.3	18.3	15.0
Combined	**8.6**	**21.3**	**17.7**
Holistic approaches			
Trajectory	10.7	25.2	28.4
HOG	15.0	32.2	35.5
HOF	15.1	29.9	36.1
MBH	16.2	37.7	39.6
Combined	**17.7**	**40.3**	**44.2**
Pose + Holistic	**19.8**	40.2	**45.0**

Table 5.4: Detection results on MPII Cooking Activities, in % (see Section 5.5.2)

for the holistic approaches to 45.0% AP (19.8% Pr, 40.2% Rc) when combining pose-based and holistic. The improvements of the combination, similar to the classification results, underlines the complementary nature of the two approaches. Even though overall the performance for the detection task (Table 5.4) is lower than for classification (Table 5.3) the relative performances are similar: pose-based approaches perform below holistic approaches and combining individual approaches improves performance, respectively. In all cases multi-class precision is significantly lower than recall, indicating a high number of false positives. Frequently short activity fragments score very high within other longer fragments or sometimes one ground truth label is fragmented into several shorter ones. We hope this dataset will provide a base for exploring how to best attack these multi-class activity detection challenges.

Table 5.5 provides detailed per-class detection results. We note a general trend when examining the combined pose + holistic approach (Table 5.5, column 5): Fine-grained activities such as *cut apart* (15.7% AP), *screw close* (31.2% AP), or *stir* (52.2% AP) tend to achieve lower performance than less fine-grained activities such as *dry* (94.8% AP), *take out from fridge* (75.5% AP) or *wash objects* (72.2% AP). This underlines our assumption that fine-grained activities are very challenging, which seem to be neglected in many other dataset.

5.6 CONCLUSION

Many different activity recognition datasets have been proposed. However, this new dataset goes beyond previous datasets by posing a detection challenge with a large number of fine-grained activity classes as required for many domains such as assisted daily living. It provides a realistic set of 65 activity classes with low inter-class and large intra-class variability.

We benchmark two approaches on this dataset. The first is based on human body joint trajectories and the second on state-of-the-art holistic features (Wang *et al.*, 2011). Combined they achieve 45.1% mean multi-class recall or accuracy and 57.9% mean average precision on the classification task and 45.0% mean average precision on the detection task. Individually the pose-based approach is outperformed by the dense trajectories which can be attributed to limitations of current articulated pose estimation approaches and the sparser and weaker feature representation. Our analysis of the detection task suggests that especially fine-grained activities are very difficult to detect.

To enable diverse directions of future work on this dataset, we provide the dataset on our website, together with intermediate representations such as body pose with trajectories to allow working on different levels of the problem of fine-grained activity recognition.

		Pose	Hol.	Pose + Holistic		
category	#	AP	AP	AP	Pr	Rc
Background activity	1861.0	31.6	47.1	**48.8**	16.9	85.0
change temperature	72.0	21.1	37.6	**49.4**	7.8	88.9
cut apart	164.0	4.2	**16.0**	15.7	8.4	38.1
cut dice	108.0	10.1	**25.1**	23.8	1.8	5.0
cut in	12.0	0.5	**22.8**	11.1	0.0	0.0
cut off ends	46.0	1.1	**7.4**	6.0	1.3	7.4
cut out inside	59.0	7.3	**16.3**	14.6	5.5	59.5
cut slices	179.0	22.7	**42.0**	39.8	24.8	33.0
cut stripes	45.0	23.1	27.6	**35.9**	23.5	33.3
dry	58.0	44.8	**95.5**	94.8	54.3	96.2
fill water from tap	9.0	67.2	**75.0**	58.3	33.3	66.7
grate	37.0	25.5	32.9	**40.2**	9.0	78.9
lid: put on	20.0	1.6	2.0	**3.5**	0.0	0.0
lid: remove	24.0	0.1	**1.9**	1.7	0.0	0.0
mix	8.0	0.3	**36.8**	35.7	0.0	0.0
move from X to Y	144.0	2.3	**15.9**	13.8	9.7	25.7
open egg	14.0	0.4	**45.2**	27.2	0.0	0.0
open tin	17.0	9.5	**79.5**	79.3	44.4	57.1
open/close cupboard	30.0	25.5	54.0	**54.2**	18.9	38.9
open/close drawer	90.0	6.1	**38.1**	37.9	15.4	37.9
open/close fridge	13.0	62.3	73.7	**73.8**	33.3	87.5
open/close oven	8.0	20.0	25.0	**100.0**	0.0	0.0
package X	22.0	1.2	31.9	**43.0**	0.0	0.0
peel	104.0	42.0	**65.2**	60.7	58.5	37.5
plug in/out	11.0	1.5	54.7	**56.4**	33.3	33.3
pour	88.0	9.3	**54.2**	50.0	16.0	70.9
pull out	7.0	2.4	**87.5**	**87.5**	16.7	75.0
puree	15.0	40.2	**67.1**	65.1	24.2	66.7
put in bowl	215.0	7.9	**18.8**	16.0	3.7	3.1
put in pan/pot	58.0	2.8	15.3	**26.0**	11.8	7.1
put on bread/dough	257.0	14.4	42.1	**42.3**	28.5	30.2
put on cutting-board	94.0	3.0	7.1	**11.6**	8.3	8.6
put on plate	102.0	1.7	**11.0**	6.1	2.2	1.8
read	23.0	1.3	34.5	**49.6**	9.5	25.0
remove from package	46.0	6.3	**39.1**	35.6	10.0	6.7
rip open	17.0	0.3	**5.8**	1.7	0.0	0.0
scratch off	14.0	0.5	**3.8**	2.8	0.0	0.0
screw close	72.0	2.2	**36.3**	31.2	19.4	47.7
screw open	73.0	3.7	19.1	**26.1**	6.9	15.6
shake	94.0	23.7	33.5	**36.7**	18.5	54.2
smell	20.0	0.3	**24.8**	22.4	4.4	15.0
spice	44.0	7.6	29.3	**32.1**	20.0	60.0
spread	24.0	3.6	11.2	**13.9**	50.0	16.7
squeeze	27.0	52.7	**90.0**	89.4	28.6	100.0
stamp	13.0	2.6	**73.3**	70.8	13.5	62.5
stir	95.0	19.0	50.0	**52.2**	18.0	63.2
strew	53.0	11.4	**39.6**	37.8	16.0	10.0
take & put in cupboard	25.0	23.8	37.2	**38.9**	0.0	0.0
take & put in drawer	14.0	0.9	**37.6**	31.8	0.0	0.0
take & put in fridge	30.0	44.2	54.6	**59.2**	31.6	66.7
take & put in oven	9.0	34.5	**100.0**	**100.0**	66.7	66.7
t. & put in spice holder	22.0	28.4	**80.2**	78.6	18.8	46.2
take ingredient apart	57.0	3.3	17.5	**20.7**	3.7	25.6
take out from cupboard	130.0	61.5	**81.5**	70.5	64.8	80.7
take out from drawer	258.0	48.2	**79.7**	70.2	63.0	70.8
take out from fridge	70.0	56.5	73.6	**75.5**	37.3	82.4
take out from oven	7.0	2.1	**83.3**	**83.3**	37.5	100.0
t. out from spice holder	31.0	10.0	67.0	**77.3**	8.5	50.0
taste	21.0	0.9	18.2	**28.8**	28.6	15.4
throw in garbage	87.0	50.0	84.4	**85.9**	43.4	84.6
unroll dough	8.0	0.6	**100.0**	83.3	66.7	66.7
wash hands	56.0	35.7	45.9	**50.6**	41.2	31.1
wash objects	139.0	51.5	67.1	**72.2**	28.8	90.1
whisk	19.0	40.6	**70.0**	60.8	15.2	77.8
wipe clean	20.0	5.5	**10.6**	7.7	5.3	10.0
Mean over all classes	86.3	17.7	44.2	**45.0**	19.8	40.2

Table 5.5: Detection results per class on MPII Cooking Activities, in % (see Section 5.5.2).

Column 2: total number of annotations; columns 3 to 5: AP for (the combined version of) pose-based, holistic, and pose + holistic approaches; column 6,7: multi-class precision and recall for the Combined pose + holistic approach.

6 SCRIPT DATA FOR ATTRIBUTE-BASED RECOGNITION OF COMPOSITE ACTIVITIES

Contents

STATE-OF-THE-ART human activity recognition methods, as discussed in the previous chapter, build on discriminative learning which requires a representative training set for good performance. This leads to scalability issues for the recognition of large sets of highly diverse activities. In this chapter we leverage the fact that many human activities are compositional and that the essential components of the activities can be obtained from textual descriptions or scripts. To share and transfer knowledge between composite activities we model them by a common set of attributes corresponding to basic actions and object participants. This attribute representation allows to incorporate script data that delivers new variations of a composite activity or even to unseen composite activities. In our experiments on 41 composite cooking tasks we found script data to successfully capture the high variability of composite activities. We show improvements in a supervised case where training data for all composite cooking tasks is available, but we are also able to recognize unseen composites by just using script data and without any manual video annotation.

6.1 INTRODUCTION

Human activity recognition in video is a fundamental problem in computer vision. State-of-the-art methods (e.g. Wang *et al.*, 2011; Kovashka and Grauman, 2010; Niebles *et al.*, 2010) achieve near perfect results for simple actions (e.g. KTH dataset, Schuldt *et al.*, 2004) and robustly recognize actions in realistic settings such as Hollywood movies (Marszalek *et al.*, 2009), videos from YouTube (Liu *et al.*, 2009), or sport scenes (Rodriguez *et al.*, 2008).

The top-performing methods typically rely on discriminative machine learning, which requires representative training data. Collecting such training sets is challenging if the number of activities is large and the activities themselves are complex. In consequence, most current research (with few exceptions: Messing *et al.*, 2009; Niebles *et al.*, 2010; Fathi *et al.*, 2011) focuses on simple basic-level activities such as *walking* or *drinking*, while the recognition of longer-term, complex, and composite activities such as *assembling furniture* or *food preparation* has been rarely addressed in computer vision.

A promising approach towards scalability of activity recognition methods to a large number of complex activities is to use intermediate representations that are shared and transferred across activities by exploiting their compositional nature. We exploit this technique and propose a new approach building on an attribute-based representation. Instead of learning a model for each composite activity we learn models for a large set of attributes shared across composite activity classes. Such approaches have been shown effective to recognize previously unseen object categories (Lampert *et al.*, 2009; Rohrbach *et al.*, 2010) and have also been applied to activity recognition (Liu *et al.*, 2011).

We evaluate our approach in the daily living domain where many tasks, such as *cleaning the house* or *preparing a dish*, are composed of several basic-level activities. A major challenge to recognize everyday activities is that these activities can often be performed in a wide variety of ways, and it is practically infeasible to create a training set with all possible alternatives.

For the purpose of this chapter we focus on the recognition of cooking activities, which share many basic-level activities, cooking tools, and ingredients. Our evaluation in Chapter 5 has shown that recognizing basic-level activities is already a challenging task and thus the recognition approach needs to be robust to failures in detection of basic-level activities. In this work we address the challenges of difficult basic-level cooking activities as well as the high variability in composite activities in three complementary ways:

1. We detect activities and objects independently but take their co-occurrence and context into account. E.g. when looking at cooking activities it is likely that *peeling* co-occurs with *carrot* or *potato* but not with *cauliflower*.

2. We model basic-level activities and participants as attributes of composite activities, allowing to easily share and transfer across composite activities. As Figure 6.1 shows a decomposition of the activities *prepare onion*, *separate egg*,

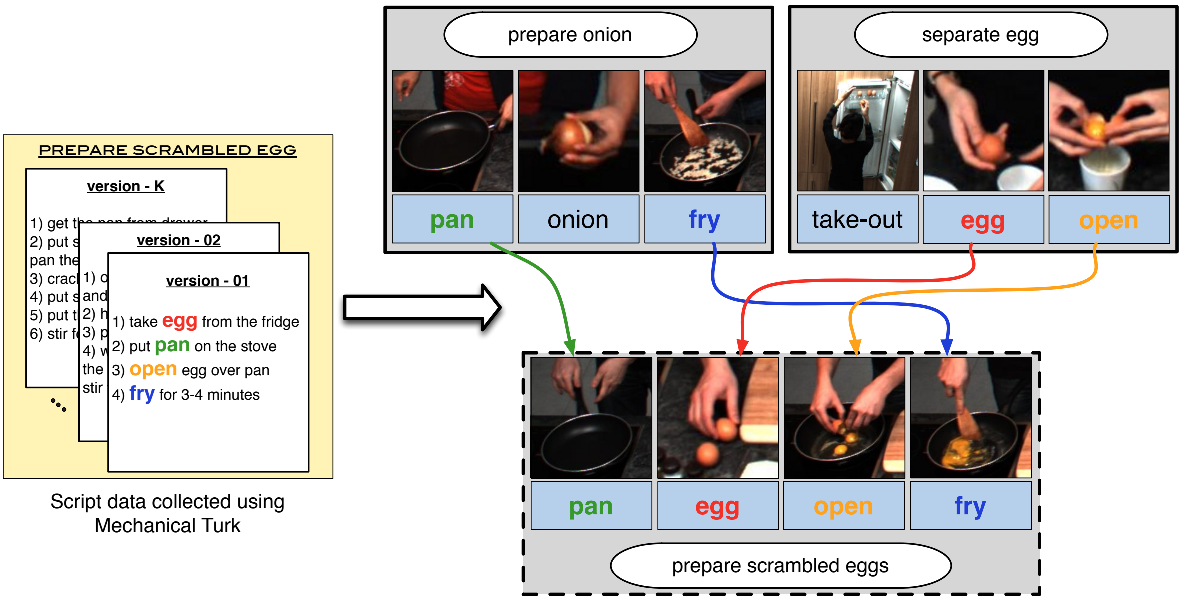


Figure 6.1: Sharing or transferring attributes of composite activities using script data. Composite activities (gray boxes) are composed of basic-level activities and their participants (light-blue boxes), modeled as attributes. These attributes can be transferred with the help of script data to unseen composite activities (dashed-line box).

and *prepare scrambled eggs* into attributes of basic-level activities such as *fry* and *open* as well as their participating ingredients (*egg*) or tools (*pan*).

3. We collect a large number of textual descriptions, instances of so called scripts, for an activity to compute how relevant a certain attribute is for a specific composite activity. Given this script data we can not only handle the variation of composites but also recognize unseen composite activities. As illustrated in Figure 6.1 the attributes *egg*, *pan*, *open*, and *fry* are determined to be important for *preparing scrambled eggs* using script data and can be transferred from known composites such as *separate egg* and *prepare onion*.

Our contributions are as follows. First, we show how to use text-based script data for handling the large variability of composite activity recognition by selecting relevant attributes. Second, we not only improve performance in the supervised case but also can transfer to unseen composite cooking activities. We achieve this by decomposing composite activities into a flexible attribute representation. Third, we show that using co-occurrence and temporal activity context can help recognizing the challenging basic-level activities. Additionally, we release the challenging recorded video dataset (called *MPII Cooking Composite Activities*, or short *MPII Composites*) allowing to evaluate recognition of activity composites and attribute transfer on a large scale.

6.2 RELATED WORK

This chapter addresses the challenging task to recognize complex everyday activities, taking cooking as running example. Our goal is to leverage on the compositional nature of human activities to enable the recognition of activities for which only few or even no training examples are available. This is in contrast to approaches that represent activities as bags of spatio-temporal features (Laptev, 2005; Wang *et al.*, 2011; Kovashka and Grauman, 2010; Chakraborty *et al.*, 2011) disregarding potential structure within the activity.

Several recent approaches (Gupta and Davis, 2007; Niebles *et al.*, 2010; Liu *et al.*, 2011) have aimed at structured representation of activities that go beyond bags-of-features. Joint modeling of actions and objects has been explored by Wu *et al.* (2007) and Gupta and Davis (2007), demonstrating improved performance for both tasks. In this work we also include both actions and objects in our representation, while aiming to recognize more complex interactions and activities. Niebles *et al.* (2010) model activities as a temporal composition of primitive actions and discriminatively learn such models. The primitive actions are learned in a data-driven manner, complicating transfer to previously unseen activities. In contrast to this we focus on semantically meaningful basic-level activities, which permit to learn the relationships between activities and objects from textual sources.

Recent work has shown that attributes are an effective intermediate representation that facilitates cross-task (Lampert *et al.*, 2009) and cross-modal learning (Rohrbach *et al.*, 2010). We build our approach on such an representation using attributes that are commonly shared between cooking activities. The attributes correspond either to basic-level activities such as *stir*, *peel*, or *grate* or to tools and ingredients used in the cooking process. Our representation is conceptually similar to the *object/action bank* representation for scene recognition (Li *et al.*, 2010a), for still image action recognition (Yao *et al.*, 2011), and video action recognition (Sadanand and Corso, 2012). Similar to these, we first train a set of detectors for a large set of attributes and then perform reasoning on top of the detector bank output.

While attributes have been used for object recognition (Ferrari and Zisserman, 2007; Lampert *et al.*, 2009; Farhadi *et al.*, 2010a; Rohrbach *et al.*, 2010) they have only recently been applied to activity recognition (Yao *et al.*, 2011; Liu *et al.*, 2011). Liu *et al.* (2011) build on a set of manually defined attributes describing various body motions such as *raise arms* and *bend torso*. The attributes are interpreted as latent variables and combined with motion trajectory features and attribute co-occurrence features within a latent SVM framework. Liu *et al.* demonstrate the effectiveness to recognize previously unseen activities, but requires manual specification of activity attributes. In contrast to this we put our main focus on investigating how attribute relationships can be automatically mined from text sources.

Language and cross-modal learning have been used for knowledge transfer (Wu *et al.*, 2007; Rohrbach *et al.*, 2010).Wu *et al.* (2007) combine visual and RFID data with common-sense knowledge to learn recognition models of complex kitchen activities. In (Rohrbach *et al.*, 2010) and Chapters 3 and 4 we relied on publicly available databases such as Wikipedia[5], WordNet (Fellbaum, 1998), or Flickr[6] to mine relationship between attributes and objects, and uses them to recognize novel object classes. These methods have not been explored for activity recognition in the past, likely because generic text corpora do not seem suitable for mining activity-related attributes as noted by Liu *et al.* (2011). To address this, we explicitly gather knowledge about activities by collecting their textual descriptions from multiple subjects. We then rely on linguistic analysis of such descriptions in order to compute statistics of the appearance of various attributes within each activity. We demonstrate that such statistics allow to significantly boost recognition performance and also facilitate recognition of previously unseen activities.

Movie scripts associated to a movie have previously been used by Laptev *et al.* (2008) to obtain automatic annotations of activities, in contrast to this we want to capture unseen variations by script data collected independent of the video. In the multimedia community, MediaMill (Snoek *et al.*, 2006) and LSCOM (Hauptmann *et al.*, 2008) are efforts to explore large scale video retrieval using mid-level concepts and exploring combination of textual and visual information.

6.3 MODELING ATTRIBUTES AND COMPOSITE ACTIVITIES

We are interested in two activity recognition tasks: First we would like to recognize different composite activities, such as preparing cucumbers. Secondly, we want to recognize the various activity attributes associated to and making up the composite activity. Those attributes characterize the composite activity and are either basic-level activities (such as *peeling* or *washing*) or the respective participants (such as *grater*, *knife*, or *cucumber*).

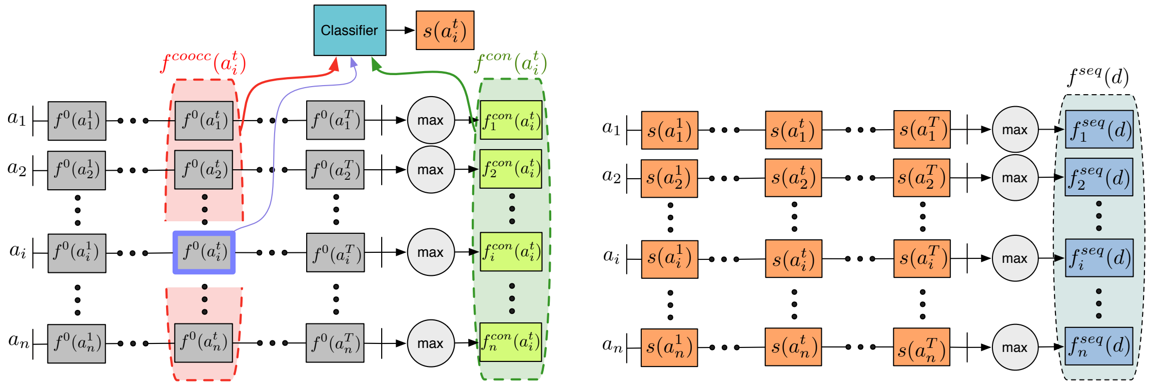
This section first describes the attribute recognition approach that equally applies to basic-level activities and participants (Section 6.3.1). Composite activities are recognized based on these attributes (Section 6.3.2). We then show how to use prior knowledge (Section 6.3.3) to improve the recognition of composite activities, overcoming the notorious lack of training data and handling the large variability of activities.

6.3.1 Recognizing activity attributes using context and co-occurrence

For a time interval t we want to classify if a particular activity attribute a_i is present. As mentioned before a_i can be any attribute including *cut*, *knife*, or *cucumber*. To obtain the final classifier score for an attribute a_i we are proposing to use three different types of features. The first type of feature is given by a video-feature-based attribute classifier providing us with confidence score $f^0(a_i^t)$ for attributes a_i at time interval t. In addition to $f^0(a_i^t)$, we define features based on context (in the same video sequence) as well as features based on the co-occurrence of other attributes (in

[5] http://www.wikipedia.org

[6] http://www.flickr.com



(a) Activity attribute recognition using contextual and co-occurrence attributes.

(b) Composite activity classification using activity attributes.

Figure 6.2: Our approach to recognition of attributes (a) and composite activities (b).

the same time interval t).

Contextual features formalize the intuition that close or adjacent time frames have strongly related attributes: e.g. if a *cucumber* is *peeled* in one time interval, the *cucumber* is probably also present in the surrounding time frames, and it is likely that the same video sequence contains a *cutting* activity as well. More formally (visualized in Figure 6.2(a)) we define a context feature vector $f^{con}(a_i^t)$ as

$$f_j^{con}(a_i^t) = \max_{u=1,\ldots,t-1,t+1,\ldots,T} f^0(a_j^u) \qquad \forall j \in \{1,2,\ldots n\}, \tag{6.1}$$

where n is the total number of attributes. Element j of the context feature vector contains evidence that attribute a_j occurs in the context of attribute a_i.

Similarly, activity attributes happening at the same time instance t are related, e.g. if we *peel* something it is more likely to observe also *carrot* or *cucumber* rather than *cauliflower*. We thus define the co-occurrence by a feature vector $f^{coocc}(a_i^t)$ of all attribute scores excluding a_i^t:

$$f_k^{coocc}(a_i^t) = f^0(a_k^t) \qquad \forall k \in \{1\ldots n\}\backslash i \tag{6.2}$$

Based on these features we train an activity attribute classifier using the features individually or by stacking them (see Figure 6.2(a)). This formulation can be easily extended to other attribute representations depending on the task and available features. In the following, $s(a_i)$ refers to the score of such an attribute classifier.

6.3.2 Composite activity classification using activity attributes

We now want to classify composite activities that span an entire video sequence, given attribute classifier scores $s(a_i^t)$. In this approach we rely on the representation that captures likelihoods of the presence or absence of a particular attribute and leave modeling temporal ordering of attributes for future work. For each sequence d

we build a feature vector $f^{seq}(d)$ by computing the maximum score of each attribute over all time intervals (see Figure 6.2(b)):

$$f_i^{seq}(d) = \max_{t=1,\ldots,T}\{s(a_i^t)\} \qquad \forall i \in \{1, 2, \ldots n\}. \tag{6.3}$$

To decide on the category of a sequence we use the feature representation $f_i^{seq}(d)$ and classify using a nearest neighbor classifier (NN) or support vector machines (SVM) given a set of labeled training sequences. The following section describes the additional incorporation of semantic relatedness to select the relevant attributes a_i, i.e. feature dimensions in $f_i^{seq}(d)$.

6.3.3 Script data for recognizing composite activities

Composite activities show a high diversity which is practically impossible to capture in a training corpus. Our system thus needs to be robust against many activity variants that are not present in the training data. The use of attributes allows to include external knowledge to determine relevant attributes for a given composite activity. For this we assume associations between attribute a_i and composite activity class z in a matrix of (normalized) weights w_i^z. Our system extracts those associations from script data (see Section 6.4), but the approach generalizes to arbitrary other external knowledge sources. We explore two options to use such information, one of which does not require any visual training data of a specific composite activity and thus enables zero-shot recognition.

Script data: To compute a confidence score $s^{scriptdata}(z|d)$ of the composite activity d being of class z we use the attribute based feature representation $f_i^{seq}(d)$. Given the weights w_i^z we compute a weighted sum

$$s^{scriptdata}(z|d) = \frac{\sum_{i=1}^{n} w_i^z f_i^{seq}(d)}{\sum_{i=1}^{n} w_i^z}, \tag{6.4}$$

This formulation is similar to the sum formulation in Chapter 4 Equation (4.5) used for image recognition with attributes, which itself is an adaption of the direct attribute prediction model introduced by Lampert *et al.* (2009). Note that the weight matrix retrieved from script data is sparse (often, $w_i^z = 0$). When mining from other corpora one might need to threshold or cut-off the weights w_i^z to achieve good performance.

NN+script data: When training data is available we can use a nearest neighbor classifier. Often, only a handful of attributes are likely to be indicative for a composite activity class, while the majority of other attributes will provide irrelevant, potentially noisy information. When searching for nearest neighbors such irrelevant attributes might dominate the distance, resulting in suboptimal performance. To reduce this effect we rely on the script data to constrain the attribute feature vector to the relevant dimensions.

More specifically, we replace the distance measure of nearest neighbor with the following training class dependent similarity function, taking weights of class-attribute associations into account. It is defined between the test attribute vector of unseen class $f_i^{seq}(d^{test})$ and the training attribute vector $f_i^{seq}(d_z^{train})$ of class z as

$$Sim(d^{test}, d_z^{train}) = \left(\frac{\sum_{i=1}^{n} w_i^z \left(f_i^{seq}(d^{test}) - f_i^{seq}(d_z^{train}) \right)^2}{\sum_{i=1}^{n} w_i^z} \right)^{0.5} . \tag{6.5}$$

To enhance robustness further, we binarize all association weights w_i^z by setting all non-zero weights to 1. This reduces the distance computation to the relevant attributes, normalized by the total number of relevant attributes. Using continues weights requires their inversion, which performed worse than binarized weights for our purposes.

6.4 MINING SCRIPT DATA

Linguistics and psychology literature knows prototypical sequences of certain activities as so-called *scripts* (Schank and Abelson, 1977; Barr and Feigenbaum, 1981). Scripts describe a certain scenario (e.g. "eating in a restaurant") with temporally ordered events (*the patron enters restaurant, he takes a seat, he reads the menu...*) and participants (*patron, waiter, food, menu,...*). Written event sequences for a scenario can be collected on a large scale using crowdsourcing (Regneri *et al.*, 2010). We make use of this method regarding our composite activities as scenarios and assembling a large number of written sequences for each of those. After a more detailed description of the data collection, we show how to match attribute labels to the text data, and what kind of statistics we use to compute the association weights w_i^z in Equations (6.4) and (6.5).

6.4.1 Data acquisition via crowdsourcing

We collect natural language sequences similar to Regneri *et al.* (2010) using Amazon's Mechanical Turk[7]. For each composite activity, we asked the subjects to give tutorial-like sequential instructions for executing the respective kitchen task. The instructions had to be divided into sequential steps with at most 15 steps per sequence. We select 53 relevant kitchen tasks as composite activities by mining the tutorials for basic kitchen tasks on the webpage "Jamie's Home Cooking Skills"[8]. All those tasks are steps to process ingredients or to use certain kitchen tools. In addition to the data we collected in this experiment, we use data from the OMICS corpus[9] and (Regneri *et al.*, 2010) for 6 kitchen-related composite activities. This results in a corpus with 2124 sequences in sum, having a total of 12958 event descriptions.

[7] http://www.mturk.com

[8] http://www.jamieshomecookingskills.com

[9] http://openmind.hri-us.com/

1.	get a large sharp knife	1.	gather your cutting board and knife.	1.	wash the cucumber
2.	get a cutting board	2.	wash the cucumber.	2.	peel the cucumber
3.	put the cucumber on the board	3.	place the cucumber flat on the cutting board.	3.	place cucumber on a cuttingboard.
4.	hold the cucumber in your weak hand	4.	slice the cucumber horizontally into round slices.	4.	take a knife and rock it back and forth on the cucumber
5.	chop it into slices with your strong hand			5.	make a clean thin slice each time.

Figure 6.3: 3 example scripts for the composite activity *cutting a cucumber*

This dataset provides much more variation than the limited number of video training examples can capture. Of course this poses also a challenge, because we need to overcome the problem of different wordings and coordinated events: Figure 6.3 shows three examples we collected for the composite activity *chopping a cucumber*. They differ in verbalization (cf. *slice*, *chop*, and *make a slice*) and granularity (*getting* something is often left out). Further, the sequences reflect different ways of preparing the vegetable, some include *peeling* it, some do not *wash* it, and so on. Some sentences contain conjugated events (*take a knife and rock it...*). While we clean the data to a certain degree by fixing spelling mistakes and resolving pronouns with the method from Bloem *et al.* (2012), we end up with both challenges and blessings of a very noisy but very big training data set.

6.4.2 Data analysis

To use the prior knowledge from the textual data, we match the attribute labels from the video annotations to the written script instances and compute several statistics: the frequency distribution give simple priors of single attributes, and tf*idf is used to find the most salient composite activity associated with certain basic-level attributes.

6.4.2.1 *Label matching*

To transfer any kind of knowledge from the script corpus to the attributes from the video annotation, we need to match attribute labels to language descriptions. The annotated attribute labels are standard English verbs (for activities, e.g. "wash") and nouns (for participating objects, e.g. "carrot"), sometimes with additional particles (e.g. "take apart" and "take out"). Because the script instances contain unrestricted natural language sentences, they do not necessarily have any correspondence with the attribute label annotations, thus we evaluate two ways of mapping between them:

- **literal**: we look for the exact matching of the attribute label within the data.
- **WordNet**: we look for attribute labels and their synonyms. We take synonyms

as members of the same *synset* according to the WordNet ontology (Fellbaum, 1998) and restrict them to words with the same part of speech, i.e. we match only verbal synonyms to activity predicates and only nouns to object terms.

6.4.2.2 *Statistics computed on the data*

We compute two different association scores between attribute labels and composite activities:

- **Freqs**: frequency distribution over all attribute labels for each composite activity.
- **tf∗idf** (term frequency ∗ inverse document frequency, Salton and Buckley, 1988) is a measure used in Information Retrieval to determine the relevance of a word for a document. Given a document collection $D = d_1, ..., d_n$, tf∗idf for a term (or word) w and a document d_i is computed as follows:

$$tf * idf(w, d_i) = freq(w, d_i) * log \frac{|D|}{|d_{w \in d}|} \tag{6.6}$$

  $d_{w \in d}$ is the set of documents containing w at least once. tf∗idf represents the distinctiveness of a term for a document: the value increases if the term occurs often in the document and rarely in other documents. In our case, one document corresponds to one composite activity, i.e. it contains all sequences collected for the same scenario.

We normalize the association scores for each composite activity over all attributes which gives the association weights used in Equations (6.4) and (6.5).

6.5 EXPERIMENTAL SETUP

This section first describes our new MPII Cooking Composite Activities dataset (MPII Composites) that is publicly available on our webpage. We then outline the experimental setup for the evaluation (Section 6.6).

6.5.1 MPII Cooking Composite Activities dataset

To evaluate composite activity recognition, we record a dataset containing different cooking activities. We discard some of the composite activities in the script corpus (Section 6.4) which are either too elementary to form a composite activity (e.g. *how to secure a chopping board*), or were duplicates with slightly different titles, or because of limited availability of the ingredients (e.g. *butternut squash*). This resulted in 41 composite cooking activities for evaluation.

The dataset recording setup is identical to Chapter 5 and, similarly, we do not tell subjects how to perform a certain cooking task. In Table 6.1 we compare MPII

	videos	subjects	categories		ground truth	attribute	video
			composites	attributes	time intervals	instances	duration
MPII Cooking	44	12	-	218	3,824	15,382	3-41 min
MPII Composites	212	22	41	218	8,818	33,876	1-23 min
combined	256	30	41	218	12,642	49,258	1-41 min

Table 6.1: Dataset statistics.

Cooking introduced in Chapter 5 and the extension MPII Composites proposed in this chapter. Recordings are made with 1624x1224 pixel resolution, with 29.4fps, recording a person at the counter from the front. We use the same annotation protocol as in Chapter 5 , but additionally distinguish participants of an activity (*cut*), namely ingredients (*carrot*), tools (*knife*), and containers (*cutting board*), for both datasets.

6.5.2 Video representation and evaluation protocol

We use a bag-of-features representations which uses HOG, HOF, and motion boundary histograms around densely sampled points, which are tracked for 15 frames by median filtering in a dense optical flow field (Wang *et al.*, 2011). This feature showed best performance on MPII Cooking (see Chapter 5). The feature extraction and training is identical to Chapter 5, i.e. we generate a codebook using k-means and train the attribute classifiers using one-vs-all SVMs trained by meanSGD (see Section 4.4.4) with a χ^2 kernel approximation (Vedaldi and Zisserman, 2010). We generate the codebook only from MPII Cooking, generating a true zero-shot setting when transferring to MPII Composites.

Recordings from subjects which appear in MPII Cooking are only used for training. The data of all remaining 17 subjects are divided into 6 cross-validation-splits. We report mean average precision (AP), taking the mean over all classes and cross validation rounds. If a class is not present in a cross-validation round, we exclude it from mean computation for this round.

In all evaluation runs for both attributes and composites, we use the same cross-validation procedure and we always evaluate on MPII Composites. Concerning training, we distinguish two settings: First we train attributes on both datasets (left columns, Tables 6.2 and 6.3). To see how well attributes can be transferred, we also train attribute classifiers only on MPII Cooking (right columns). In the SVM case, composites are trained using meanSGD on the attribute classifier output score vector $f^{seq}(d)$.

6.6 EVALUATION

In this section we first evaluate our attributes enhanced with context and co-occurrence, and then evaluate recognition of composite cooking activities using

Attribute Training on:	MPII Cooking + MPII Composites	MPII Cooking
Base (f^0)	32.3	18.4
Context only (f^{con})	13.1	10.1
Base+Context	34.2	13.3
Co-occurrence only (f^{coocc})	27.3	20.3
Base+Co-occurrence	30.9	21.5
Base+Context+Co-occurrence	37.7	17.3

Table 6.2: Attribute recognition using context and co-occurrence, AP in %

different levels of supervision, including a zero-shot approach using script data.

6.6.1 Attribute recognition using context and co-occurrence

Table 6.2 summarizes the results for recognizing activities and their participants, modeled as attributes. For a certain time window, multiple attributes can be activated, e.g. because a person is *mixing* a *salad* with *fork* and *spoon* in a *bowl*, resulting in 5 attributes activated at the same time.

The left column of Table 6.2 shows the results for training on both, MPII Cooking and MPII Composites, but evaluating on MPII Composites only. The performance of the base classifier trained on the dense trajectory feature representation achieves 32.3% mean average precision (AP) for the 218 attribute classifiers on MPII Composites.

Using only temporal context to recognize activity attributes performance drops significantly (13.1% AP). This is the expected result, because the context is similar for all activities of the same sequence and thus cannot discriminate attributes. In contrast, when using co-occurrence only, the performance drops only by 5.0% compared to the base classifiers due to the high relatedness between the attributes, namely between activities and their participants.

Combining context and co-occurrence information with the base classifier gives 34.2% and 30.9%, respectively. This is below the base classifier's performance for co-occurrence, but a combination of all training modes achieves a performance of 37.7% AP, improving the base classifier's result by 5.4%.

In a second setting, we restrict the training dataset to MPII Cooking but still evaluate on MPII Composites (right column of Table 6.2), requiring the activity attributes to transfer to different composite activities. When comparing the right to the left column, we notice a significant performance drop for all classifiers. This decrease can mainly be attributed to the strong reduction of training data to about one third. Co-occurrence and Base+Co-occurrence achieve the best results with 20.3% and 21.5% accuracy. Co-occurrence stand out compared to the other individual attribute classifiers: Because the activity context changes from MPII Cooking to MPII Composites (having different composite activities), context leads to tremendous performance drops in all combinations.

6.6.2 Composite cooking activity classification

After evaluating attribute recognition performance in Section 6.6.1, we now show the results for recognizing composites using the attributes as described in Section 6.3.2. We only use the combination of base, context, and co-occurrence. Although this is not the best choice for recognizing attributes for the attribute transfer setting we found it to work better or similar to alternatives for composite recognition.

The results are shown in Table 6.3, which, similar to Table 6.2, shows results for training the attributes on both, MPII Cooking and MPII Composites, on the left and reduced attribute training on MPII Cooking only on the right. In the first (top) section of the table we use MPII Composites as training data for the composite cooking activities with 6-fold cross-validation as done before. For training of composite activities, we are limited to MPII Composites, because MPII Cooking is not structured into different composite cooking activities. In the second (bottom) section of the table we use *n*o training data for the composite cooking activities, often referred to zero-shot learning. This is enabled by the use of script data as motivated before.

The results in the top left quarter of Table 6.3 show the fully supervised setup. The first setup uses an SVM trained directly on the video feature representation rather than basic level attributes. This is the same setup as in Chapter 5 and as our Base (f^0) classifier, but this time trained and tested on complete composite activity videos. It achieves 38.4% AP, showing how challenging the dataset is. However, an SVM, trained on the attribute feature vectors (f_i^{seq}), achieves 51.2% AP, while NN classification reaches slightly better performance of 51.7%. This demonstrates that our attribute representation is a good way model for the video. To restrict NN to relevant attributes, we reduce the feature vector using script data (see Section 6.3.3). We distinguish four options: The first two use normalized frequency counts, while the third and fourth use tf*idf to determine the relevance of an attribute for a given composite. For both we mine words in the collected scripts either literally or using a WordNet (WN) expansion (see Section 6.4 for details). We first notice that tf*idf for WN (53.9%) outperforms the purely training data based methods SVM and NN. tf*idf obviously selects the right attributes for a given composite activity, making the problem of finding the nearest neighbor simpler. In comparison to the frequency counts (50.9% and 51.2%), tf*idf performs slightly better, because tf*idf activates only the most distinctive attributes for a specific composite cooking activity, while frequency counts activate less selectively based on co-occurrence of task and attribute. Comparing WordNet expansion vs. literal, we find that the expansion helps (0.3% and 2.4% increase) as it activates a broader attribute inventory.

Next we compare these results to the reduced attribute training set, leading to disjoint training set for attributes and composite cooking activities (Table 6.3, upper

Attribute Training on:	MPII Cooking + MPII Composites	MPII Cooking
Training composite cooking activities on MPII Composites		
SVM (on features)	38.4	-
SVM (on attributes)	51.2	32.2
NN (on attributes)	51.7	34.6
NN+Script data		
- freqs-literal	50.9	**36.2**
- freqs-WN	51.2	35.6
- tf*idf-literal	51.5	32.1
- tf*idf-WN	**53.9**	30.7
No training data for composite cooking activities		
Script data		
- freqs-literal	42.6	**22.9**
- freqs-WN	38.0	22.1
- tf*idf-literal	**49.3**	22.4
- tf*idf-WN	48.7	21.5

Table 6.3: Composite cooking activity classification, AP in %. Top left quater: fully supervised, right column: reduced attribute training data, bottom section: no composite cooking activity training data, right bottom: true zero shot.

right quarter). Similar to the previously observed drop of performance of 20.4% for the combined attribute representation (Table 6.2, last row), we also see a significant drop in composite recognition of 19.0% and 17.1%, for SVM and NN, and 14.7% to 23.2% for the different NN+Script data versions. While the best performing approach is again based on NN+Script data, this time literal frequencies perform best with 36.2% AP. Presumably the attribute classifiers are all too weak and select only the semantically most relevant attributes like tf*idf, but this strategy fails if these few happen to be very noisy.

In the third part (Table 6.3, bottom left quarter), we evaluate the case when we do not have any training labels for the composite cooking tasks which does not allow using SVM or NN. We rely on script data for selecting relevant attributes instead. Using weighted attributes (Section 6.3.3) with the same measures, we again find tf*idf to perform best with 49.3% AP for the literal version, which is a drop by only 4.6% compared to the best fully supervised case. When using frequency statistics instead of tf*idf, performance drops to 42.6% and 38.0% AP.

Finally, we show our results on a true zero-shot setting (Table 6.3, right bottom part). We would like to stress that the attributes have only been trained on MPII Cooking and not as part of the unseen composites, nor are feature representations or composite cooking activities trained for the new MPII Composites, and also subjects are disjoint. Associations to unseen data is only provided by script data and not manually defined. For this challenging setting, we achieve a performance of 22.9% AP for the freqs-literal measure outperforming again the others like for the supervised case above.

Overall we found that script data improves performance by 2.2% AP to 53.9% AP in the fully supervised case and by 1.6% to 36.2% AP for reduced attribute training data. It also enables recognizing highly varied cooking tasks without training data close to supervised performance (49.3%) and obtains encouraging 22.9% for the complete zero-shot case where training happens entirely on a different dataset, different people, and different cooking tasks.

6.7 CONCLUSION

Composite activities are difficult to recognize because of their inherent variability and the lack of training data for specific composites. This chapter shows that attribute-based activity recognition allows recognizing composite activities well. Most notably, we have shown how textual script data, which is easy to collect, enables an improvement of the composite activity recognition when only little training data is available, and even allows for complete zero-shot transfer. We have also shown that activity attribute recognition can be improved by using context and co-occurrence attributes.

A direction for future work is to exploit the script structure for activity recognition by modeling the temporal structure of the video. In the following chapter we will improve the zero-shot recognition of composite activities by exploiting visual similarities within the unlabeled data. Chapter 9 shows how to produce detailed textual descriptions for the videos dataset presented in this chapter.

7 TRANSFER LEARNING IN A TRANSDUCTIVE SETTING

Contents

We showed in the previous chapters how zero-shot knowledge transfer can be achieved without manual supervision for object and video recognition using language resources and script data. In this chapter, we extend these ideas with semi-supervised learning to exploit unlabeled instances of (novel) categories with no or only a few labeled instances. Our proposed approach *Propagated Semantic Transfer* combines three techniques. First, we transfer information from known to novel categories by incorporating external knowledge, such as linguistic or expert-specified information, e.g., by a mid-level layer of semantic attributes. Second, we exploit the manifold structure of novel classes. More specifically we adapt a graph-based learning algorithm – so far only used for semi-supervised learning – to zero-shot and few-shot learning. Third, we improve the local neighborhood in such graph structures by replacing the raw feature-based representation with a mid-level object- or attribute-based representation. We evaluate our approach on three challenging datasets in two different applications, namely on *Animals with Attributes* and *ImageNet* for image classification and on *MPII Composites* for activity recognition. Our approach consistently outperforms state-of-the-art transfer and semi-supervised approaches on all datasets.

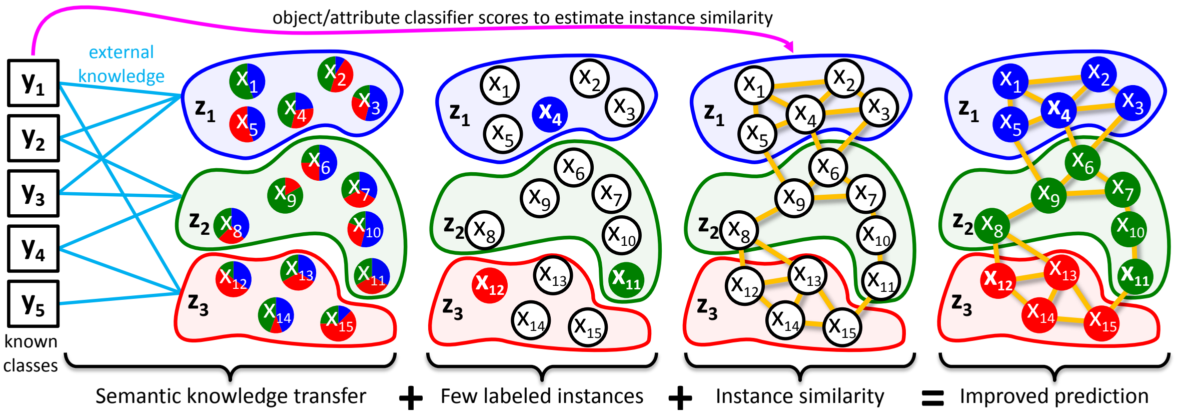


Figure 7.1: Conceptual visualisation of our approach Propagated Semantic Transfer. Known categories y, novel categories z, instances x (colors denote predicted category affiliation). See Figure 7.2 for an example result.

7.1 INTRODUCTION

While supervised training is an integral part of building visual, textual, or multimodal category models, more recently, knowledge transfer between categories has been recognized as an important ingredient to scale to a large number of categories as well as to enable fine-grained categorization. This development reflects the psychological point of view that humans are able to generalize to novel[10] categories with only a few training samples (Moses *et al.*, 1996; Bart and Ullman, 2005b). This has recently gained increased interest in the computer vision and machine learning literature, which look at zero-shot recognition (with no training instances for a class) (Lampert *et al.*, 2013; Farhadi *et al.*, 2009; Palatucci *et al.*, 2009; Parikh and Grauman, 2011; Fu *et al.*, 2013; Mensink *et al.*, 2012; Frome *et al.*, 2013), and one- or few-shot recognition (Thrun, 1996; Bart and Ullman, 2005b; Raina *et al.*, 2007). Knowledge transfer is particularly beneficial when scaling to large numbers of classes (Mensink *et al.*, 2012; Frome *et al.*, 2013), distinguishing fine-grained categories (Farrell *et al.*, 2011; Duan *et al.*, 2012), or analyzing compositional activities in videos (Fu *et al.*, 2013).

Recognizing categories with no or only few labeled training instances is challenging. To improve existing transfer learning approaches, we exploit several sources of information. Our approach allows using (1) trained category models, (2) external knowledge, (3) instance similarity, and (4) labeled instances of the novel classes if available. More specifically we learn category or attribute models based on labeled training data for known categories y (see also Figure 7.1) using supervised training. These trained models are then associated with the novel categories z using, e.g. expert or automatically mined semantic relatedness (cyan lines in Figure 7.1).

[10] We use "novel" throughout this chapter to denote categories with no or few labeled training instances.

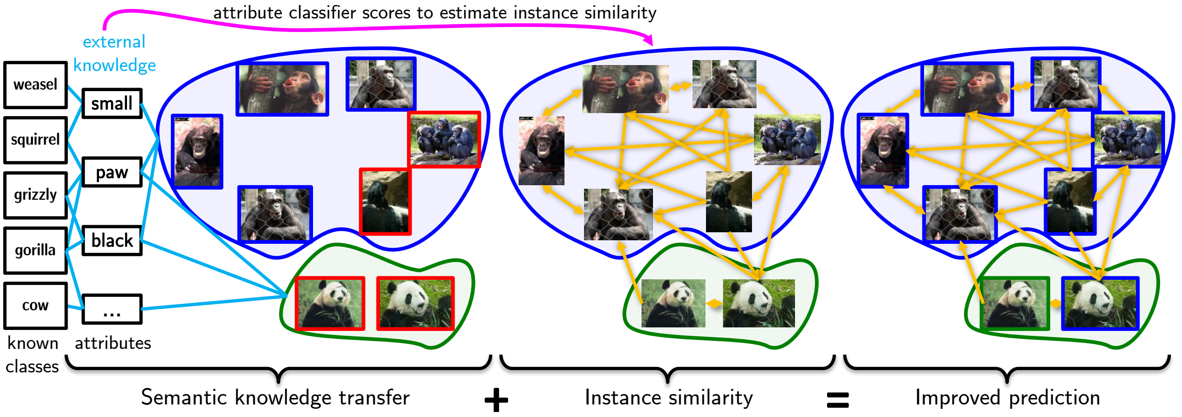


Figure 7.2: Visualisation of our approach Propagated Semantic Transfer with examples images (without few-shot). Graph structure and classification as determined by our system: while initially after knowledge transfer two chimpanzees and two giant pandas are incorrectly classified as another class (red boxes). Using instance similarity significantly improves this, and only one giant panda is now wrongly classified as a chimpanzee (blue box).

Similar to unsupervised learning Weber *et al.* (2000); Sivic *et al.* (2005) our approach exploits similarities in the data space via a graph structure to discover dense regions that are associated with coherent categories or concepts (orange graph structure in Figure 7.1). However, rather than using the raw input space, we map our data into a semantic output space with the models trained on the known classes (pink arrow) to benefit from their discriminative knowledge. Given the uncertain predictions and the graph structure we adapt semi-supervised label propagation (Zhu *et al.*, 2003; Zhou *et al.*, 2004) to generate more reliable predictions. If labeled instances are available they can be seamlessly added. Note, attribute or category models do not have to be retrained if novel classes are added which is an important aspect e.g. in a robotic scenario. In Figure 7.2 we show how approach works on a few sample images.

The main contribution of this chapter is threefold. First, we propose a novel approach that extends semantic knowledge transfer to the transductive setting, exploiting similarities in the unlabeled data distribution. The approach allows to do zero-shot recognition but also smoothly integrate labels for novel classes (Section 7.3). Second, we improve the local neighborhood structure in the raw feature space by mapping the data into a low dimensional semantic output space using the trained attribute and category models. Third, we validate our approach on three challenging datasets for two different applications, namely on *Animals with Attributes* and *ImageNet* for image classification and on *MPII Composites* for activity recognition (Section 7.4). We also provide a discussion of related work (Section 7.2) and conclusions for future work (Section 7.5).

7.2 RELATED WORK

Knowledge transfer or transfer learning has the goal to transfer information of learned models to changing or unknown data distributions while reducing the need and effort to collect new training labels. It refers to a variety of tasks, including domain adaptation (Saenko *et al.*, 2010) or sharing of knowledge and representations (Torralba *et al.*, 2004; Blanke and Schiele, 2010) (a recent categorization can be found in Pan and Yang (2010)).

In this work we focus on transferring knowledge from known categories with sufficient training instances to novel categories with limited training data. In computer vision or machine learning literature this setting is normally referred to as zero-shot learning (Lampert *et al.*, 2013; Palatucci *et al.*, 2009; Rohrbach *et al.*, 2010; Fu *et al.*, 2013; Mensink *et al.*, 2012) if there are no instances for the test classes available and one- or few-shot learning (Mensink *et al.*, 2012; Fu *et al.*, 2013; Fink, 2004) if there are one or few instances available for the novel classes.

To recognize novel categories zero-shot recognition uses additional information, typically in the form of an intermediate attribute representation (Lampert *et al.*, 2013; Fu *et al.*, 2013), direct similarity (Rohrbach *et al.*, 2010) between categories, or hierarchical structures of categories (Zweig and Weinshall, 2007). The information can either be manually specified (Lampert *et al.*, 2013; Fu *et al.*, 2013) or mined automatically from knowledge bases as we have shown in (Rohrbach *et al.*, 2010) and in Chapters 3 and 4. Our approach builds on these works by using a semantic knowledge transfer approach as the first step. If one or a few training examples are available, these are typically used to select or adapt known models (Bart and Ullman, 2005b; Fu *et al.*, 2013; Sharmanska *et al.*, 2012). In contrast to related work, our approach uses the above mentioned semantic knowledge transfer also when few training examples are available to reduce the dependency on the quality of the samples. Also, we still use the labeled examples to propagate information.

Additionally, we exploit the neighborhood structure of the unlabeled instances to improve recognition for zero- and few-shot recognition. This is in contrast to previous works with the exception of the zero-shot approach of Fu *et al.* (2013) who learn a discriminative, latent attribute representation and applies self-training on the unseen categories. While conceptually similar, the approach is different to ours, as we explicitly use the local neighborhood structure of the unlabeled instances. A popular choice to integrate local neighborhood structure of the data are graph-based methods. These have been used to discover a grouping by spectral clustering (Ng *et al.*, 2002; Luxburg, 2007), and to enable semi-supervised learning (Zhu *et al.*, 2003; Zhou *et al.*, 2004). Our setting is similar to the semi-supervised setting. To transfer labels from labeled to unlabeled data *label propagation* is widely used (Zhu *et al.*, 2003; Zhou *et al.*, 2004) and has shown to work successfully in several applications (Liu *et al.*, 2011; Fergus *et al.*, 2009). In this work, we extend transfer learning by considering the neighborhood structure of the novel classes. For this we adapt the well-known label propagation approach of Zhou *et al.* (2004). We build a k-nearest neighbor graph to capture the underlying manifold structure as it has shown to provide the most robust structure (Maier *et al.*, 2008). Nevertheless, the quality of the graph structure is key to success of graph-based methods and strongly dependents on the feature representation (Ebert *et al.*, 2010). We thus improve the graph structure by replacing the noisy raw input space with the more compact semantic output space which has shown to improve recognition (Sharmanska *et al.*, 2012).

To improve image classification with reduced training data, Choi *et al.* (2013) and Shrivastava *et al.* (2012) use attributes as an intermediate layer and incorporate unlabeled data, however, both works are in a classical semi-supervised learning setting similar to Ebert *et al.* (2010), while our setting is transfer learning. More specifically Shrivastava *et al.* (2012) propose to bootstrap classifiers by adding unlabeled data. The bootstrapping is constrained by attributes shared across classes. In contrast, we use attributes for transfer and exploit the similarity between instances of the novel classes. Choi *et al.* (2013) automatically discover a discriminative attribute representation, while incorporating unlabeled data. This notion of attributes is different to ours as we want to use semantic attributes to enable transfer from other classes. Other directions to improve the quality of the intermediate representation include integrating metric learning (Tran and Sorokin, 2008; Mensink *et al.*, 2012) or online methods (Kankuekul *et al.*, 2012) which we defer to future work.

7.3 propagated semantic transfer (pst)

Our main objective is to robustly recognize novel categories by transferring knowledge from known classes and exploiting the similarity of the test instances. More specifically our novel approach called *Propagated Semantic Transfer* consists of the following four components: we employ semantic knowledge transfer from known classes to novel classes (Section 7.3.1); we combine the transferred predictions with labels for the novel classes (Section 7.3.2); a similarity metric is defined to achieve a robust graph structure (Section 7.3.3); we propagate this information within the novel classes (Section 7.3.4).

7.3.1 Semantic knowledge transfer

We first transfer knowledge using a semantic representation. This allows to include external knowledge sources. We model the relation between a set of K known classes $y_1, \ldots, y_K$ to the set of N novel classes $z_1, \ldots, z_N$. Both sets are disjoint, i.e. $\{y_1, \ldots, y_K\} \bigcap \{z_1, \ldots, z_N\} = \emptyset$. We use two strategies to achieve this transfer: i) an attribute representation that employs an intermediate representation of $a_1, \ldots, a_M$ attributes or ii) direct similarities calculated among the known object classes. Both work without any training examples for z_n, i.e. also for zero-shot recognition (Lampert *et al.*, 2013; Rohrbach *et al.*, 2010).

7.3.1.1 *Attribute representation*

We use the Direct-Attribute-Prediction (DAP) model (Lampert *et al.*, 2013), using the formulation from Chapter 3. An intermediate level of M attribute classifiers $p(a_m|x)$ is trained on the known classes y_k to estimate the presence of attribute a_m in the instance x. The subsequent knowledge transfer requires an external knowledge source that provides class-attribute associations $a_m^{z_n} \in \{0,1\}$ indicating if attribute a_m is associated with class z_n. Options for such association information are discussed in Section 7.4.2. Given this information the probability of the novel classes z_n to be present in the instance x can then be estimated:

$$p(z_n|x) \propto \prod_{m=1}^{M} (2p(a_m|x))^{a_m^{z_n}}. \tag{7.1}$$

7.3.1.2 *Direct similarity*

As an alternative to attributes, we can use the U most similar training classes $y_1, ..., y_U$ as a predictor for novel class z_n given an instance x (see Chapter 3):

$$p(z_n|x) \propto \prod_{u=1}^{U} (2p(y_u|x))^{y_u^{z_n}}, \tag{7.2}$$

where $y_u^{z_n}$ provides continuous normalized weights for the strength of the similarity between the novel class z_n and the known class y_u. To comply with Chapters 4 and 6 we slightly diverge from these models for the ImageNet and MPII Composites dataset by using a sum formulation instead of the probabilistic expression, i.e. for attributes $p(z_n|x) \propto \frac{\sum_{m=1}^{M} a_m^{z_n} p(a_m|x)}{\sum_{m=1}^{M} a_m^{z_n}}$, and for direct similarity $p(z_n|x) \propto \frac{\sum_{u=1}^{U} p(y_u|x)}{U}$. Note that in this case we do not obtain probability estimates, however, for label propagation the resulting scores are sufficient.

7.3.2 Combining transferred and ground truth labels

In the following we treat the multi-class problem as N binary problems, where N is the number of binary classes. For class z_n the semantic knowledge transfer provides $p(z_n|x) \in [0,1]$ for all instances x. We combine the best predictions per class, scaled to $[-1,1]$, with labels $\hat{l}(z_n|x) \in \{-1,1\}$ provided for some instances x in the following way:

$$l(z_n|x) = \begin{cases} \gamma \hat{l}(z_n|x) & \text{if there is a label for } x \\ (1-\gamma)(2p(z_n|x)-1) & \text{if } p(z_n|x) \text{is among top-}\delta \text{ fraction of predictions for } z_n \\ 0 & \text{otherwise.} \end{cases} \tag{7.3}$$

γ provides a weighting between the true labels and the predicted labels. In the zero-shot case we only use predictions, i.e. $\gamma = 0$. The parameters $\delta, \gamma \in [0,1]$ are chosen, similar to the remaining parameters, using cross-validation on the training set.

7.3.3 Similarity metric based on discriminative models for graph construction

We enhance transfer learning by exploiting also the neighborhood structure within novel classes, i.e. we assume a transductive setting. Graph-based semi-supervised learning incorporates this information by employing a graph structure over all instances. In this section we describe how to improve the graph structure as it has a strong influence on the final results (Ebert *et al.*, 2010). The k-NN graph is usually built on the raw feature descriptors of the data. Distances are computed for each pair (x_i, x_j) by

$$d(x_i, x_j) = \sum_{d=1}^{D} |x_{i,d} - x_{j,d}|, \tag{7.4}$$

where D is the dimensionality of the raw feature space. We note that the visual representation used for label propagation can be independent of the visual representation used for transfer. While the visual representation for transfer is required to provide good generalization abilities in conjunction with the employed supervised learning strategy, the visual representation for label propagation should induce a good neighborhood structure. Therefore we propose to use the more compact output space trained on the known classes which we found to provide a much better structure, see Figure 7.4. We thus compute the distances either on the M-dimensional vector of the attribute classifiers $p(a_m|x)$ with $M \ll D$, i.e.,

$$d(x_i, x_j) = \sum_{m=1}^{M} |p(a_m|x_i) - p(a_m|x_j)|, \tag{7.5}$$

or on the K-dimensional vector of object-classifiers $p(y_k|x)$ with $K \ll D$, i.e.

$$d(x_i, x_j) = \sum_{\kappa=1}^{K} |p(y_\kappa|x_i) - p(y_\kappa|x_j)|. \tag{7.6}$$

These distances are transformed into similarities with a RBF kernel: $s(x_i, x_j) = \exp\left(\frac{-d(x_i,x_j)}{2\sigma^2}\right)$. Finally, we construct a k-NN graph that is known for its good performance (Maier *et al.*, 2008; Ebert *et al.*, 2010), i.e.,

$$W_{ij} = \begin{cases} s(x_i, x_j) & \text{if } s(x_i, x_j) \text{ is among the k largest similarities of } x_i \\ 0 & \text{otherwise.} \end{cases} \tag{7.7}$$

7.3.4 Label propagation with certain and uncertain labels

In this work, we build upon the label propagation by Zhou *et al.* (2004). The k-NN graph with RBF kernel gives the weighted graph W (see Section 7.3.3). Based on this graph we compute a normalized graph Laplacian, i.e., $S = D^{-1/2}WD^{-1/2}$ with the diagonal matrix D summing up the weights in each row in W. Traditional semi-supervised label propagation uses sparse ground truth labels. In contrast we have dense labels $l(z_n|x)$ which are a combination of uncertain predictions and certain labels (see Equation (7.3)) for all instances $\{x_1, \dots, x_i\}$ of the novel classes z_n. Therefore, we modify the initialization by setting

$$L_n^{(0)} = [l(z_n|x_1), \dots, l(z_n|x_i)] \tag{7.8}$$

for the N novel classes. For each class, labels are propagated through this graph structure converging to the following closed form solution in the limit

$$L_n^* = (I - \alpha S)^{-1} L_n^{(0)} \quad \text{for } 1 \leq n \leq N, \tag{7.9}$$

with the regularization parameter $\alpha \in (0, 1]$. The resulting framework makes use of the manifold structure underlying the novel classes to regulate the predictions from transfer learning. In practice we use an interative procedure as the algorithm converges after a few iterations:

$$L_n^{(t+1)} = \alpha S L_n^{(t)} + (1 - \alpha) L_n^{(0)} \quad \text{for } 1 \leq n \leq N. \tag{7.10}$$

7.4 EVALUATION

We validate our approach on three datasets: first, the Animals with Attributes dataset (AwA), which is one of the first and most widely used datasets for semantic knowledge transfer and zero-shot recognition; second, for large scale and fine-grained recognition, the ImageNet 2010 challenge dataset (ImageNet); and, in the domain of activity recognition in videos, the MPII Cooking Composite Activities dataset (MPII Composite Activities). Sample images of these datasets are shown in Figure 7.3.

7.4.1 Datasets

We shortly outline the most important properties of the examined datasets in the following paragraphs and show example images/frames in Figure 7.3.

7.4.1.1 *AwA*

The Animals with Attributes dataset (AwA) (Lampert *et al.*, 2013) is one of the first and most widely used datasets for semantic knowledge transfer and zero-shot recognition and evaluated our approaches on it in Chapter 3. It consists of 50 mammal classes, 40 training (24,395 images) and 10 disjoint test classes (6,180 images). We use the provided pre-computed 6 image descriptors, which are concatenated.

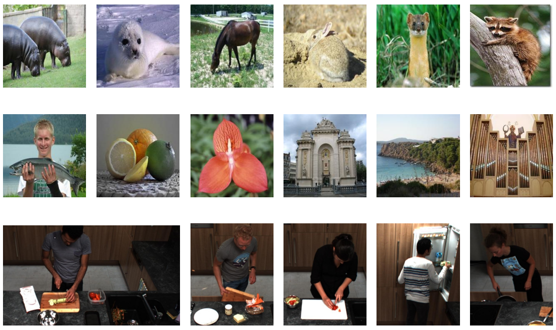


Figure 7.3: Example images/frames from AwA (first row), ImageNet (second row), and MPII Composite Activities (third row)

7.4.1.2 *ImageNet*

The ImageNet 2010 challenge (Berg *et al.*, 2010a) requires large scale and fine-grained recognition. It consists of 1,000 image categories which are split into 800 training and 200 test categories according as in Chapter 4. We use the LLC and Fisher-Vector encoded SIFT descriptors as introduced in Chapter 4.

7.4.1.3 *MPII Composite Activities*

The MPII Composite Cooking Activities dataset introduced in Chapter 6 distinguishes 41 basic cooking activities, such as *prepare scrambled egg* or *prepare carrots* with video recordings of varying length from 1 to 41 minutes. It consists of a total of 256 videos, 44 are used for training the attribute representation, 170 are used as test data. We use the same dense-trajectory representation and train/test split as in Chapter 6.

7.4.2 External knowledge sources and similarity measures

Our approach incorporates external knowledge to enable semantic knowledge transfer from known classes y to unseen classes z. We use the class-attribute associations $a_m^{z_n}$ for attribute-based transfer (Equation (7.1)) or inter-class similarity $y_u^{z_n}$ for direct-similarity-based transfer (Equation (7.2)) provided with the datasets. In the following we shortly outline the knowledge sources and measures.

7.4.2.1 *Manual (AwA)*

AwA is accompanied with a set of 85 attributes and associations to all 40 training and all 10 test classes. The associations are provided by human judgments (Lampert *et al.*, 2013).

7.4.2.2 *Hierarchy (ImageNet)*

For ImageNet the manually constructed WordNet/ImagNet hierarchy is used to find the most similar of the 800 known classes (leaf nodes in the hierarchy). Furthermore, the 370 *inner nodes* can group several classes into attributes, see Section 4.3.1 for details.

7.4.2.3 *Linguistic knowledge bases (AwA, ImageNet)*

An alternative to manual association are automatically mined associations. We use similarity matrices which are extracted using different linguistic similarity measures as described in Chapters 3 and 4. They are either based on linguistic corpora, namely *Wikipedia* and *WordNet*, or on hit-count statistics of web search. One can distinguish basic web search (*Yahoo Web*), web search refined to part associations (*Yahoo Holonyms*), image search (*Yahoo Image* and *Flickr Image*), or use the information of the summary snippets returned by web search (*Yahoo Snippets*). As ImageNet does not provide attributes, we mined 811 part-attributes from the associated WordNet hierarchy.

7.4.2.4 *Script data (MPII Composites)*

To associate composite cooking activities such as *preparing carrots* with attributes of fine-grained activities (e.g. *wash*, *peel*), ingredients (e.g. *carrots*), and tools (e.g. *knife*, *peeler*), we collected textual description (*Script data*) of these activities with AMT as described in detail in Chapter 6. The provided associations are computed based on either the *frequency statistics* or, more discriminate, by term frequency times inverse document frequency (*tf*idf*). Words in the text can be matched to labels either *literally* or by using *WordNet* expansion, see Section 6.4.2.1 for details.

7.4.3 Results

To enable a direct comparison, we closely follow the experimental setups of the respective datasets as introduced in Chapters 3, 4, and 6. Only for AwA we use in contrast to Chapter 3 all images for training instead of a subset of 92 and train sigmoid functions to estimate probabilities using the code provided by Lin *et al.* (2007). On all datasets we train attribute or object classifiers (for direct similarity) with one-vs-all SVMs using Mean Stochastic Gradient Descent (see Section 4.4.4) and, for AwA and MPII Composites, with a χ^2 kernel approximation (Vedaldi and Zisserman, 2010).

The hyper-parameters of our new *Propagated Semantic Transfer* algorithm are estimated using 5-fold cross-validation on the respective training set, splitting them into 80% known and 20% novel classes: We determine the parameters for our approach on the AwA training set and then set them for all datasets to $\alpha = 0.8$, $\gamma = 0.98$, the number of neighbors $k = 50$, the number of iterations for propagation to 10, and use $L1$ distance. Due to the different recognition precision of the datasets

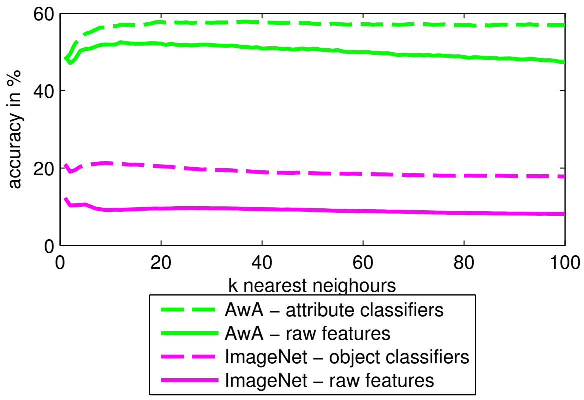


Figure 7.4: Accuracy of the majority vote from kNN (kNN-Classifier) on the test sets' ground truth.

we determine $\delta = 0.15/0.04$ separately for AwA/ImageNet. For MPII Composites we only do zero-shot recognition and use all samples due to the limited number of samples of ≤ 7 per class. For few-shot recognition we report the mean over 10 runs where we pick examples randomly. The labeled examples are included in the evaluation to make it comparable to the zero-shot case.

We validate our claim that the classifier output space induces a better neighborhood structure than the raw features by examining the k-Nearest-Neighbour (kNN) quality for both. In Figure 7.4 we compare the kNN quality on two datasets for both feature representation. We observe that the attribute (Equation (7.5)) and object (Equation (7.6)) classifier-based representations (green and magenta dashed line) achieve a significantly higher accuracy than the respective raw feature-based representation (Equation (7.4), Figure 7.4 solid lines). We note that a good kNN-quality is required but not sufficient for good propagation, as it also depends on the distribution and quality of initial predictions. In the following, we compare the resulting final performance of the raw features with the attribute classifier representation.

7.4.3.1 *AwA - image classification*

We start by comparing the performance of related work to our approach on AwA (see Section 7.4.1) in Figure 7.5. We start by examining the zero-shot results in Table 7.1, where no training examples are available for the novel or in this case unseen classes. The best results to our knowledge for on this dataset are reported by Lampert *et al.* (2013). On this 10-class zero-shot task they achieve 81.4% area under ROC-curve (AUC) and 41.4% multi-class accuracy (Acc) with DAP, averaged over the 10 test classes. Additionally we report results from Zero-Shot Learning (Fu *et al.*, 2013) which achieves 41.3% Acc. Our *Propagated Semantic Transfer*, using the raw image descriptors to build a neighborhood structure, achieves 81.2% AUC and 40.5% Acc. However, when propagating on the 85-dimensional attribute space, we improve over Lampert *et al.* (2013) and Fu *et al.* (2013) to 83.7% AUC and 42.7% Acc. To understand the difference in performance between the attribute and the image

Approach	Performance AUC	 Acc.
DAP (Lampert *et al.*, 2013)	81.4	41.4
IAP (Lampert *et al.*, 2013)	80.0	42.2
Zero-Shot Learning (Fu *et al.*, 2013)	n/a	41.3
PST (ours)		
on image descriptors	81.2	40.5
on attributes	83.7	42.7

Table 7.1: Zero-shot results on AwA dataset, see Section 7.4.3.1. Predictions with attributes and manual defined associations, in %.

descriptor space we examine the neighborhood quality used for propagating labels shown in Figure 7.4. The k-NN accuracy, measured on the ground truth labels, is significantly higher for the attribute space (green dashed curve) compared to the raw features (solid green). The information is more likely propagated to neighbors of the correct class for the attribute-space leading to a better final prediction. Another advantage is the significantly reduced computation and storage costs for building the k-NN graph which scales linearly with the dimensionality. We believe that such an intermediate space, in this case represented by attributes, might provide a better neighborhood structure and could be used in other label-propagation tasks.

Next we compare our approach in the few-shot setting, i.e. we add labeled examples per class. In Figure 7.4.3.1 we compare our approach (PST) to two label propagation (LP) baselines. We first note that PST (red curves) seamlessly moves from zero-shot to few-shot, while traditional LP (blue and black curves) needs at least one training example. We first examine the three solid lines. The black curve is the best LP variant from Ebert *et al.* (2010) evaluated on the 10 test classes of AwA rather than all 50 as done by Ebert *et al.* We also compute LP in combination with the similarity metric based on the attribute classifier scores (blue curves). This transfer of knowledge residing in the classifier trained on the known classes already gives a significant improvement in performance. Our approach (red curve) additionally transfers labels from the known classes and improves further. Especially for few labels our approach benefits from the transfer, e.g. for 5 labeled samples per class PST achieves 43.9% accuracy, compared to 38.1% for LP with attribute classifiers and 32.2% for LP according to Ebert *et al.* (2010). For less samples LP drops significantly while our approach has nearly stable performance. For large amounts of training data, PST approaches - as expected - LP (red vs. blue in Figure 7.4.3.1).

The dashed lines in Figure 7.4.3.1 provide results for automatically mined associations $a_m^{z_n}$ between attributes and classes. It is interesting to note that these automatically mined associations achieve performance very close to the manual defined associations (dashed vs. solid). In this plot we use Yahoo Image as base for the semantic relatedness.

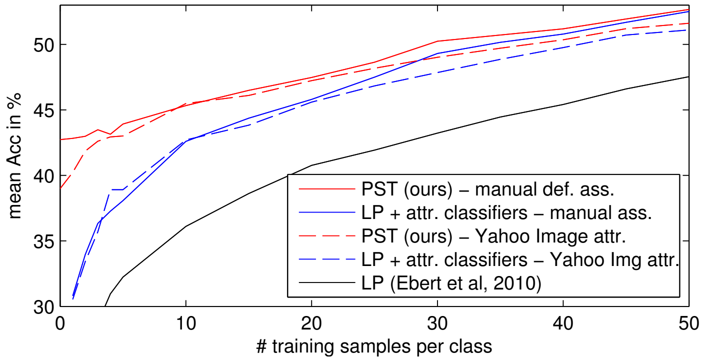
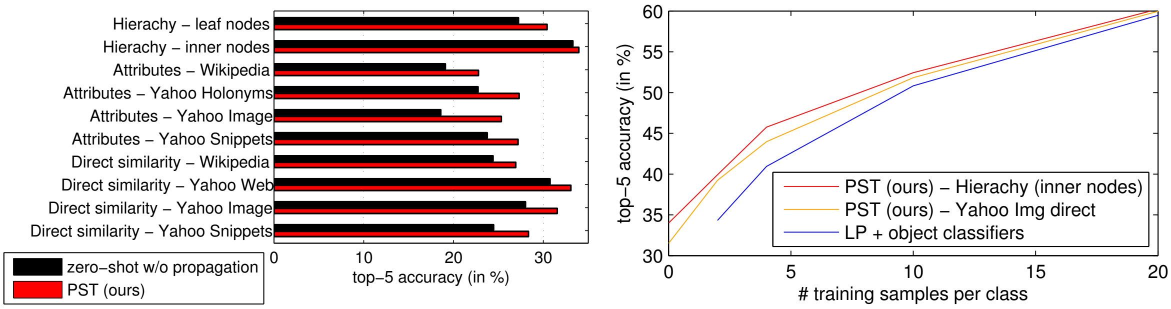


Figure 7.5: Few-Shot Results on AwA Dataset, see Section 7.4.3.1.



(a) Zero-Shot results on ImageNet.

(b) Few-Shot results on ImageNet.

Figure 7.6: Results on ImageNet, see Section 7.4.3.2.

7.4.3.2 *ImageNet - large scale image classification*

In this section we evaluate our Propagated Semantic Transfer approach on a large image classification task with 200 unseen image categories using the setup as in Chapter 4. We report the top-5 accuracy[11] (Berg *et al.*, 2010a) which requires one of the best five predictions for an image to be correct.

In Figure 7.6(a) we compare our results to zero-shot without propagation presented in Chapter 4 and published in (Rohrbach *et al.*, 2011). For zero-shot recognition our PST (red bars) improves performance over zero-shot without propagation (black bars). The largest improvement in top-5 accuracy is achieved for Yahoo Image with Attributes which increases by 6.7% to 25.3%. The absolute performance of 34.0% top-5 accuracy is achieved by using the inner nodes of the WordNet hierarchy for transfer, closely followed by Yahoo Web with direct similarity, achieving 33.1% top-5 accuracy. Similar to the AwA dataset we improve PST over the LP-baseline for few-shot recognition (Figure 7.6(b)).

[11] *top-5 accuracy* = 1 - *top-5 error* as defined by Berg *et al.* (2010a)

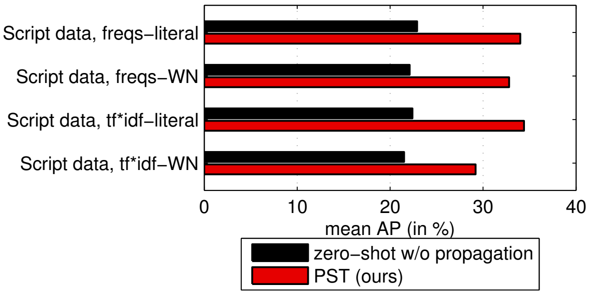


Figure 7.7: Results on MPII Composite Activities, see Section 7.4.3.3.

7.4.3.3 *MPII composite - activity recognition*

In the last two subsections, we showed the benefit of *Propagated Semantic Transfer* on two image classification challenges. We now evaluate our approach on the video-activity recognition dataset MPII Composite Cooking Activities.

We compute mean AP using the features and the setup from Chapter 6. In Figure 7.7 we compare the performance of Propagated Semantic Transfer (red bars) to the results of zero-shot recognition without propagation as we presented in Chapter 6 and published in Rohrbach *et al.* (2012b) (black bars). We distinguish the four variants of Script data based transfer. Our approach achieves significant performance improvements in all four cases, increasing mean AP by 11.1%, 10.7%, 12.0%, and 7.7% to 34.0%, 32.8%, 34.4%, and 29.2%, respectively. This is especially impressive as it reaches the level of supervised training: for the same set of attributes (and very few, ≤ 7 training categories per class) we achieve 32.2% for SVM, 34.6% for NN-classification, and up to 36.2% for a combination of NN with script data, see Table 6.3.

We find these results encouraging as it is much more difficult to collect and label training examples for this domain than for image classification and the complexity and compositional nature of activities frequently requires recognizing unseen categories (Fu *et al.*, 2013).

7.5 CONCLUSION

In this chapter we address a frequently occurring setting where there is large amount of training data for some classes, but other, e.g. novel classes, have no or only few labeled training samples. We propose a novel approach named *Propagated Semantic Transfer*, which integrates semantic knowledge transfer with the visual similarities of unlabeled instances within the novel classes. We adapt a semi-supervised label-propagation approach by building the neighborhood graph on expressive, low-dimensional semantic output space and by initializing it with predictions from knowledge transfer.

We evaluated this approach on three diverse datasets for image and video-activity recognition, consistently improving performance over the state-of-the-art for *zero-shot* and *few-shot* prediction. Most notably we achieve 83.7% AUC / 42.7% multi-class accuracy on the Animals with Attributes dataset for zero-shot recognition, scale to 200 unseen classes on ImageNet, and achieve up to 34.4% (+12.0%) mean AP on MPII Composite Activities which is on the level of supervised training on this dataset. We show that our approach consistently improves performance independent of factors such as (1) the specific datasets and descriptors, (2) different transfer approaches: direct vs. attributes, (3) types of transfer association: manually defined, linguistic knowledge bases, or script data, (4) domain: image and video activity recognition, or (5) model: probabilistic vs. sum formulation.

8 GROUNDING ACTION DESCRIPTIONS IN VIDEOS

Contents

PREVIOUS chapters focused on visual recognition tasks by improving object and activity classification using knowledge minded from linguistic knowledge bases and text descriptions. In this chapter we turn the focus to a computational linguistic task: we show how visual information can improve the semantic similarity estimate between sentences. This problem is referred to as grounding in computational linguistics and recent work has shown that the integration of visual information into text-based models can substantially improve model predictions, but so far only visual information extracted from static images has been used. More specifically we consider the problem of grounding *sentences describing actions* in the *cooking videos* we presented in Chapter 6. While we collected cooking instructions independently of the video in Chapter 6 to capture the variability of composite activities, in this chapter we collect descriptions specific for each video. The natural language descriptions of the actions are aligned with the videos and we annotate how similar the action descriptions are to each other. Experimental results demonstrate that a text-based model of similarity between actions improves substantially when combined with visual information from videos depicting the described actions.

8.1 INTRODUCTION

The estimation of semantic similarity between words and phrases is a basic task in computational semantics. Vector-space models of meaning are one standard approach. Following the distributional hypothesis, frequencies of context words are recorded in vectors, and semantic similarity is computed as a proximity measure in the underlying vector space. Such distributional models are attractive because they are conceptually simple, easy to implement and relevant for various NLP tasks (Turney and Pantel, 2010). At the same time, they provide a substantially incomplete picture of word meaning, since they ignore the relation between language and extra-linguistic information, which is constitutive for linguistic meaning. In the last few years, a growing amount of work has been devoted to the task of grounding meaning in visual information, in particular by extending the distributional approach to jointly cover texts and images (Feng and Lapata, 2010; Bruni *et al.*, 2011). As a clear result, visual information improves the quality of distributional models. Bruni *et al.* (2011) show that visual information drawn from images is particularly relevant for concrete common nouns and adjectives.

A natural next step is to integrate visual information from *videos* into a semantic model of event and action verbs. Psychological studies have shown the connection between action semantics and videos (Glenberg, 2002; Howell *et al.*, 2005), but to our knowledge, we are the first to provide a suitable data source and to implement such a model.

The contribution of this chapter is three-fold:

- We present a *multimodal corpus* containing textual descriptions aligned with high-quality videos. Starting from the video corpus introduced in Chapter 6, which contains high-resolution video recordings of basic cooking tasks, we collected multiple textual descriptions of each video via Mechanical Turk. We also provide an accurate sentence-level alignment of the descriptions with their respective videos. We expect the corpus to be a valuable resource for computational semantics, and moreover helpful for a variety of purposes, including video understanding and generation of text from videos. The latter one we explore in the next chapter.

- We provide a *gold-standard dataset* for the evaluation of similarity models for action verbs and phrases. The dataset has been designed as analogous to the Usage Similarity dataset of Erk *et al.* (2009) and contains pairs of natural-language action descriptions plus their associated video segments. Each of the pairs is annotated with a similarity score based on several manual annotations.

- We report an experiment on *similarity modeling of action descriptions* based on the video corpus and the gold standard annotation, which demonstrates the impact of scene information from videos. Visual similarity models outperform text-based models; the performance of combined models approaches the upper bound indicated by inter-annotator agreement.

The chapter is structured as follows: We first place ourselves in the landscape of related work (Section 8.2), then we introduce our corpus (Section 8.3). Section 8.4 reports our action similarity annotation experiment and Section 8.5 introduces the similarity measures we apply to the annotated data. We outline the results of our evaluation in Section 8.6, and conclude the chapter with a summary and directions for future work (Section 8.7).

8.2 RELATED WORK

A large multimodal resource combining language and visual information resulted from the ESP game (von Ahn and Dabbish, 2004). The dataset contains many images tagged with several one-word labels.

The Microsoft Video Description Corpus (MSVD Chen and Dolan, 2011) is a resource providing textual descriptions of videos. It consists of multiple crowd-sourced textual descriptions of short video snippets. The MSVD corpus is much larger than our corpus, but most of the videos are of relatively low quality and therefore very challenging for state-of-the-art video processing to extract precise information. The videos are typically short and summarized with a single sentence. Our corpus contains coherent textual descriptions of longer video sequences, where each sentence is associated with a timeframe.

Gupta *et al.* (2009) present another useful resource: their model learns the alignment of predicate-argument structures with videos and uses the result for action recognition in videos. However, the corpus contains no natural language texts.

The connection between natural language sentences and videos has so far been mostly explored by the computer vision community, where different methods for improving action recognition by exploiting linguistic data have been proposed (among others Gupta and Mooney, 2010; Motwani and Mooney, 2012; Cour *et al.*, 2008; Tzoukermann *et al.*, 2011). Our resource is intended to be used for action recognition as well, but in this chapter, we focus on the inverse effect of visual data on language processing.

Feng and Lapata (2010) were the first to enrich topic models for newspaper articles with visual information, by incorporating features from article illustrations. They achieve better results when incorporating the visual information, providing an enriched model that pairs a single text with a picture.

Bruni *et al.* (2011) used the ESP game data to create a visually grounded semantic model. Their results outperform purely text-based models using visual information from pictures for the task of modeling noun similarities. They model single words, and visual features lead only to moderate improvements, which might be due to the mixed quality and random choice of the images. Dodge *et al.* (2012) recently investigated which words can actually be grounded in images at all, producing an automatic classifier for visual words.

An interesting in-depth study by Mathe *et al.* (2008) automatically learnt the semantics of motion verbs as abstract features from videos. The study captures 4 actions with 8-10 videos for each of the actions, and would need a perfect object recognition from a visual classifier to scale up.

Steyvers (2010) and later Silberer and Lapata (2012) present an alternative approach to incorporating visual information directly: they use so-called *feature norms*, which consist of human associations for many given words, as a proxy for general perceptual information. Because this model is trained and evaluated on those feature norms, it is not directly comparable to our approach.

The *Restaurant Game* by Orkin and Roy (2009) grounds written chat dialogues in actions carried out in a computer game. While this work is outstanding from the social learning perspective, the actions that ground the dialogues are clicks on a screen rather than real-world actions. The dataset has successfully been used to model determiner meaning (Reckman *et al.*, 2011) in the context of the *Restaurant Game*, but it is unclear how this approach could scale up to content words and other domains.

8.3 THE TACOS CORPUS

We build our corpus on top of the "MPII Cooking Composite Activities" video corpus (*MPII Composites*) introduced in Chapter 6, which contains videos of different activities in the cooking domain, e.g., *preparing carrots* or *separating eggs*. We extend the existing corpus with multiple textual descriptions collected by crowd-sourcing via Amazon Mechanical Turk[12] *(MTurk)*. To facilitate the alignment of sentences describing activities with their proper video segments, we also obtained approximate timestamps, as described in Section 8.3.2.

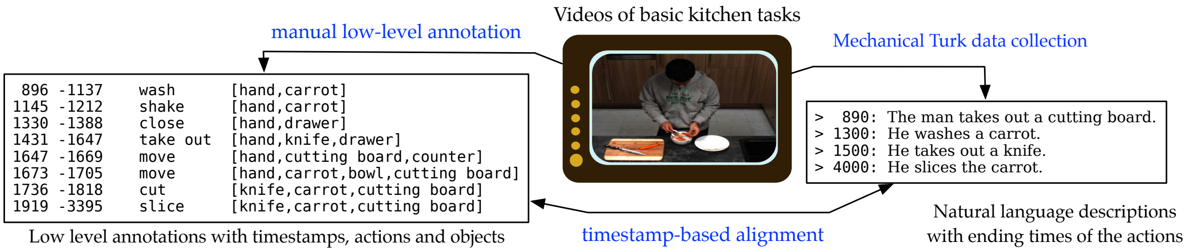
MPII Composites comes with timed gold-standard annotation of low-level activities and participating objects (e.g. OPEN [HAND, DRAWER] or TAKE OUT [HAND, KNIFE, DRAWER]). By adding textual descriptions (e.g., *The person takes a knife from the drawer*) and aligning them on the sentence level with videos and low-level annotations, we provide a rich multimodal resource (cf. Figure 8.1), the "Saarbrücken Corpus of Textually Annotated Cooking Scenes" (TACOS). In particular, the TACoS corpus provides:

- A collection of coherent *textual descriptions for video recordings* of activities of medium complexity, as a basis for empirical discourse-related research, e.g., the selection and granularity of action descriptions in context.

- A high-quality *alignment of sentences with video segments*, supporting the grounding of action descriptions in visual information.

- *Collections of paraphrases* describing the same scene, which result as a by-product from the text-video alignment and can be useful for text generation from videos (among other things).

[12] mturk.com


manual low-level annotation
Videos of basic kitchen tasks
Mechanical Turk data collection
Low level annotations with timestamps, actions and objects
timestamp-based alignment
Natural language descriptions
with ending times of the actions

```
 896 -1137    wash       [hand,carrot]
1145 -1212    shake      [hand,carrot]
1330 -1388    close      [hand,drawer]
1431 -1647    take out   [hand,knife,drawer]
1647 -1669    move       [hand,cutting board,counter]
1673 -1705    move       [hand,carrot,bowl,cutting board]
1736 -1818    cut        [knife,carrot,cutting board]
1919 -3395    slice      [knife,carrot,cutting board]
```

```
>  890: The man takes out a cutting board.
> 1300: He washes a carrot.
> 1500: He takes out a knife.
> 4000: He slices the carrot.
```

Figure 8.1: TACoS corpus overview

- The alignment of textual activity descriptions with *sequences of low-level activities*, which may be used to study the decomposition of action verbs into basic activity predicates.

We expect that our corpus will encourage and enable future work on various topics in natural language and video processing. In this chapter, we will make use of the second aspect only, demonstrating the usefulness of the corpus for the grounding task.

After a more detailed description of the basic video corpus and its annotation (Section 8.3.1) we describe the collection of textual descriptions with MTurk (Section 8.3.2), and finally show the assembly and some benchmarks of the final corpus (Section 8.3.3).

8.3.1 The video corpus

MPII Composites contains 212 high resolution video recordings of 1-23 minutes length (4.5 minutes on average). 41 basic cooking tasks such as *cutting a cucumber* were recorded, each between 4 and 8 times. The selection of cooking tasks is based on those proposed at "Jamie's Home Cooking Skills".[13] The corpus is recorded in a kitchen environment with a total of 22 subjects. Each video depicts a single task executed by an individual subject.

The dataset contains expert annotations of low-level activity tags. Annotations are provided for segments containing a semantically meaningful cooking related movement pattern. The action must go beyond single body part movements (such as *move arm up*) and must have the goal of changing the state or location of an object. 60 different activity labels are used for annotation (e.g. PEEL, STIR, TRASH). Each low-level activity tag consists of an activity label (PEEL), a set of associated objects (CARROT, DRAWER,...), and the associated timeframe (starting and ending points of the activity). Associated objects are the participants of an activity, namely tools (e.g. KNIFE), patient (CARROT) and location (CUTTING-BOARD). We provide the coarse-grained role information for *patient*, *location* and *tool* in the corpus data, but we did not use this information in our experiments. The dataset contains a total of 8818 annotated segments, on average 42 per video.

[13] www.jamieshomecookingskills.com

8.3.2 Collecting textual video descriptions

We collected textual descriptions for a subset of the videos in *MPII Composites*, restricting collection to tasks that involve manipulation of cooking ingredients. We also excluded tasks with fewer than four video recordings in the corpus, leaving 26 tasks to be described. We randomly selected five videos from each task, except the three tasks for which only four videos are available. This resulted in a total of 127 videos. For each video, we collected 20 different textual descriptions, leading to 2540 annotation assignments. We published these assignments (HITs) on MTurk, using an adapted version[14] of the annotation tool Vatic (Vondrick *et al.*, 2012).

In each assignment, the subject saw one video specified with the task title (e.g. *How to prepare an onion*), and then was asked to enter at least five and at most 15 complete English sentences to describe the events in the video. The annotation instructions contained example annotations from a kitchen task not contained in our actual dataset.

Annotators were encouraged to watch each video several times, skipping backward and forward as they wished. They were also asked to take notes while watching, and to sketch the annotation before entering it. Once familiarized with the video, subjects did the final annotation by watching the entire video from beginning to end, without the possibility of further non-sequential viewing. Subjects were asked to enter each sentence as soon as the action described by the sentence was completed. The video playback paused automatically at the beginning of the sentence input. We recorded pause onset for each sentence annotation as an approximate ending timestamp of the described action. The annotators resumed the video manually.

The tasks required a HIT approval rate of 75% and were open only to workers in the US, in order to increase the general language quality of the English annotations. Each task paid 1.20 USD. Before paying we randomly inspected the annotations and manually checked for quality. The total costs of collecting the annotations amounted to 3,353 USD. The data was obtained within a time frame of 3.5 weeks.

8.3.3 Putting the TACoS corpus together

Our corpus is a combination of the MTurk data and MPII Composites, created by filtering out inappropriate material and computing a high-quality alignment of sentences and video segments. The alignment is done by matching the approximate timestamps of the MTurk data to the accurate timestamps in MPII Composites.

We discarded text instances if people did not time the sentences properly, taking the association of several (or even all) sentences to a single timestamp as an indicator. Whenever we found a timestamp associated with two or more sentences, we discarded the whole instance. Overall, we had to filter out 13% of the text instances, which left us with 2206 textual video descriptions.

For the alignment of sentence annotations and video segments, we assign a

[14] github.com/marcovzla/vatic/tree/bolt

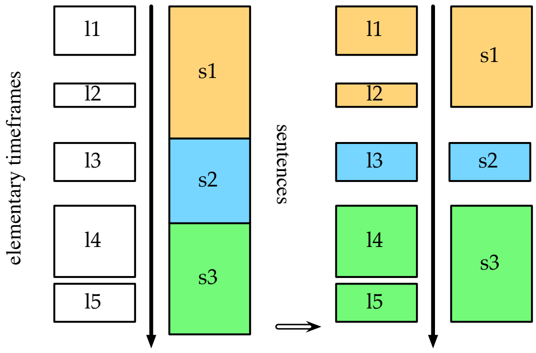


Figure 8.2: Aligning action descriptions with the video.

precise timeframe to each sentence in the following way: We take the timeframes given by the low-level annotation in MPII Composites as a gold standard micro-event segmentation of the video, because they mark all distinct frames that contain activities of interest. We call them *elementary frames*. The sequence of elementary frames is not necessarily continuous, because idle time is not annotated.

The MTurk sentences have end points that constitute a coarse-grained, noisy video segmentation, assuming that each sentence spans the time between the end of the previous sentence and its own ending point. We refine those noisy timeframes to gold frames as shown in Figure 8.2: Each elementary frame (*l1*-*l5*) is mapped to a sentence (*s1*-*s3*) if its noisy timeframe covers at least half of the elementary frame. We define the final gold sentence frame then as the timespan between the starting point of the first and the ending point of the last elementary frame.

The alignment of descriptions with low-level activities results in a table as given in Figure 8.3. Columns contain the textual descriptions of the videos; rows correspond to low-level actions, and each sentence is aligned with the last of its associated low-level actions. As a side effect, we also obtain multiple paraphrases for each sentence, by considering all sentences with the same associated time frame as equivalent realizations of the same action.

The corpus contains 17,334 action descriptions (tokens), realizing 11,796 different sentences (types). It consists of 146,771 words (tokens), 75,210 of which are content word instances (i.e. nouns, verbs and adjectives). The verb vocabulary comprises 28,292 verb tokens, realizing 435 lemmas. Since verbs occurring in the corpus typically describe actions, we can note that the linguistic variance for the 60 different low-level activities is quite large. Figure 8.4 gives an impression of the action realizations in the corpus, listing the most frequent verbs from the textual data, and the most frequent low-level activities.

On average, each description covers 2.7 low-level activities, which indicates a clear difference in granularity. 38% of the descriptions correspond to exactly one low-level activity, about a quarter (23%) covers two of them; 16% have 5 or more low-level elements, 2% more than 10. The corpus shows how humans vary the granularity of their descriptions, measured in time or number of low-level activities,

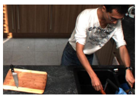
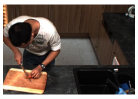
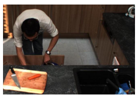
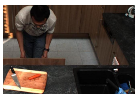
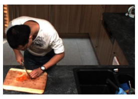
Sample frame	Start	End	Action	Participants	NL Sequence 1	NL Sequence 2	NL Sequence 3
	743	911	wash	hand, carrot	He washed carrot	The person rinses the carrot.	He rinses the carrot from the faucet.
	982	1090	cut	knife, carrot, cutting board	He cut off ends of carrots	The person cuts off the ends of the carrot.	He cuts off the two edges.
	1164	1257	open	hand, drawer			
	1679	1718	close	hand, drawer			He searches for something in the drawer, failed attempt, he throws away the edges in trash.
	1746	1799	trash	hand, carrot		The person searches for the trash can, then throws the ends of the carrot away.	
	1854	2011	wash	hand, carrot			He rinses the carrot again.
	2011	2045	shake	hand, carrot	He washed carrot	The person rinses the carrot again.	He starts chopping the carrot in small pieces.
	2083	2924	slice	knife, carrot, cutting board			
	2924	2959	scratch off	hand, carrot, knife, cutting board			
	3000	3696	slice	knife, carrot, cutting board	He diced carrots		He finished chopping the carrots in small pieces.

Figure 8.3: Excerpt from the TACoS corpus for a video on PREPARING A CARROT. Example frames, low-level annotation (*Action* and *Participants*) is shown along with three of the MTurk sequences (*NL Sequence 1-3*).

Top 10 Verbs	cut, take, get, put, wash, place, rinse, remove, **pan*, peel
Top 10 Activities	move, take out, cut, wash, take apart, add, shake, screw, put in, peel

Figure 8.4: 10 most frequent verbs and low-level actions in the TACoS corpus. **pan* is probably often mis-tagged as verb.

and it shows how they vary the linguistic realization of the same action. For example, Figure 8.3 contains *dice* and *chop into small pieces* as alternative realizations of the low-level activity sequence slice - scratch off - slice.

The descriptions are of varying length (9 words on average), reaching from two-word phrases to detailed descriptions of 65 words. Most sentences are short, consisting of a reference to the person in the video, a participant and an action verb (*The person rinses the carrot*, *He cuts off the two edges*). People often specified an instrument (*from the faucet*), or the resulting state of the action (*chop the carrots in small pieces*). Occasionally, we find more complex constructions (support verbs, coordinations).

As Figure 8.3 indicates, the timestamp-based alignment is pretty accurate; occasional errors occur like *He starts chopping the carrot...* in NL Sequence 3. The data contains some typos and ungrammatical sentences (*He washed carrot*), but for our own experiments, the small number of such errors did not lead to any processing problems.

8.4 the action similarity dataset

In this section, we present a gold standard dataset, as a basis for the evaluation of visually grounded models of action similarity. We call it the "Action Similarity Dataset" (ASim) in analogy to the Usage Similarity dataset (USim) of Erk *et al.* (2009, 2012). Similarly to USim, ASim contains a collection of sentence pairs with numerical similarity scores assigned by human annotators. We asked the annotators to focus on the similarity of the activities described rather than on assessing semantic similarity in general. We use sentences from the TACoS corpus and record their timestamps. Thus each sentence comes with the video segment which it describes (these were not shown to the annotators).

8.4.1 Selecting action description pairs

Random selection of annotated sentences from the corpus would lead to a large majority of pairs which are completely dissimilar, or difficult to grade (e.g., *He opens the drawer – The person cuts off the ends of the carrot*). We constrained the selection process in two ways: First, we consider only sentences describing activities of manipulating an ingredient. The low-level annotation of the video corpus helps us identify candidate descriptions. We exclude rare and special activities, ending up with CUT, SLICE, CHOP, PEEL, TAKE APART, and WASH, which occur reasonably frequently, with a wide distribution over different scenarios. We restrict the candidate set to those sentences whose timespan includes one of these activities. This results in a conceptually more focussed repertoire of descriptions, and at the same time admits full linguistic variation (*wash an apple under the faucet – rinse an apple, slice the cucumber – cut the cucumber into slices*).

Second, we required the pairs to share some lexical material, either the head verb or the manipulated ingredient (or both).[15] More precisely, we composed the ASim dataset from three different subsets:

Different activity, same object: This subset contains pairs describing different types of actions carried out on the same type of object (e.g. *The man washes the carrot. – She dices the carrot.*). Its focus is on the central task of modeling the semantic relation between *actions* (rather than the objects involved in the activity), since the object head nouns in the descriptions are the same, and the respective video segments show the same type of object.

Same activity, same object: Description pairs of this subset will in many cases, but not always, agree in their head verbs. The dataset is useful for exploring the degree to which action descriptions are underspecified with respect to the precise manner of their practical realization. For example, peeling an onion will mostly be done in a rather uniform way, while *cut* applied to *carrot* can mean that the carrot is chopped up, or sliced, or cut in halves.

Same activity & verb, different object: Description pairs in this subset share head verb and low-level activity, but have different objects (e.g. *The man washes the carrot. – A girl washes an apple under the faucet.*). This dataset enables the exploration of the objects' meaning contribution to the complete action, established by the variation of equivalent actions that are done to different objects.

We assembled 900 action description pairs for annotation: 480 pairs share the object; 240 of which have different activities, and the other 240 pairs share the same activity. We included paraphrases describing the same video segment, but we excluded pairs of identical sentences. 420 additional pairs share their head verb, but have different objects.

8.4.2 Manual annotation

Three native speakers of English were asked to judge the similarity of the action pairs with respect to *how they are carried out*, rating each sentence pair with a score from 1 (not similar at all) to 5 (the same or nearly the same). They did not see the

[15] We refer to the latter with the term *object*; we don't require the ingredient term to be the actual grammatical object in the action descriptions, we rather use "object" in its semantic role sense as the entity affected by an action.

Part of Gold Standard	Sim	σ	ρ
different activity, same object	2.20	1.07	0.73
same activity, same object	4.19	1.04	0.73
all with same object	3.20	1.44	0.84
same activity & verb, different object	3.34	0.69	0.43
complete dataset	3.27	1.15	0.73

Figure 8.5: Average similarity ratings (*Sim*), their standard deviation (σ)) and annotator agreement (ρ) for ASim.

respective videos, but we noted the relevant kitchen task (i.e. which vegetable was prepared). We asked the annotators explicitly to ignore the actor of the action (e.g. whether it is a man or a woman) and score the similarities of the underlying actions rather than their verbalizations. Each subject rated all 900 pairs, which were shown to them in completely random order, with a different order for each subject.

We compute inter-annotator agreement (and the forthcoming evaluation scores) using Spearman's rank correlation coefficient (ρ), a non-parametric test which is widely used for similar evaluation tasks (Mitchell and Lapata, 2008; Bruni *et al.*, 2011; Erk and McCarthy, 2009). Spearman's ρ evaluates how the samples are ranked relative to each other rather than the numerical distance between the rankings.

Figure 8.5 shows the average similarity ratings in the different settings and the inter-annotator agreement. The average inter-rater agreement was $\rho = 0.73$ (averaged over pairwise rater agreements), with pairwise results of $\rho = 0.77$, 0.72, and 0.69, respectively, which are all highly significant at $p < 0.001$.

As expected, pairs with the same activity and object are rated very similar (4.19) on average, while the similarity of different activities on the same object is the lowest (2.2). For both subsets, inter-rater agreement is high ($\rho = 0.73$), and even higher for both same object subsets together (0.84).

Pairs with identical head verbs and different objects have a small standard deviation, at 0.69. The inter-annotator agreement on this set is much lower than for pairs from the same object set. This indicates that similarity assessment for different variants of the same activity is a hard task even for humans.

8.5 models of action similarity

In the following, we demonstrate that visual information contained in videos of the kind provided by the TACoS corpus (Section 8.3) substantially contributes to the semantic modeling of action-denoting expressions. In Section 8.6, we evaluate several methods for predicting action similarity on the task provided by the ASim dataset. In this section, we describe the models considered in the evaluation. We use two different models based on visual information, and in addition two text based models. We will also explore the effect of combining linguistic and visual information and

Model		same object	same verb	overall
text	jaccard	0.28	0.25	0.25
	textual vectors	0.30	0.25	0.27
	text combined	0.39	**0.35**	0.36
video	visual raw vectors	0.53	-0.08	0.35
	visual classifier	0.60	0.03	0.44
	video combined	0.61	-0.04	0.44
mix	All unsupervised	0.58	0.32	0.48
	All combined	**0.67**	0.28	**0.55**
upper bound		0.84	0.43	0.73

Figure 8.6: Evaluation results in Spearman's ρ. All values > 0.11 are significant at $p < 0.001$.

investigate which mode is most suitable for which kinds of similarity.

8.5.1 Text-based models

We use two different models of textual similarity to predict action similarity: a simple word-overlap measure (Jaccard coefficient) and a state-of-the-art model based on "contextualized" vector representations of word meaning (Thater *et al.*, 2011).

8.5.1.1 *Jaccard coefficient.*

The Jaccard coefficient gives the ratio between the number of (distinct) words common to two input sentences and the total number of (distinct) words in the two sentences. Such simple surface-oriented measures of textual similarity are often used as baselines in related tasks such as recognizing textual entailment (Dagan *et al.*, 2005) and are known to deliver relatively strong results.

8.5.1.2 *Vector model.*

We use the vector model of Thater *et al.* (2011), which "contextualizes" vector representations for individual words based on the particular sentence context in which the target word occurs. The basic intuition behind this approach is that the words in the syntactic context of the target word in a given input sentence can be used to refine or disambiguate its vector. Intuitively, this allows us to discriminate between different actions that a verb can refer to, based on the different objects of the action.

We first experimented with a version of this vector model which predicts action similarity scores of two input sentences by computing the cosine similarity of the contextualized vectors of the verbs in the two sentences only. We achieved better performance with a variant of this model which computes vectors for the two sentences by summing over the contextualized vectors of all constituent content words.

In the experiments reported below, we only use the second variant. We use the same experimental setup as Thater *et al.* (2011), as well as the parameter settings they reported to work best.

8.5.2 Video-based models

We distinguish two approaches to compute the similarity between two video segments. In the first, unsupervised approach we extract a video descriptor and compute similarities between these raw features (Wang *et al.*, 2011). The second approach builds upon the first by additionally learning higher level attribute classifiers as in Chapter 6 on a held out training set. The similarity between two segments is then computed between the classifier responses. In the following we detail both approaches:

8.5.2.1 *Raw visual features.*

We use the state-of-the-art video descriptor *Dense Trajectories* (Wang *et al.*, 2011) which extracts visual video features, namely histograms of oriented gradients, flow, and motion boundary histograms, around densely sampled and tracked points.

This approach is especially suited for this data as it ignores non-moving parts in the video: we are interested in activities and manipulation of objects, and this type of feature implicitly uses only information in relevant image locations. Using a bag-of-words representation we encode the features using a 16,000 dimensional codebook. In Chapter 5 we showed that this feature representation is superior to human pose-based approaches for our videos.

We compute the similarity between two encoded features by computing the intersection of the two (normalized) histograms.

8.5.2.2 *Visual classifiers.*

Visual raw features tend to have several dimensions in the feature space which provide unreliable, noisy values and thus degrade the strength of the similarity measure. Intermediate level attribute classifiers can learn which feature dimensions are distinctive and thus significantly improve performance over raw features. In Chapter 6 we showed that using such an attribute classifier representation can significantly improve performance for composite activity recognition. The relevant attributes are all activities and objects annotated in the video data (cf. Section 8.3.1). For the experiments reported below we use the same setup as in Chapter 6 and use all videos in MPII Composites and MPII Cooking introduced in Chapter 5 and 6, excluding the 127 videos used during evaluation. The real-valued SVM-classifier output provides a confidence how likely a certain attribute appeared in a given video segment. This results in a 218-dimensional vector of classifier outputs for each video segment. To compute the similarity between two vectors we compute the cosine between them.

8.6 EVALUATION

We evaluate the different similarity models introduced in Section 8.5 by calculating their correlation with the gold-standard similarity annotations of ASim (cf. Section 8.4). For all correlations, we use Spearman's ρ as a measure. We consider the two textual measures (JACCARD and TEXTUAL VECTORS) and their combination, as well as the two visual models (VISUAL RAW VECTORS and VISUAL CLASSIFIER) and their combination. We also combined textual and visual features, in two variants: The first includes all models (ALL COMBINED), the second only the unsupervised components, omitting the visual classifier (ALL UNSUPERVISED). To combine multiple similarity measures, we simply average their normalized scores (using z-scores).

Figure 8.6 shows the scores for all of these measures on the complete ASim dataset (OVERALL), along with the two subparts, where description pairs share either the object (SAME OBJECT) or the head verb (SAME VERB). In addition to the model results, the table also shows the average human inter-annotator agreement as UPPER BOUND.

On the complete set, both visual and textual measures have a highly significant correlation with the gold standard, whereas the combination of both clearly leads to the best performance (0.55). The results on the SAME OBJECT and SAME VERB subsets shed light on the division of labor between the two information sources. While the textual measures show a comparable performance over the two subsets, there is a dramatic difference in the contribution of visual information: On the SAME OBJECT set, the visual models clearly outperform the textual ones, whereas the visual information has no positive effect on the SAME VERB set. This is clear evidence that the visual model does not capture the similarity of the participating objects but rather genuine action similarity, which the visual features (Wang *et al.*, 2011) we employ were designed for. A direction for future work is to learn dedicated visual object detectors to recognize and capture similarities between objects more precisely.

The numbers shown in Table 8.1 support this hypothesis, showing the two groups in the SAME OBJECT class: For sentence pairs that share the same activity, the textual models seem to be much more suitable than the visual ones. In general, visual models perform better on actions with different activity types, textual models on closely related activities.

Overall, the supervised classifier contributes a good part to the final results. However, the supervision is not strictly necessary to arrive at a significant correlation; the raw visual features alone are sufficient for the main performance gain seen with the integration of visual information.

	Model (Same Object)	*same action*	*diff. action*
text	jaccard	0.44	0.14
	text vectors	0.42	0.05
	text combined	**0.52**	0.14
video	vis. raw vectors	0.21	0.23
	vis. classifier	0.21	**0.45**
	video combined	0.26	0.38
mix	All unsupervised	0.49	0.24
	all combined	0.48	0.41
upper bound		0.73	0.73

Table 8.1: Results for sentences with the same object, with either the same or different low-level activity.

8.7 conclusion

We presented the TACoS corpus, which provides coherent textual descriptions for high-quality video recordings, plus accurate alignments of text and video on the sentence level. We expect the corpus to be beneficial for a variety of research activities in natural-language and visual processing.

In this chapter we focused on the task of grounding the meaning of action verbs and phrases. We designed the ASim dataset as a gold standard and evaluated several text- and video-based semantic similarity models on the dataset, both individually and in different combinations.

We are the first to provide semantic models for action-describing expressions, which are based on information extracted from videos. Our experimental results show that these models are of considerable quality, and that predictions based on a combination of visual and textual information even approach the upper bound given by the agreement of human annotators.

In this work we used existing similarity models that had been developed for different applications. We applied these models without any special training or optimization for the current task, and we combined them in the most straightforward way. There is room for improvement by tuning the models to the task, or by using more sophisticated approaches to combine modality-specific information (Silberer and Lapata, 2012).

We built our work on an existing corpus of high-quality video material, which is restricted to the cooking domain. As a consequence, the corpus covers only a limited inventory of activity types and action verbs. Note, however, that our models are fully unsupervised (except the Visual Classifier model), and thus can be applied without modification to arbitrary domains and action verbs, given that they are about observable activities. Also, corpora containing information comparable to the TACoS corpus but with wider coverage (and perhaps a bit noisier) can be obtained with a moderate amount of effort. One needs videos of reasonable quality and some sort of alignment with action descriptions. In some cases such alignments even come for free, e.g. via subtitles, or descriptions of short video clips that depict just a single action.

The TACoS corpus and all other data described in this chapter (videos, low-level annotation, aligned textual descriptions, the ASim-Dataset and visual features) are publicly available at http://www.coli.uni-saarland.de/projects/smile/page.php?id=tacos.

For future work, we will further investigate the compositionality of action-describing phrases. We also want to leverage the multimodal information provided by the TACoS corpus for the improvement of high-level video understanding. In the next chapter we use this corpus to learn the generation of natural language text from videos.

TRANSLATING VIDEO CONTENT TO NATURAL LANGUAGE DESCRIPTIONS 9

Contents

After Chapters 3, 4 and 5 showed the benefits of linguistic knowledge for visual recognition and Chapter 8 the benefit of visual recognition for language processing, this chapter shows how to convert from the visual to the linguistic modality. While humans use rich natural language to describe and communicate visual perceptions, it is challenging to automate this process. In order to learn how to generate natural language descriptions for visual content we combine two important ingredients. First, we generate a rich semantic representation of the visual content including e.g. object and activity labels. To predict the semantic representation we learn a CRF to model the relationships between different components of the visual input. And second, we propose to formulate the generation of natural language as a machine translation problem using the semantic representation as source language and the generated sentences as target language. For this we exploit the power of a parallel corpus of videos and textual descriptions and adapt statistical machine translation to translate between our two languages. We evaluate our video descriptions on the TACoS dataset introduced in the previous chapter, which contains video snippets aligned with sentence descriptions. Using automatic evaluation and human judgments we show significant improvements over several baseline approaches, motivated by prior work. Our translation approach also shows improvements over related work on an image description task.

9.1 INTRODUCTION

Computer vision has advanced to detect people, classify their actions, or to distinguish between a large number of objects and specify their attributes. The output is often a semantic representation encoding activities and objects categories. While such representations can be well processed by automated systems, the natural way to communicate this information with humans is natural language. Thus, this work addresses the problem of generating textual descriptions for videos. This task has a wide range of applications in the domain of human-computer/robot interaction, generating summary descriptions of (web-)videos, and automating movie descriptions for visually impaired people. Furthermore, being able to convert visual content to language is an important step in understanding the relationship between visual and linguistic information which are the richest interaction modalities available to humans.

Generating natural language descriptions of visual content is an intriguing task but requires combining the fundamental research problems of visual recognition and natural language generation (NLG). While for descriptions of images, recent approaches have proposed to statistically model the conversion from images to text (Farhadi *et al.*, 2010b; Kulkarni *et al.*, 2011; Kuznetsova *et al.*, 2012; Mitchell *et al.*, 2012), most approaches for video description use rules and templates to generated video descriptions (Kojima *et al.*, 2002; Gupta *et al.*, 2009; Barbu *et al.*, 2012; Hanckmann *et al.*, 2012; Khan *et al.*, 2011; Tan *et al.*, 2011; Das *et al.*, 2013b; Guadarrama *et al.*, 2013a). Although these works have started exploring the domain of describing visual content, important research questions remain: (1) How to best approach the conversion from visual information to linguistic expressions? (2) Which part of the visual information is verbalized by humans and what is verbalized even though it is not directly present in the visual information? (3) What is a good semantic representation (SR) of visual content and what is the limit of such a representation given perfect visual recognition?

Answering these questions is clearly beyond the scope of this thesis but we aim to address them jointly here. To address the first question we suggest to learn the conversion from video to language descriptions in a two-step approach. In the first step we learn an intermediate SR using a probabilistic model, following ideas used to generate image descriptions (Farhadi *et al.*, 2010b; Kulkarni *et al.*, 2011). Then, given the SR, we propose to phrase the problem of NLG as a *translation problem*, which means translating the SRs to natural language descriptions. In contrast to related work on video description, we learn both the SR as well as the language descriptions from an aligned parallel corpus containing videos, semantic annotations and textual descriptions. We compare our approach to related work and baselines using no intermediate SR and/or language model.

Second, we do not want to define manually the right level of verbalization. Instead we learn from a parallel training corpus the most relevant information to verbalize and how to verbalize it. For this we employ the methods from statistical machine translation (Koehn, 2010). (a) We learn the correct ordering of words

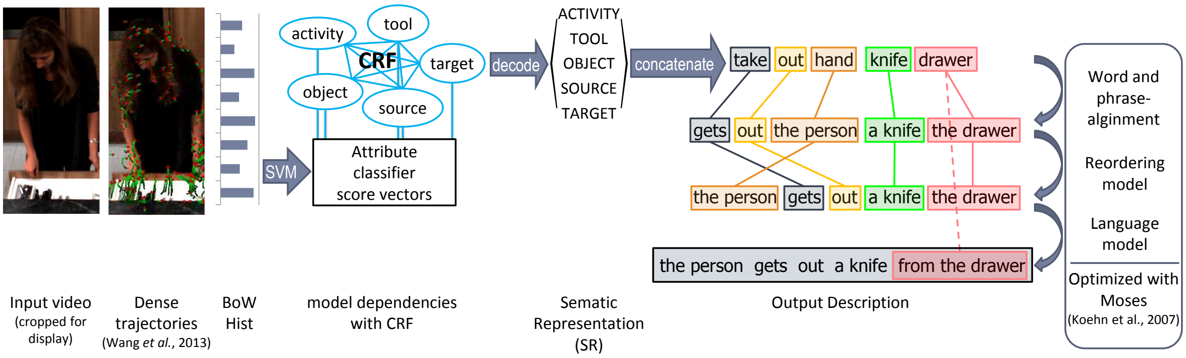

activity
tool
CRF
target
object
source
SVM
Attribute classifier score vectors
decode
ACTIVITY
TOOL
OBJECT
SOURCE
TARGET
concatenate
take out hand knife drawer
gets out the person a knife the drawer
the person gets out a knife the drawer
the person gets out a knife from the drawer
Word and phrase-alignment
Reordering model
Language model
Optimized with Moses (Koehn et al., 2007)
Input video (cropped for display)
Dense trajectories (Wang et al., 2013)
BoW Hist
model dependencies with CRF
Sematic Representation (SR)
Output Description

Figure 9.1: Overview of our approach for describing videos with natural language at test time. We first extract dense trajectories from a full-frame, but pre-segmented video snipped and encode them in a Bag-Of-Words histogram. This feature vector is classified into attributes of activities of objects. These classifiers build the unary features of a CRF which predicts the semantic representation of ⟨activity, tool, object, source, target⟩. We concatenate this representations to a string which serves as input language for a translation pipeline which translates it to a natural language sentence. All steps are learned from a parallel corpus of video, semantic representation, and descriptions. Details are described in Section 9.3.

and phrases, referred to as surface realization in NLG. (b) We can learn which SR should be realized in language. When describing a video, using "cooking" as a running example, the visually recognized object peeler would normally not be mentioned when describing that *a person is peeling a carrot* but can still contribute to the verbalization of *peeling*. (c) We learn the proper correspondence between semantic concepts and verbalization, i.e. we do not have to define how semantic concepts are realized. For example the concepts ⟨move, pan, counter, hob⟩ could be realized as *He puts the frying pan on the stove* rather than being limited to *He moves the pan from the counter to the hob* when just adding function words.

Although NLG can be defined purely by rules and templates which might provide a more robust approach for limited domains, we believe that learning these parameters from data is a much more attractive approach. For any sufficiently rich domain, the required complexity of rules and templates is likely to make the rule engineering task either infeasible or prohibitively expensive. This has been shown for language translation, where statistical machine translation has generally replaced rule-based approaches (Koehn, 2010).

To address the third question of the right visual input we compare three different visual representations, namely a raw video descriptor (Wang *et al.*, 2013), our attribute based representation from Chapter 6, and our CRF model. To understand the limits of our SR we also run the translation on ground truth annotations.

In Figure 9.1 we give an overview of our two-step approach at test time. The details are described in Section 9.3.

The main contributions are as follows. First, we phrase video description as a translation problem from video content to natural language descriptions (Section 9.3). As intermediate step we employ a SR of the video content. Second, in Section 9.5 we evaluate our approach on the TACoS video-description dataset introduced in Chapter 8. Using automatic as well as human evaluation, the proposed approach outperforms several baseline methods inspired by previous work. The SR, when using ground truth annotations, allows generating language that is close to human performance. Additionally our approach also compares favorably to Farhadi *et al.* (2010b) on the Pascal-sentence dataset for an image description task (Section 9.6). Third, annotations as well as intermediate outputs and final descriptions to allow for comparisons to our work or building on our SR are released on our website.

9.2 RELATED WORK

9.2.1 Statistical machine translation (SMT)

Machine translation aims to translate from one natural language to another. SMT formulates this problem as data-driven machine learning problem. SMT is a mature field with existing approaches achieving respectable results across many language pairs, see e.g. Lopez (2008) for a review and tutorial. Based on sentence-aligned corpora of source and target language a translation model is estimated. Additionally, a model for the target language is learnt to generate a fluent and grammatical output. The open source Moses toolkit (Koehn *et al.*, 2007) optimizes this pipeline on a training set (see Section 9.3.2).

Duygulu *et al.* (2002) propose to approach object recognition in analogy to machine translation by learning a lexicon from images segments to associated keywords from images with keywords. Rather than translating to words or labels we translate from a SR to full descriptions. Matuszek *et al.* (2010) apply statistical machine translation to translate natural language instruction to a formal language which is used direct robots using the Word Alignment-based Semantic Parser (WASP, Wong and Mooney, 2006) .

9.2.2 Natural language generation from images and video

Generating descriptions of visual content can be roughly divided in four different directions according to: (1) generating descriptions for (test) images or videos which already contain some associated text, (2) generating descriptions by using manually defined rules or templates, (3) retrieving existing descriptions from similar visual content, or (4) learning a language model from a training corpus to generate descriptions.

(1) Assuming the availability of text associated with the image at test time one can effectively use summarization techniques (Aker and Gaizauskas, 2010; Feng and Lapata, 2010) which benefit from visual content. This setting is different from ours as we want to generate descriptions at test time from visual content only.

(2) Given a SR extracted from visual content it is possible to generate language using manually defined rules and templates. To describe images, Kulkarni *et al.* (2011) extract objects and their attributes as well as their spatial prepositions from images. These entities are modeled in a Conditional Random Field (CRF). From the CRF predictions they generate descriptions based on simple templates (or n-gram model, which falls into (4)). We also use a CRF to predict an intermediate SR but we show that our translation system generates descriptions more similar to human descriptions. For videos, Kojima *et al.* (2002) build a concept hierarchy of actions which is manually defined and associated with different body, hand and head movements. Our setting is visually more challenging and varied making manual definitions challenging. Tan *et al.* (2011) learn audio-visual concepts and generates a video description for three different activities using rules to combine action, scene, and audio concepts with glue words. Gupta *et al.* (2009) extract an AND-OR graph from sports videos to model causal relationships. Using the graph, sentences can then be constructed using simple templates. Hanckmann *et al.* (2012) and Barbu *et al.* (2012) extract actions, body-pose, objects and their tracks on the DARPA Mind's eye corpus which depict 48 different verbs. Using a set of templates they generate text for their SR. Similarly, Khan *et al.* (2011) use templates to describe videos on the TREC Video summarization task. Das *et al.* (2013b) follow a different route and uses a topic model to jointly model textual and visual words and a tripartite graph based on object/concept detectors. Text generation is done with manually defined templates and retrieval from the training corpus. Guadarrama *et al.* (2013a) predict multiple subject-verb-object triples for a video snippet. These are reweighed according to the confidence along a classifier hierarchy and a language model. The best suited triple is used to generate multiple sentences based on a template which are again scored against a n-gram language model. Similarly, our translation approach weights resulting sentences according to a language model. However, using templates limits the natural flexibility of language, as noted by Kuznetsova *et al.* (2012).

(3) The third group of approaches reduces the generation process to retrieving sentences from a training corpus based on locally (Ordonez *et al.*, 2011) or globally (Farhadi *et al.*, 2010b) similar images. Farhadi *et al.* (2010b) learn an intermediate SR of object, action, and scenes using a Markov Random Field. We compare to their retrieval results by applying our translation approach to their SR.

(4) The fourth line of work, which also includes this work, goes beyond retrieving existing descriptions by learning a language model to compose novel descriptions. Kulkarni *et al.* (2011) learn an n-gram language model to predict function words for their SR. One of our baselines is based on this idea (Section 9.4.2). Two recent approaches use an aligned corpus of images and descriptions as a basis for generating novel descriptions for images using state-of-the art language generation techniques. Kuznetsova *et al.* (2012) retrieves candidate phrases from an image-caption database based on object, scene, and region recognition. Using an Integer Linear Programming formulation for content planning and surface realization they construct the most relevant and linguistically coherent descriptions. While they hand craft constraints to translate from the image, we learn a statistical translation model. Mitchell *et al.* (2012)

use a corpus of 700,000 Flickr images with associated descriptions. Based on the visual recognition system of Kulkarni *et al.* (2011) they learn to predict sets of nouns and their order and add necessary prepositions, predicates, and determiners to form syntactically well-formed phrases. In contrast to their Tree-adjoining-grammar (TAG)-like natural language generation approach we use flat, co-occurrence based techniques from SMT.

9.3 VIDEO DESCRIPTION AS A TRANSLATION PROBLEM

In this section we present a two-step approach which describes video content with natural language. We assume that for training we have a parallel corpus which contains a set of video snippets and sentences. Video snippets represented by the video descriptor x_i are aligned with a sentence z_i, i.e. we have (x_i, z_i). In case there is an extra description for the same video snippet we treat it as an independent alignment (x_k, z_k) with $x_k = x_i$. Additionally we introduce an intermediate level semantic representation (SR) in form of labels y_i.

At test time we first predict the SR y^* for a new video (descriptor) x^* and then generate a sentence z^* from y^*.

In the following we present our proposed approach using human-activity videos in a kitchen scenario based on the TACoS corpus, where people are recorded preparing different kinds of ingredients. However, we show in Section 9.6 that our approach can also be applied to translate images to descriptions.

We build the SR based on the annotations provided with the TACoS corpus which we introduced in Chapter 8. It distinguishes *activities*, *tools*, *ingredients/objects*, *(source) location/container*, and *(target) location/container*. This directly converts to our SR y in the form of ⟨ACTIVITY, TOOL, OBJECT, SOURCE, TARGET⟩. As a tool, object or location can be missing, we represent this with an additional NULL label for the respective node.

The SR annotations in TACoS have sometimes a finer granularity than the sentences, i.e. $(y_i^1, \ldots, y_i^{l_i}, \ldots, y_i^{L_i}, z_i)$ where L_i is the number of SR annotations for sentence z_i. For learning the SR we just extract the corresponding video snippet for the SR, i.e. $(x_i^{l_i}, y_i^{l_i})$. As there are no annotations at test time, there exist no alignment problem when predicting y^*. In Section 9.3.2 we discuss several variants how to handle the different granularity of the SR and the sentences.

9.3.1 Predicting a SR from visual content

In the first step we extract a SR from the visual content. Typically different visual information is highly correlated with each other. E.g. for cooking activities, the activity *slice* is more correlated with the object *carrot* and tool *knife* than with *milk* and *spoon*. We model these relationships with a CRF where the visual entities are modeled as nodes n_j observing the video descriptors x as unaries. In our case we use a fully connected graph and learn linear pairwise (p) and unary (u) weights, using the following standard energy formulation for the structured model:

$$E(n_1, ..., n_N; x_i) = \sum_{j=1}^{N} E^u(n_j; x_i) + \sum_{j \sim k} E^p(n_j, n_k) \tag{9.1}$$

with $E^u(n_j; x_i) = \langle w_j^u, x_i \rangle$, where w_j^u is a vector of the size of the video representation x_i and $E^p(n_j, n_k) = w_{j,k}^p$.

We learn the model with training videos $x_i^{l_i}$ and SR labels $y_i^{l_i} = \langle n_1, n_2, \ldots, n_N \rangle$ using loopy belief propagation (LBP) with the implementation from Schmidt (2013). We model the five SR categories as nodes ($N = 5$), the different states are based on the provided labels of TACoS (for samples see Table 9.1).

9.3.2 Translating from a SR to a description

Converting a SR to descriptions ($SR \rightarrow D$) has many similarities to translating from a source to a target language ($L_S \rightarrow L_T$) in machine translation.

1. For $SR \rightarrow D$ we have to find the verbalization of a label n_i, e.g. HOB→*stove*, similar to translating a word from L_S to L_T.
2. For $SR \rightarrow D$ we have to determine the ordering of the concepts of the SR in D, which is similar to finding the alignment between two languages.
3. In a natural description of video not necessarily all semantic concepts are verbalized, e.g. KNIFE might not be verbalized when we describe *He cuts a carrot*. There exists a similar problem for $L_S \rightarrow L_T$, where certain words in L_S, e.g. articles, are either not represented in L_T or multiple ones are combined to one.
4. The inverse problem also exist, e.g. adding function words to the SR to form a full sentence, e.g. CUT, CARROT→*He cuts the carrots.*
5. When translating $L_S \rightarrow L_T$ a language model of L_T is used to achieve a grammatically correct and fluent target sentence, same for D in $SR \rightarrow D$.

Motivated by these similarities, we propose to use established techniques for statistical machine translation (SMT) to learn a translation model from a parallel corpus of SRs and descriptions. We use the widely used Moses toolkit (Koehn *et al.*, 2007) to learn a translation model and in the following shortly layout the steps taken.

First we have to build a parallel corpus. In TACoS we encounter the problem that one sentence can be aligned to multiple SRs, i.e. $(y_i^1, \ldots, y_i^{L_i}, z_i)$. However, the input for SMT is aligned single sentences. We propose the following variants to handle the different granularity levels of SRs and descriptions:

All. For all SR annotations aligned to a sentence we create a separate training example, i.e. $(y_i^1, z_i), \ldots, (y_i^{L_i}, z_i)$.

Last. We only use the last SR as this frequently is the most important one, which is an artifact of the recording of the TACoS dataset, where users indicate only the ending time of their description in the video, i.e. $(y_i^{L_i}, z_i)$.

Semantic overlap. We estimate the highest word overlap between the sentence and the string of the SR: $\frac{|y_i \cap Lemma(z_i)|}{|y_i|}$, where *Lemma* refers to lemmatizing, i.e. reducing to base forms, e.g. *took* to *take*, *knives* to *knife*.

Sentence level prediction. While we do not have an annotated SR for the sentence level, we can predict one SR for each sentence, i.e. y_i^* for z_i. While this will be noisier during training time it also reflects better the situation at test time where we also have predictions at sentence level as annotations are unavailable.

SMT expects an input string as source language expression. We convert our SR ⟨ACTIVITY, TOOL, OBJECT, SOURCE, TARGET⟩ in a string by concatenating the concepts using spaces as delimiters to indicate word boundaries, i.e. *activity tool object source target*, where NULL states are converted to empty strings.

Next we use giza++ (Och and Ney, 2003) to learn a word-level alignment, i.e. in our case concepts-word alignment. This is the basis for the phrase-based translation model learned by Moses, which does not look at single words but tries to find multiple words (phrases) which correspond to each other and the corresponding probability. Additionally a reordering model is learned based on the training data alignment statistics (Koehn *et al.*, 2007).

To estimate the fluency of the descriptions we use IRSTLM (Federico *et al.*, 2008) which is based on n-gram statistics of TACoS.

The final step involves optimizing a linear model between the probabilities from the language model, phrase tables, and reordering model, as well as word, phrase, and rule counts (Koehn *et al.*, 2007). For this we use 10% of the training data as a validation set. In the optimization, the BLEU@4 score is used to compute the difference between predicted and provided reference descriptions.

For testing, we apply our translation model to the SR y^* predicted by the CRF for a given input video x^*. This decoding results in the description z^*.

9.4 BASELINES

In the following we describe baselines which are motivated by related work and which fully or partially replace our translation approach. For all these variants we use the same setup as for our translation system, see Section 9.5.

9.4.1 Sentence retrieval

An alternative to generating novel descriptions is to retrieve the most likely sentence from a training corpus (Farhadi *et al.*, 2010b). Given a test video x^* we search for the closest training video x_i and output the sentence $z^* = z_i$ (in case there are several we choose the first). To measure the distance between videos we distinguish three variants:

Raw video features. We use the L2-distance between BoW quantized dense trajectory representations (Wang *et al.*, 2013). This requires no intermediate level annotation of the data.

Attribute classifiers. While the raw video features tend to be too noisy to compute reliable distances, using the vector of attribute classifier outputs instead of the raw video features improves similarity estimates between videos as shown in Chapter 8.

CRF predictions. We use the estimated configuration to find the most similar SR in training data using hamming distance. This is the most similar variant to Farhadi *et al.* (2010b) which also use a probabilistic graphical model to represent the intermediate representation.

9.4.2 Natural language generation with N-grams

While we keep the same SR we replace the SMT pipeline by learning a n-gram language model on the training set of the descriptions. It predicts function words between the content words from the SR-labels, similar to one of the approaches discussed by Kulkarni *et al.* (2011). For the n-gram model to work we have do manually define the following steps: 1) the order of the content words has to be identical to the ones in the target sentence; 2) for our corpus, tool and location is frequently not verbalized, thus our model could only find a sensible string when we reduced it to ACTIVITY and OBJECT; 3) to further improve performance we only use the verb in the activity, e.g. CUT DICE→*cut*, and the root word for noun phrases, e.g. PLASTIC BAG→*bag*.

9.5 EVALUATION: TRANSLATING VIDEO TO TEXT

We evaluate our video description approach on the TACoS dataset introduced in Chapter 8 which contains videos with aligned SR annotations and sentence descriptions. We use an updated version of TACoS with a total of 18,227 video/sentence pairs on 7,206 unique time intervals. There are 5609 intermediate level annotations, which form our semantic representation (SR) and consists of the tuple ⟨ACTIVITY, TOOL, OBJECT, SOURCE, TARGET⟩.

To describe the video we use the dense trajectory features (Wang *et al.*, 2013) which extract trajectory information, HOG, HOF, and MBH to form a descriptor which has shown state-of-the art performance on many activity recognition datasets, including our dataset as shown in Chapter 5. As our final video descriptor and input for the CRF we use our attribute-classifier representation from Chapter 6 which includes both actions and objects on top of the dense trajectory features.

We test our approach on a subset of 490 video snippet / sentence pairs. There is no overlap in the human subjects to the training data. The CRF and Moses are trained on the remaining TACoS corpus, using 10% as a validation set for parameter estimation. The attribute classifiers are trained on the remaining videos of our MPII Cooking Composite Activity dataset, which is a superset of TACoS. We preprocess all text data by substituting gender specific identifiers with "the person" as we do not distinguish male and female with our visual system.

Node	states	Example states	SVM	LBP
ACTIVITY	66	cut dice, pour, stir, peel	58.7	**60.8**
TOOL	43	fork, hand, knife, towel	81.6	**82.0**
OBJECT	109	bread, carrot, salt, pot	32.5	**33.2**
SOURCE	51	fridge, plate, cup, pot	**76.0**	71.0
TARGET	35	counter, plate, hook	**74.9**	70.3
All nodes correct			18.7	**21.6**

Table 9.1: CRF nodes of our SR. SVM vs. LBP inference: Node accuracy in % over all test sentences.

We evaluate automatically using the BLEU score which is widely used to evaluate machine translations against reference translations (Papineni *et al.*, 2002). It computes the geometric mean of n-gram word overlaps for n=1,. . . ,N, weighted by a brevity penalty. While BLEU@4 (N=4) has shown to provide the best correlation with human judgments, we also provide BLEU@1 to comply with results reported in (Kuznetsova *et al.*, 2012; Kulkarni *et al.*, 2011). For manual evaluation, we follow (Kuznetsova *et al.*, 2012) and ask 10 human subjects to rate grammatical correctness (independent of video content), correctness, and relevance (latter two independent of grammatical correctness). Correctness rates if the sentences are correct with respect to the video, and relevance judges if the sentence describes the most salient activity and objects. We additionally ask the judges to separately rate the correctness of the activity, objects (tools and ingredients), and locations described. We ask to rate on a scale from 1 to 5 with 5: perfect, 4: almost perfect, 3:70-80% good, 2: 50-70% good, 1: totally bad (Kuznetsova *et al.*, 2012).

We present the human judges with different sentences of our systems in a random order for each video and ask explicitly to make consistent relative judgment between different sentences. If needed, continuous scores (e.g. 3.5) can be assigned. We limit our human evaluation to the best and most discriminant approaches.

In Table 9.1 we evaluate our visual recognition system, reporting accuracy over all test sentences for the different nodes.

9.5.1 Results: Translating video to text

Results of the various baselines and from our translation system are provided in Table 9.2 and typical sample outputs of our approach and baseline systems are shown in Table 9.4. We start by comparing the evaluation according to BLEU scores which is available for all approaches. We first examine the baseline approaches. When retrieving the closest sentence from the training data based on the raw video features (first row in Table 9.2), we obtain BLEU@4 of 6.0%. By replacing the raw features with the higher level representations of attribute classifier outputs and the CRF prediction we improve to 12.0% and 13.0% BLEU@4 respectively, where the latter one is similar to the concept presented by Farhadi *et al.* (2010b) for image

	BLEU in %		Human judgments		
Approach	@4	@1	Grammar	Correctness	Relevance
Baselines					
Sentence retrieval (raw video features)	6.0	32.3			
Sentence retrieval (attributes classifiers)	12.0	39.9	4.6	2.3 (3.1/2.0/2.7)	2.1
Sentence retrieval (CRF predictions)	13.0	40.0	4.6	2.8 (3.7/2.5/3.0)	2.6
CRF + N-gram generation	16.0	56.2	4.7	2.9 (3.9/2.6/2.7)	2.5
Translation (this work)					
CRF + Training on					
annotations (All)	11.2	38.5			
annotations (Last)	16.9	44.5			
annotations (Semantic overlap)	18.9	48.1	4.6	2.9 (3.7/2.6/3.2)	2.6
sentence level predictions	22.1	49.6	4.6	3.1 (3.9/2.9/3.3)	2.8
Upper Bounds					
CRF + Training & test on					
annotations (Last)	27.7	58.2			
annotations (Semantic overlap)	34.2	66.9	4.8	4.5 (4.5/4.7/4.0)	4.1
Human descriptions	36.0[16]	66.9[16]	4.6	4.6 (4.6/4.7/3.7)	4.3

Table 9.2: Evaluating generated descriptions on TACoS video-description corpus. Human judgments from 1-5, where 5 is best. For correctness judgments we additionally report correctness of activity, objects, and location.

description. Modeling the language statistics with a n-gram model to fill function words between predicted keywords of the SR leads to a further improvement to 16% with $n = 3$ and a search span of up to 10 words. Other n-gram models with smaller search span or different n perform worse.

Next we compare the baselines to our translation system. We first notice that most variants improve over the various baseline approaches, up to 22.1% BLEU@4. This is a significant improvement over the best baseline achieving 16.0% which uses a 3-gram language model. From this we can conclude two things. First, with respect to the SR, it seems that the CRF provides a strong intermediate representation, compared to representing the video with only raw or attribute features. Second, using our translation approach clearly improves over sentence retrieval (+9.1%) or a pure n-gram model (+6.1%). We note that the n-gram model could not be applied directly to the SR, but we had to manually select a subset of the SR and preprossess the data (see Section 9.4.2) which can be learned from data using SMT.

Comparing our different variants it is interesting to see that it is important how to match a SR with descriptions during training SMT model. When a sentence is aligned to multiple SRs, just matching all SRs to it leads to a noisy model (11.2%). It is better to use the last SR (16.9%), or the largest semantic overlap between a SR and training sentence (18.9%). Best is training on the predictions rather than ground truth SRs (22.1%) which is impressive given that it is learned on noisy predictions. In contrast to the SRs based on annotations, the predictions are on sentence intervals. This indicates that a SR on the same level of the sentence granularity is most powerful.

To answer the question what is the limit of our SR, we test on the ground truth SR, i.e. we model perfect visual recognition. This results in 27.7% / 34.2% for the last/overlap variant. This is a significant improvement and can be explained by the noisy visual predictions (see Table 9.1). As an upper bound we report the BLEU score for the human descriptions which is 36.0%[16].

While BLEU is a good indicator for performance, it cannot level with human judgments summarized in the last three columns of Table 9.2. Starting with the last column (relevance, 6th column) the two main trends suggested by the BLEU scores are confirmed: our proposed approach using *training on sentence level predictions* outperforms all baselines; and using our SR based on annotations is encouragingly close to human performance (4.1 vs. 4.3, on a scale from 1 to 5, where 5 is best). The human judgments about correctness (5th column) show scores for overall correctness (first number) followed by the scores for activities, objects (including tools and ingredients), and location (covering source and target location, see Table 9.1). Again the two main trends are confirmed. All approaches based on CRF perform similar (2.8-2.9), only our *training on sentence level predictions* performs higher with a average score of 3.1 as it can recover from errors by learning typical errors by the CRF during training (see also examples in Table 9.4). It is interesting to look at the 4th column which judges the grammatical correctness of the produced sentences disregarding the visual input. Training and testing on annotations (score 4.8) outperforms the score for human descriptions (4.6), indicating that our system learned a better language model than most human descriptions have. Our translation system achieves the same score as human descriptions. The n-gram generation receives a slightly better score of 4.7 which is however due to the shorter sentences produced by this model, leading to less grammatical errors.

9.6 EVALUATION: TRANSLATING IMAGES TO TEXT

We perform a second evaluation to compare with related work and show that our approach for video description can also be applied for image description. For our evaluation we choose the Pascal sentence dataset (Farhadi *et al.*, 2010b) which consist of 1,000 images, each paired with 5 different descriptions of one sentence. Rather than building our own SR we use the predictions provided by Farhadi *et al.* (2010b)

The SR consists of object-activity-scene triples which we annotate for the training set as they are not provided. We learn our translation approach on the training set of triples and image descriptions. We evaluate on a subset of 323 images where there are predicted descriptions available for both related approaches (Farhadi *et al.*, 2010b; Kulkarni *et al.*, 2011). We use the first predicted triple (with highest score) from Farhadi *et al.* (2010b). Mitchell *et al.* (2012) also predict sentences for this dataset but

[16] Computed only on a 272 sentence subset where the corpus contains more than a single reference sentence for the same video. This reduces the number of references by one which leads to a lower BLEU score.

Approach	BLEU @4	 @1
Related Work		
Template-based generation (Kulkarni *et al.*, 2011)	0.0	14.9
MRF + sentence retrieval (Farhadi *et al.*, 2010b)	1.1	25.6
Translation (this work)		
MRF + translation	4.6	34.6
MRF + adjective extension + translation	5.2	32.7
Upper Bound		
Human descriptions	15.2	56.7

Table 9.3: Evaluating generated descriptions on the Pascal Sentence dataset.

only example sentences were available to us.

9.6.1 Results: Translating Images to Text

We start by comparing our computed results to numbers reported by related work. Kulkarni *et al.* (2011) reports 15% BLEU@1 for their template-based generation and 50% for human descriptions. On our test subset we receive 14.9% and 56.7%, respectively, indicating that the results on the different subsets are comparable. Next we compare the two baselines with our approach shown in Table 9.3. For BLEU@4 the template approach Kulkarni *et al.* (2011) achieves 0.0 as the 4-gram precision is 0 (n-gram precision for 2- and 3-gram are very low (0.2%, 1.4%). This is not surprising as the templates produce very different text compared to descriptions by humans.

The sentences retrieved by Farhadi *et al.* (2010b) achieve a higher BLEU@4 of 1.1% and BLEU@1 of 25.6%. As these are sentences produced by humans this improvement is not surprising, but indicates that errors in the prediction cannot be recovered. Using the predicted triples from Farhadi *et al.* (2010b) together with our translation approach significantly improves performance to 4.6% BLEU@4 and 34.6% @1. Still, we found the SR not to be rich enough to produce good predictions. Adding adjectives and counts from the SR predicted by Kulkarni *et al.* (2011) could slightly increase to 5.2% BLEU@4 but decreasing to 32.7% @1. The BLEU@4 of only 15.2% for humans indicates the difficulty and diversity of the dataset. Never-the-less we outperform the best reported BLEU-score result on this dataset of 30% @1 by 5% (note the not identical test set) for language model based generation or meaning representation Kulkarni *et al.* (2011). In this case Kulkarni *et al.* (2011) allow synonyms which our translation system determines automatically from the training data.

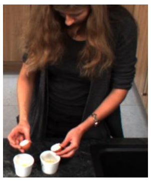
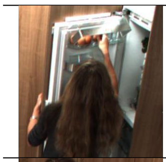
	**(1)** SR predicted by CRF	⟨ open egg, hand, egg, bowl, null ⟩
	Sentence retrieval (CRF predictions)	the person slices the avocado
	CRF + N-gram generation	the person opens up egg over
	CRF+Train on annotations (Overlap)	the person cracks the eggs into the bowl
	CRF+Train on sentence level predictions	the person cracks the eggs
	Human description	the person dumps any remaining whites of the eggs from the shells into the cup with the egg whites
	(2) SR predicted by CRF	⟨ take out, hand, plastic-bag, fridge, cutting-board⟩
	Sentence retrieval (CRF predictions)	the person took out cucumber
	CRF + N-gram generation	the person takes out a bag of chilies
	CRF+Train on annotations (Overlap)	the person gets out a package of limes from the fridge and places it on the cutting board
	CRF+Train on sentence level predictions	the person gets out a cutting board from the loaf of bread from the fridge
	Human description	the person gets the lime, a knife and a cutting board
	**(3)** SR predicted by CRF	⟨ put in, hand, wrapping-paper, null, fridge⟩
	Sentence retrieval (CRF predictions)	person then places cucumber on plate
	CRF + N-gram generation	the person puts the bread with existing plastic paper
	CRF+Train on annotations (Overlap)	the person rinses and puts away the butter back in the fridge
	CRF+Train on sentence level predictions	the person takes out a carrot from the fridge
	Human description	the person procures an egg from the fridge
	(4) SR predicted by CRF	⟨ remove from package, kiwi, hand, plastic-bag, null⟩
	Sentence retrieval (CRF predictions)	the person selects five broad beans from the package
	CRF + N-gram generation	the person removes a kiwi
	CRF+Train on annotations (Overlap)	the person takes the package of beans out of the kiwi
	CRF+Train on sentence level predictions	the person goes to the refrigerator and takes out the half kiwi
	Human description	using her hands, the person splits the orange in hald over the saucer

Table 9.4: Example output of our translation system (blue) compared to baseline approaches and human descriptions, errors in red. (1, 2) our system provides the best output; (2, 3) our system partially recovers from a wrong SR; (4) failure case.

9.7 CONCLUSION

Automatically describing videos with natural language is both a compelling as well as a challenging task. This work proposes to learn the conversion from visual content to natural descriptions from a parallel corpus of videos and textual descriptions rather than using rules and templates to generate language. Our model is a two-step approach, first learning an intermediate representation of semantic labels from the video, and then translating it to natural language adopting techniques from statistical machine translation. This allows training which part of the visual content to verbalize and in which order. In order to form a natural description of the content as humans would give it our model learns which words should be added although they are not directly present in the visual content.

In an extensive experimental evaluation we show improvements of our approach compared to retrieval and n-gram based sentence generation used in prior work. The improvements are consistent across automatic evaluation with BLEU scores and human judgments of correctness and relevance. The application of our approach to sentence descriptions shows clear improvements over Kulkarni *et al.* (2011) and Farhadi *et al.* (2010b) using BLEU sore evaluation, indicating that we produce descriptions more similar to human descriptions.

To handle the different levels of granularity in the SR compared to the description we compare different variants of our model, showing that an estimation of the largest semantic overlap between the SR and the description during training performs best.

While we show the benefits of phrasing video description as a translation problem, there are many possibilities to improve our work. Further directions include modeling temporal dependencies in both the SR and the language generation, as well as modeling the uncertainty of the visual input explicitly in the generation process, which has similarities to translating from uncertain speech input. This work could be combined with approaches which automatically extract a semantic representation from a text description, which has recently been proposed by Ramanathan *et al.* (2013) for activities.

CONCLUSIONS AND FUTURE PERSPECTIVES 10

Contents

Significant progress has been achieved in visual recognition and computational linguistics in recent years. While computer vision has made significant steps to reliable recognition for up to 1,000 object categories (Krizhevsky *et al.*, 2012) and to widely-applicable activity recognition (Wang *et al.*, 2013), computational linguistics have developed robust approaches to estimate semantic similarity (Szarvas *et al.*, 2011) and automatic translation systems between natural languages (Koehn, 2010) which are deployed in widely used web applications such as translate.google.com. However, despite the success within the domains, comparatively little work has been done to combine both modalities. This is where this thesis sets in to explore how the two modalities could benefit from each other and how to convert from one to the other modality. More specifically we focused on three directions, (1) *visual knowledge transfer using linguistic semantic relatedness*, (2) *script data for activity recognition*, and (3) *natural language descriptions of visual content*. After a summary of the thesis with respect to the three directions in the following, we discuss our contributions and future perspectives.

First, we examined *visual knowledge transfer using linguistic semantic relatedness*. More specifically we looked at the task of recognizing unseen visual object classes by transferring knowledge from known to unseen classes. For replacing manual supervision, we combined different language resources to achieve robust semantic relatedness estimates which provide the association between known and unseen classes. To understand the scalability of current semantic knowledge transfer approaches, we conducted a large scale study where we compared hierarchical, attribute-based, and direct similarity-based knowledge sharing and knowledge transfer. The knowledge sharing experiments consisted of 1,000 classes of the ImageNet (Deng *et al.*, 2009) challenge. In comparison to standard one-vs-all classification we found that the examined knowledge sharing approaches could only minimally improve. In the knowledge transfer experiments we were able to scale to 200 unseen classes and found that direct similarity approaches can achieve performance close to approaches using a manually created hierarchy. Additionally, we proposed a novel approach, Propagated Semantic Transfer, which exploits unlabeled data of the unseen/novel classes and allows benefiting from few labeled instances if available. This is realized by combining semantic knowledge transfer with label propagation (Zhou *et al.*, 2004).

The second direction of this thesis is concerned with using *script data for activity recognition*. Confronted with the unavailability of a suitable dataset, we recorded, labeled, and released two datasets of fine-grained cooking activities. The first one, the *MPII Cooking Activities Dataset*, focuses on a diverse set of complex dishes. The dataset provides a benchmark for fine-grained activity classification and detection. We evaluated our novel human pose-based features as well as holistic dense-trajectory (Wang *et al.*, 2011) features on it. The *MPII Cooking Composite Activities dataset* is the second one and focuses on recognizing dishes (composite activities) rather than individual activities. It consists of 256 videos showing different composite activities, such as *preparing carrots*, or *preparing scrambled egg*. To handle the inherent variability and the lack of training data of a specific composite activity, we propose to leverage textual script data which is easy to obtain. For this we exploit the decomposability of composite activities in smaller components termed attributes. The attributes represent objects and fine grained actions such as *peel* or *fry* which are shared across different composites. This enabled zero-shot transfer and improved recognition of composite activities when little training data is available. Furthermore, we found that context and co-occurrence are beneficial for attribute recognition.

The third part of the thesis looked at *natural language descriptions of visual content*. In contrast to both previous parts which focused on exploiting statistics in separately collected visual and linguistic datasets, in this part we focused on tightly coupled visual and linguistic modality, i.e. aligned data. We collected the *Saarbrücken Corpus of Textually Annotated Cooking Scenes* (TACoS corpus) which contains aligned sentence descriptions with the videos from the *MPII Cooking Composite Activities dataset*. This data allowed us to examine how grounding action descriptions in video helps to understand their semantic similarity. More specifically we compared a pure text based model with a visual model and showed that the visual information can significantly improve performance when combined with the text based model, especially for understanding similarity of activities. Finally, based on this parallel corpus, we proposed a two-step approach to learn how to automatically describe video snippets with natural language sentences. In the first step we predict an intermediate semantic representation from the visual input using SVM attribute classifiers paired with a conditional random field (CRF) to model co-occurrence between activities and objects. In the second step we adapt statistical machine translation to learn how to convert from the intermediate semantic representation to natural language sentences. We show the success of our approach not only for video description, but also for an image description task.

In summary we have shown that combining visual recognition with techniques from computational linguistics allows for unsupervised knowledge transfer and improved visual recognition of objects and composite activities. At the same time visual recognition can also be beneficial for grounding linguistic expressions. Finally we proposed an approach for the intruding but challenging task of automatically describing visual content, learning all steps from our parallel corpus.

The field of research looking at the interactions between visual recognition and computational linguistics was still at its infancy at the onset of the thesis. This required proposing several new datasets to study the field. We took the effort to release these datasets publicly combined with software and intermediate results to allow researches working at different challenges of the field. During the course of the thesis the field has received increasing interest by many other researches, leading to proposals of many novel approaches for knowledge transfer for object and activity recognition (e.g. Frome *et al.*, 2013; Fu *et al.*, 2013), grounding (e.g. Yu and Siskind, 2013), and image and video descriptions (e.g. Kuznetsova *et al.*, 2012; Guadarrama *et al.*, 2013a).

10.1 DISCUSSION OF CONTRIBUTIONS

The overall goal of this thesis was to exploit the mutual benefits when combining visual and linguistic modalities. Towards this goal we investigated how linguistic knowledge can be used for visual object and activity recognition and we explored natural language descriptions of visual content. In the following we will discuss the contributions and steps we made towards these goals and tasks with respect to the individual chapters.

First, we improved the robustness of semantic relatedness to enable unsupervised knowledge transfer in Chapter 3. We showed improvements over our earlier work (Rohrbach *et al.*, 2010) by adding four ingredients: we used novel semantic relatedness measures which we found to be more appropriate for attribute-based knowledge transfer; we combined individual semantic relatedness measures to exploit the different weakness and strength of different language sources; we showed how to expand the attribute inventory and generate more robust attributes; and finally, in combination with classifier level fusion we achieved performance on the level human supervision (Lampert *et al.*, 2009). One important aspect was to show that the improvements were not only attained for the case were we only distinguish unseen classes, but also when moving to the more realistic setting when the known classes are present at test time. Also other researches have recently argued that this is an important aspect (Socher *et al.*, 2013; Frome *et al.*, 2013).

Second, in Chapter 4 we scale sharing and transfer of semantic knowledge to a large scale setting of the 1,000 classes of the ImageNet 2010 challenge. To enable such large scale learning we develop a Mean Stochastic Gradient Decent (SGD) approach, which averages the models of different SGD epochs and converges significantly faster than SGD. We successfully used it as a learning technique in all the following work in the thesis and provide a parallel implementation of it publicly. Although current approaches use deep learning and convolutional networks for learning known classes which outperform our results in the supervised setting, when compared in the same knowledge transfer setting they cannot improve over our work (Frome *et al.*, 2013). We thus expect that our approach has still significantly room of improvement when combined with recent object classification models.

Third, in Chapter 5 we recorded, labeled, and publicly released a fine-grained activity recognition set, with the challenge to distinguished 65 cooking activities in both, a classification and detection scenario. While it is limited with respect to domain and to a single kitchen with fixed camera, it focuses on different challenges which we believe are typical for assisted daily living, human-robot interaction, or industrial scenarios: It requires to distinguish between activities with low inter-class but high intra-class variability, frequently determined only by subtle changes of movement such as *open* vs. *close lid* and *cut apart* vs. *cut dices*. The activities are sometimes occluded and require to generalize over different object categories, e.g. *peel* cucumber and *peel* pineapple. We benchmarked two approaches on it: the state-of-the-art holistic activity descriptor based on dense trajectories from Wang *et al.* (2011) and pose-based features. For the latter we proposed a novel approach based on body pose tracks which are described using features motivated from work in the sensor-based activity recognition community (Zinnen *et al.*, 2009). While we found the holistic approach outperforming the pose based approach, we showed slight improvements when combing both. Despite showing inferior performance we believe that pose-based approaches can be beneficial but currently lack with respect to unreliable pose estimation and feature representation. In (Amin *et al.*, 2013) we extend the 2D pictorial structures model (Mykhaylo *et al.*, 2011) with color features, more effective spatial terms, generalize it to a mixture model, and propose a novel approach for mixture component selection. We further extend it to a multi-view model that jointly reasons over humans seen from multiple viewpoints. It does not only recover 3D pose but also provides improved 2D pose estimations.

Fourth, in Chapter 6 we show how to use text-based script data for handling the large variability of composite activity recognition by selecting relevant attributes. Our flexible attribute representation enables transfer to unseen composite cooking activities and allows significantly improving performance in a supervised setting. The latter point is interesting as semantic attributes can typically not improve over discriminative approaches. We attribute this to the challenging task of learning composite activities with limited training data where knowledge sharing on an intermediate level of attributes is vital for successful recognition. We furthermore found that for the fully supervised case as well as for the zero-shot case, selecting the relevant attributes according to the discriminative tf*idf scores improves over simple co-occurrence counts. Recently, Elhoseiny *et al.* (2013) used textual descriptions to recognize unseen image categories. While our approach could also be applied to their setting, their idea of using domain adaptation to map from textual to visual domain is an interesting option for our work.

Fifth, Chapter 7 connects the previous chapters for object and activity recognition by proposing Propagated Semantic Transfer which extends zero-shot recognition to few-shot recognition and exploits the similarity between unlabeled data. We realized the approach by combining our knowledge transfer approach with label propagation (Zhou *et al.*, 2004). An important observation we made is that using attribute or object classifier scores rather than raw image or video features to compute similarity between classes significantly improves the nearest neighbor quality, one of the most important aspects for label propagation (Ebert *et al.*, 2010). Given that recognizing novel classes with little or no training data is very hard task, our approach is flexible and exploits different sources of available information, namely knowledge from known classes, external knowledge provided by humans or linguistic information, few labeled examples of novel classes, and unlabeled data. We note that our approach is also flexible to the amount of information available for novel classes, namely if some classes have no labeled examples and some have a few.

Sixth, we presented the TACoS corpus, which provides multi-sentence coherent textual descriptions for our MPII Cooking dataset. The descriptions are aligned on sentence level with the video, providing a unique parallel corpus of video, semantic representation, and natural language descriptions. We hope that the publicly available resource will foster research for combining visual and linguistic understanding, such as visual grounding and describing videos with natural language, which we explored in this thesis, but also for other tasks such as paraphrasing as the corpus contains multiple descriptions for the same video. In Chapter 8 we focused on the task of grounding the meaning of action verbs and phrases. For this we collected a gold standard of sentence similarity (ASim dataset) and evaluated several text- and video-based semantic similarity models on the dataset, both individually and in different combinations. Interestingly the differences with respect to handled objects are better captured by the linguistic models, while difference in activity are better captured by the visual features.

Finally, in Chapter 9 we learn a model to automatically describe video from the aligned TACoS corpus. We showed that our approach improves over several baseline approaches motivated by related work, including retrieval based and n-gram based generation. While our approach requires an annotated intermediate semantic representation our results showed that the representation can be learned independently from the translation model, i.e. we can learn the translation model on the predicted semantic representation. This additional opens the possibility for the language model to recover from systematic errors of the visual recognition. We also applied the approach to image description where it improved performance over Farhadi *et al.* (2010b) and Kulkarni *et al.* (2011).

10.2 FUTURE PERSPECTIVES

In the following we first discuss items of future work with respect to the different directions of the thesis. We focus on the most recent work in the thesis, *generating natural language description for visual content*. In the last section we give a broader outlook for the field.

10.2.1 Visual knowledge transfer using linguistic semantic relatedness

While attributes are beneficial for knowledge transfer and sharing, current visual and linguistic approaches are not very powerful recognizing attributes and retrieving attribute relationships, compared to standard categories. We thus layout several directions for future work.

Mining and localizing visual attributes (ongoing work). In this thesis we used supervised training to learn attributes on image-level. To scale attribute discovery and increase their specificity at the same time, an option is to localize them using semi-supervised data. We made a first step towards this in the Master Thesis of Gholamreza Bahmanyar (2011) which I co-advised. On the Attribute Discovery Dataset (Berg *et al.*, 2010b) we used a topic model to model the relationship between textual and visual words. While different attributes were frequently grouped together in a topic, the visual words were able to localize the attributes without being fixed to a predefined grid as in (Berg *et al.*, 2010b).

Category or domain aware attributes. A main motivation of using semantic attributes in this thesis was to share them across classes as originally proposed by Lampert *et al.* (2009). However, while this works well for visually very related classes such within mammals, it is questionable how well visual attribute representations are sharable across diverse domains, e.g. how well "neck" of a bottle can transfer to a "neck" of a horse, or "leg" of an animal versus a "leg" of a table. While they are in both cases visually not very similar, in case of "neck" they have similar notion of being thinner and on top of the body or, in the case of leg, being a support for body; thus a partial transfer might be possible (Stark *et al.*, 2009). It thus would be interesting to look into modals for attributes which are able to have a generic part and more specialized versions targeted to certain domains and categories.

Attribute grounding in visual data. Semantic similarity between object categories is a well-studied problem in computational linguistics. Some approaches have been able to ground these in visual information (e.g. Feng and Lapata, 2010; Leong and Mihalcea, 2011) and in this thesis we grounded activity descriptions. As we have shown in Chapters 3 and 4, estimating object class-attribute relations is harder which has also been reported by Baroni and Lenci (2008). It might thus be beneficial to also ground these relations in visual information or use joint models to understand automatically the difference and similarities between e.g. a "neck" of a bottle and the "neck" of a horse.

10.2.2 Script data for activity recognition

For relating textual script data to visual activities we found that the quality of the visual recognition for objects and activities is still a major limitation, we thus propose several ideas for future work. Another future direction is to relate scripts and videos in a more structured way which could be beneficial for both modalities.

Visual recognition of manipulated objects (ongoing work). Recognizing object is a bottleneck of our visual recognition approach as found in the last two chapters. While we are typically interested in the object the person is interacting with, the objects are typically partially occluded during the interaction. It might thus be easier to recognize the objects before or after they are manipulated and identify when a manipulation started and ended by tracking. An alternative but also complementary approach is to recognize the person's hands and extract strong visual features around it. This allows capturing the manipulated objects during the manipulation, put also during pick up and set down. In addition to capturing the object it also implicitly captures the hand pose which can be indicative for the manipulated object as well as the performed activity. When integrated probabilistically in a pose model we achieve first promising result.

Deep learning for activity recognition. While deep network architectures have been used for activity recognition (Taylor *et al.*, 2010; Le *et al.*, 2011), the impressive performance gains realized for visual object recognition in images (Krizhevsky *et al.*, 2012) have not yet been realized. This is most likely due to the added complexity for feature learning given the additional temporal domain as well as missing corresponding large scale datasets. Thus an obvious ingredient to exploit the power of deep convolutional networks will be to collect and use significantly larger annotated datasets than currently used. However, a step to simplify the complexity of the task could be to combine deep learning with tracked feature points, in the simplest form by replacing the manually designed HOG, HOF, and MBH features used in dense trajectories (Wang *et al.*, 2013) with a learned representation. This would exploit point tracking which would be hard to learn from data.

Phrases of objects and activities. Object and activity recognition is typically approached individually. However, many activities are specific to a certain object. As an example consider the following activity-object pairs: OPEN TIN, OPEN BOTTLE, and CLOSE BOTTLE. Ideally we would like to learn these specific concepts, similar in spirit to the idea of visual phrases (Sadeghi and Farhadi, 2011), but at the same time we would still like to generalize to the concepts of OPEN and BOTTLE and exploit the connection between specific and general concepts during learning and at test time. It would also be interesting to see how these ideas would affect our grounding approach which tried to estimate these similarities but used individual classifiers for the components of phrases. A first idea to approach this would be to learn additional edge features in our CRF model, which represent the specificity of object and activities.

Relate the structure of scripts with structure of videos. We use script data in form of cooking instructions to improve composite activity recognition, however we do not exploit the structure encoded in script knowledge (Regneri *et al.*, 2010). While we found that structure of script data cannot directly be mapped to the temporal structure of a video, the script knowledge is still a good indicator of temporal order and necessary steps. Thus relating the structure of scripts with the temporal structure of videos can improve visual recognition, but would also allow to ground and verify automatically collected script knowledge in visual data.

10.2.3 Natural language descriptions of visual content

As part of future work we plan to extend our work to a more unconstrained setting as e.g. handled by Guadarrama *et al.* (2013a) by working on longer videos with multi-sentence descriptions which are *not constraint* by manually defined templates, but rather compose language in a natural way. In the following we will discuss several aspects to achieve this goal.

Video segmentation (ongoing work). Most activity recognition approaches concentrate on activity classification, disregarding the temporal segmentation of the video, which is important when describing a video with multiple sentences. The problem has been approached with sliding window and non-maximum suppression (Rohrbach *et al.*, 2012a) or over-segmentation and then classifying those segments (Das *et al.*, 2013b). Both approaches seem unsatisfactory as they either lead to a significant drop in performance or remove the possibility to recognize larger segments for long-term activities. We are currently working on a first approach that exploits fine-grained activity and object classifiers learned on video snippets to form a similarity measure for agglomerative clustering.

Multi-sentence generation (ongoing work). While most current systems focus on short videos described with a single or a few sentences, we want to produce multi-sentence descriptions for more complex, long videos. While this has been explored for videos previously (Das *et al.*, 2013b; Khan *et al.*, 2011; Tan *et al.*, 2011), it remains a challenge to generate a linguistically coherent description. Currently we try to enforce consistency of multiple sentences by ensuring they are about the same topic by adding an additional topic node in our CRF. Furthermore we want to look at the following three ingredients. First, we hope to exploit the visual knowledge to improve anaphora resolution, which is a challenging problem when solely text information is used. Second, we also want to combine multiple sentences to more complex expressions by exploiting person and object identity and similarity information provided by the semantic representation. Finally, we want to adopt the recent advancements of statistical machine translation to produce multi-sentence translations (Hardmeier *et al.*, 2013) using constraints from our visual representation, ideally incorporating the above two aspects.

Correct usage of anaphors. For most languages, including English, one requires to know the identity, the gender, and the number, to correctly refer to people and objects as well as to refer back to them at a later point in the text using anaphors (*he, she, it, they, them*). Especially tracking the identity of (potentially small) objects and counting (at least singular and plural) are challenging in a general domain video, given that already people re-identification is a challenging problem (Zhao *et al.*, 2013).

Learning a semantic representation. An intermediate semantic representation is an important aspect for generating high quality descriptions of video as it allows abstraction, sharing, transfer, and including world knowledge. Furthermore, the semantic representation is important to reason across different sentence intervals and to ensure consistency within and across sentences. So far we used a manually defined semantic representation (Rohrbach *et al.*, 2013b), while Krishnamoorthy *et al.* (2013) learned the subject-verb-object states from textual data. To be able to capture the diversity in an unconstrained domain and not be restricted by a too simple semantic representation for natural verbalization it would be good to learn the intermediate semantic concepts from data. One way to attack this would be an integrated approach which mines visual-textual concepts which are a combination of visual features and linguistic phrases. Apart from immediately being able to guide the video description process, this will hopefully allow for a better disambiguation of similar visual concepts and ground synonymic linguistic expressions.

This could also be phrased as weakly supervised learning, treating the semantic representation as latent and build a latent visual-text mapping (Socher and Fei-Fei, 2010). This would allow to leverage corpora without semantic representation such as movies with script data (Laptev and Pérez, 2007) or web videos with captions (Ramanathan *et al.*, 2013). Furthermore, one could use unaligned visual datasets and text corpora by learning how to map them to the latent semantic space as has been proposed for building bilingual text-lexicons from mono-lingual text corpora (Haghighi *et al.*, 2008).

Coupling of visual recognition and language generation. In this thesis we have approached learning the semantic representation and text generation separately. Currently we are working on exploiting the probabilistic output of the visual recognition during language generation rather than relying on a single maximum prediction. However, it could be better to use structured learning when optimizing the semantic representation. This allows the integration of a loss which measures the quality of the generated language rather than optimizing for the semantic representation. This should be done for single sentences, but ideally also for multi-sentence descriptions, as discussed above.

A further idea for improving the coupling between visual learning and learning how to generate a description is to connect deep learning approaches from vision (Taylor *et al.*, 2010) with concepts from deep learning for machine translation (Le *et al.*, 2012; Schwenk, 2012).

Interacting and instructing robots. While there has been several works on instructing robots with natural language (e.g. Wong and Mooney, 2006), for many utterances it is important that the robot is able to ground and relate this information to its visual perception to understand what is meant. While initial work has been done in this area (e.g. Tellex *et al.*, 2011; Guadarrama *et al.*, 2013b) the handled complexity of language and diversity of visual examples is still limited.

10.2.4 A broader view on the topic

While the previous sections discussed concrete items to approach limitations and ideas for future steps with respect to the contributions of this thesis, in this section we step back and outline broader challenges for combining visual recognition and computational linguistics.

More labelled data. As mentioned already above for automatic video description, the next big advancement for relating visual and linguistic information is likely by using powerful machine learning techniques on large supervised data sets, similar to the success of using large parallel corpora in statistical machine translation and visual recognition using ImageNet (Deng *et al.*, 2009) with deep convolutional networks (Krizhevsky *et al.*, 2012). Thus collecting large supervised corpora of parallel visual and linguistic information and adapting existing machine learning techniques is one of the most promising routes for improving over what is currently possible.

Tighter coupling of visual and linguistic elements. Most works in this field including this thesis link visual and linguistic domain on a rather loose level by exploiting strong co-occurrence statistics in both domains independently to aid solving visual recognition or grounding tasks. To achieve a better understanding how elements of the different modalities relate to each other one could aim to learn semantic units which are associated with both, visual and linguistic elements, ideally sharing visual and linguistic concepts between different semantic units to handle different levels of abstraction and handle visual uncertainty and polysemous expressions on the linguistic side.

Larger semantic units. When relating or converting between visual and linguistic information, approaches typically focus on single elements such as activities↔ verbs or objects↔nouns rather than relating more complex scenarios such as ⟨UNSUCCESSFULLY, OPEN, TIN, WITH KNIFE⟩, similar to ideas discussed above for visual recognition using visual phrases of objects (Sadeghi and Farhadi, 2011). In this thesis we have done initial steps in the last two chapters, but the focus was more on combining statistics of individual elements rather than modeling them jointly. While modeling elements individually allows exploiting much better occurrence statistics (during training), we believe that for a more complete understanding it is important to model elements jointly. However, this is only feasible if approaches allow transferring and share knowledge across complex scenarios and model the change in meaning or at least the effect of the relation to the other modality if elements are combined to more complex visual and linguistic entities.

Better understanding. While basic sciences try to understand the universe and its principles, applied science is more driven by applications. Both these principles are important to advance the field and are likely succeed together. On the one hand we need a deeper understanding how words, sentences, and text documents are related to visual objects, scenes, and films to truly advance the field. On the other hand applications drive research as well as interest quickly – and would it not be amazing if we eventually could automatically describe to blind people what they are not able see?

LIST OF FIGURES

LIST OF TABLES

BIBLIOGRAPHY

J. Aggarwal and M. Ryoo (2011). Human activity analysis: A review, *ACM Computing Surveys*, vol. 43. Cited on page 71.

M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa (2011). Action dataset - A survey, in *Proceedings of the International conference on Instrumentation, Control, Information Technology and System Integration 2011*. Cited on page 71.

Z. Akata, F. Perronnin, Z. Harchaoui, and C. Schmid (2013). Label-Embedding for Attribute-Based Classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 15.

A. Aker and R. J. Gaizauskas (2010). Generating Image Descriptions Using Dependency Relational Patterns, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2010*. Cited on pages 32 and 138.

S. Amin, M. Andriluka, M. Rohrbach, and B. Schiele (2013). Multi-view Pictorial Structures for 3D Human Pose Estimation, in *British Machine Vision Conference (BMVC) 2013*. Cited on pages 6, 9, 24, and 154.

M. Andriluka, S. Roth, and B. Schiele (2009). Pictorial Structures Revisited: People Detection and Articulated Pose Estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on page 77.

O. Aubert and Y. Prié (2007). Advene: an open-source framework for integrating and visualising audiovisual metadata, in *ACM Multimedia 2007*. Cited on page 76.

Y. Aytar, M. Shah, and J. Luo (2008). Utilizing semantic word similarity measures for video retrieval, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on page 18.

M. Baccouche, F. Mamalet, C. Wolf, C. Garcia, and A. Baskurt (2011). Sequential deep learning for human action recognition, in *Human Behavior Understanding 2011*, pp. 29–39, Springer. Cited on page 21.

G. Bahmanyar (2011). *Semi-Supervised Discovery of Visual Attributes*, Masters thesis, Universität des Saarlandes, Saarbrücken. Cited on page 156.

A. Barbu, A. Bridge, Z. Burchill, D. Coroian, S. Dickinson, S. Fidler, A. Michaux, S. Mussman, S. Narayanaswamy, D. Salvi, L. Schmidt, J. Shangguan, J. M. Siskind, J. Waggoner, S. Wang, J. Wei, Y. Yin, and Z. Zhang (2012). Video in sentences out, in *UAI 2012*. Cited on pages 29, 33, 136, and 139.

K. Barnard, P. Duygulu, D. A. Forsyth, N. de Freitas, D. M. Blei, and M. I. Jordan (2003). Matching Words and Pictures, *Journal of Machine Learning Research (JMLR)*. Cited on page 19.

K. Barnard and K. Yanai (2006). Mutual information of words and pictures, in *Information Theory and Applications 2006*. Cited on page 27.

M. Baroni and A. Lenci (2008). Concepts and properties in word spaces, *Italian Journal of Linguistics*, vol. 20(1), pp. 55–88. Cited on page 156.

M. Baroni and A. Lenci (2010). Distributional memory: A general framework for corpus-based semantics, *Computational Linguistics*, vol. 36(4), pp. 673–721. Cited on page 27.

A. Barr and E. Feigenbaum (1981). *The Handbook of Artificial Intelligence, Volume 1*, William Kaufman Inc., Los Altos, CA. Cited on page 94.

E. Bart and S. Ullman (2005a). Cross-Generalization: Learning Novel Classes from a Single Example by Feature Replacement, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2005*. Cited on page 37.

E. Bart and S. Ullman (2005b). Single-example learning of novel classes using representation by similarity, in *Proceedings of the British Machine Vision Conference (BMVC) 2005*. Cited on pages 16, 37, 52, 53, 54, 104, and 106.

R. Benenson, M. Mathias, T. Tuytelaars, and L. V. Gool (2013). Seeking the strongest rigid detector, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 1.

A. Berg, J. Deng, and L. Fei-Fei (2010a). *ILSVRC 2010, www.image-net.org/challenges/LSVRC/2010/*. Cited on pages 20, 52, 58, 59, 66, 111, and 115.

T. L. Berg, A. C. Berg, and J. Shih (2010b). Automatic Attribute Discovery and Characterization from Noisy Web Data, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 18 and 156.

S. Bergsma and B. Van Durme (2011). Learning bilingual lexicons using the visual similarity of labeled web images, in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2011*. Cited on page 28.

M. Berland and E. Charniak (1999). Finding parts in very large corpora, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 1999*. Cited on page 40.

V. Bettadapura, G. Schindler, T. Plötz, and I. Essa (2013). Augmenting Bag-of-Words: Data-Driven Discovery of Temporal and Structural Information for Activity Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 21.

U. Blanke and B. Schiele (2010). Remember and Transfer what you have Learned - Recognizing Composite Activities based on Activity Spotting, in *Proceedings of the 14th IEEE International Symposium on Wearable Computers (ISWC'10) 2010*. Cited on page 106.

J. Bloem, M. Regneri, and S. Thater (2012). Robust processing of noisy web-collected data, in *Proceedings of KONVENS 2012 2012*. Cited on page 95.

P. Bodesheim, A. Freytag, E. Rodner, M. Kemmler, and J. Denzler (2013). Kernel Null Space Methods for Novelty Detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 16.

E. Boiy, K. Deschacht, and M.-F. Moens (2008). Learning Visual Entities and Their Visual Attributes from Text Corpora, in *DEXA 2008*. Cited on page 27.

P. Bojanowski, F. Bach, I. Laptev, J. Ponce, C. Schmid, J. Sivic, *et al.* (2013). Finding Actors and Actions in Movies, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on page 26.

A. Bordes, L. Bottou, and P. Gallinari (2009). SGD-QN: Careful Quasi-Newton Stochastic Gradient Descent, *JMLR*. Cited on pages 59 and 60.

L. Bottou (2010), *http://leon.bottou.org/projects/sgd*. Cited on page 61.

L. D. Bourdev and J. Malik (2009). Poselets: Body part detectors trained using 3D human pose annotations, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on page 22.

W. Brendel and S. Todorovic (2011). Learning Spatiotemporal Graphs of Human Activities, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 22, 24, and 74.

E. Bruni, G. B. Tran, and M. Baroni (2011). Distributional semantics from text and images, in *Workshop on GEometrical Models of Natural Language Semantics (GEMS) 2011*. Cited on pages 27, 32, 120, 121, and 129.

B. Chakraborty, M. Holte, T. Moeslund, J. Gonzalez, and F. Roca (2011). A Selective Spatio-Temporal Interest Point Detector for Human Action Recognition in Complex Scenes, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 21, 70, 74, and 90.

O. Chapelle, B. Schölkopf, A. Zien, *et al.* (2006). *Semi-supervised learning*, vol. 2, MIT press Cambridge. Cited on page 17.

D. L. Chen and W. B. Dolan (2011). Collecting Highly Parallel Data for Paraphrase Evaluation, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2011*. Cited on pages 23, 26, 30, 32, and 121.

H.-H. Chen, M.-S. Lin, and Y.-C. Wei (2006). Novel association measures using web search with double checking, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2006*. Cited on pages 37, 40, and 41.

X. Chen, A. Shrivastava, and A. Gupta (2013). NEIL: Extracting visual knowledge from web data, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on page 19.

J. Choi, M. Rastegari, A. Farhadi, and L. S. Davis (2013). Adding Unlabeled Samples to Categories by Learned Attributes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 107.

T.-S. Chua, J. Tang, R. Hong, H. Li, Z. Luo, and Y.-T. Zheng (2009). NUS-WIDE: A Real-World Web Image Database from National University of Singapore, in *CIVR 2009*. Cited on page 58.

R. L. Cilibrasi and P. M. Vitanyi (2007). The google similarity distance, *IEEE Transactions on Knowledge and Data Engineering*, vol. 19(3), pp. 370–383. Cited on page 18.

T. Cour, C. Jordan, E. Miltsakaki, and B. Taskar (2008). Movie/Script: Alignment and Parsing of Video and Text Transcription, in *Proceedings of the European Conference on Computer Vision (ECCV) 2008*, vol. 5305 of *Lecture Notes in Computer Science*, pp. 158–171, Springer Berlin Heidelberg. Cited on pages 26 and 121.

I. Dagan, O. Glickman, and B. Magnini (2005). The PASCAL Recognising Textual Entailment Challenge, in *Proceedings of MLCW 2005 2005*. Cited on page 130.

N. Dalal, B. Triggs, and C. Schmid (2006). Human Detection Using Oriented Histograms of Flow and Appearance, in *Proceedings of the European Conference on Computer Vision (ECCV) 2006*. Cited on pages 21, 71, and 80.

P. Das, R. K. Srihari, and J. J. Corso (2013a). Translating Related Words to Videos and Back through Latent Topics, in *Proceedings of Sixth ACM International Conference on Web Search and Data Mining 2013*. Cited on page 30.

P. Das, C. Xu, R. F. Doell, and J. Corso (2013b). 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 (CVPR) 2013*. Cited on pages 26, 30, 32, 33, 136, 139, and 158.

F. de la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey (2009). Guide to the CMU Multimodal Activity Database, Technical report CMU-RI-TR-08-22, Robotics Institute. Cited on pages 72 and 74.

M.-C. de Marneffe, B. MacCartney, C. D. Manning, *et al.* (2006). Generating typed dependency parses from phrase structure parses, in *Proceedings of the International Conference on Language Resources and Evaluation (LREC) 2006*. Cited on page 23.

T. Dean, J. Yagnik, M. Ruzon, J. Segal, Mark Shlens, and S. Vijayanarasimhan (2013). Fast, Accurate Detection of 100,000 Object Classes on a Single Machine, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 1.

B. Delezoide, G. Pitel, H. Le Borgne, G. Greffenstette, P.-A. Moëllic, and C. Millet (2008). Object/Background Scene Classification in Photographs Using Linguistic Statistics from the Web, in *OntoImage 2008*. Cited on pages 18, 19, 37, and 40.

J. Deng, A. Berg, K. Li, and L. Fei-Fei (2010). What does classifying more than 10,000 image categories tell us?, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 52, 53, 55, and 61.

J. Deng, W. Dong, R. Socher, L.-J. Li, K. Li, and L. Fei-Fei (2009). ImageNet: A Large-Scale Hierarchical Image Database, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 5, 32, 52, 53, 58, 151, and 160.

J. Deng, J. Krause, A. Berg, and L. Fei-Fei (2012). Hedging Your Bets: Optimizing Accuracy-Specificity Trade-offs in Large Scale Visual Recognition, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 30.

L. R. Dice (1945). Measures of the amount of ecologic association between species, *Ecology*, vol. 26(3), pp. 297–302. Cited on page 18.

J. Dodge, A. Goyal, X. Han, A. Mensch, M. Mitchell, K. Stratos, K. Yamaguchi, Y. Choi, H. D. III, A. C. Berg, and T. L. Berg (2012). Detecting Visual Text, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics 2012*. Cited on pages 27 and 121.

K. Duan, D. Parikh, D. Crandall, and K. Grauman (2012). Discovering Localized Attributes for Fine-grained Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 18 and 104.

P. Duygulu, K. Barnard, N. de Freitas, and D. A. Forsyth (2002). Object Recognition as Machine Translation: Learning a Lexicon for a Fixed Image Vocabulary, in *Proceedings of the European Conference on Computer Vision (ECCV) 2002*. Cited on page 138.

S. Ebert (2012). *Semi-Supervised Learning for Image Classification*, Ph.D. thesis, Saarland University. Cited on page 17.

S. Ebert, D. Larlus, and B. Schiele (2010). Extracting Structures in Image Collections for Object Recognition, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 17, 20, 107, 109, 114, and 155.

M. Elhoseiny, B. Saleh, and A. Elgammal (2013). Write a Classifier: Zero-Shot Learning Using Purely Textual Descriptions, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 19, 20, 24, and 154.

K. Erk and D. McCarthy (2009). Graded Word Sense Assignment, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2009*. Cited on page 129.

K. Erk, D. McCarthy, and N. Gaylord (2009). Investigations on Word Senses and Word Usages, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2009*. Cited on pages 120 and 127.

K. Erk, D. McCarthy, and N. Gaylord (2012). Measuring Word Meaning in Context, *CL*. Cited on page 127.

M. Everingham, L. V. Gool, C. Williams, J. Winn, and A. Zisserman (2009). The PASCAL Visual Object Classes (VOC) challenge. Cited on page 25.

M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman (2011). *The PASCAL Action Classification Taster Competition*. Cited on page 71.

A. Farhadi, I. Endres, and D. Hoiem (2010a). Attribute-Centric Recognition for Cross-Category Generalization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 37, 52, 53, 74, and 90.

A. Farhadi, I. Endres, D. Hoiem, and D. Forsyth (2009). Describing objects by their attributes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on page 104.

A. Farhadi, M. Hejrati, M. Sadeghi, P. Young, C. Rashtchian, J. Hockenmaier, and D. Forsyth (2010b). Every Picture Tells a Story: Generating Sentences from Images, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 10, 25, 28, 31, 136, 138, 139, 142, 143, 144, 146, 147, 149, and 155.

R. Farrell, O. Oza, V. Morariu, T. Darrell, and L. S. Davis (2011). Birdlets: Subordinate Categorization Using Volumetric Primitives and Pose-Normalized Appearance, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on page 104.

A. Fathi, A. Farhadi, and J. M. Rehg (2011). Understanding egocentric activities, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 74 and 88.

M. Federico, N. Bertoldi, and M. Cettolo (2008). IRSTLM: an open source toolkit for handling large scale language models, in *Interspeech 2008*. Cited on page 142.

C. Fellbaum (1998). *WordNet: An Electronical Lexical Database*, The MIT Press. Cited on pages 5, 40, 52, 91, and 96.

P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan (2010). Object Detection with Discriminatively Trained Part-Based Models, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 32. Cited on pages 22, 30, and 78.

Y. Feng and M. Lapata (2010). How Many Words Is a Picture Worth? Automatic Caption Generation for News Images, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2010*. Cited on pages 32, 120, 121, 138, and 156.

R. Fergus, Y. Weiss, and A. Torralba (2009). Semi-supervised Learning in Gigantic Image Collections, in *Advances in Neural Information Processing Systems (NIPS) 2009*. Cited on page 106.

V. Ferrari, M. Marin, and A. Zisserman (2008). Progressive Search Space Reduction for Human Pose Estimation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 75 and 78.

V. Ferrari and A. Zisserman (2007). Learning Visual Attributes, in *Advances in Neural Information Processing Systems (NIPS) 2007*. Cited on pages 36, 37, and 90.

J. M. Ferryman (Ed.) (2007). *PETS*. Cited on pages 72 and 73.

M. Fink (2004). Object Classification from a Single Example Utilizing Class Relevance Pseudo-Metrics, in *Advances in Neural Information Processing Systems (NIPS) 2004*. Cited on pages 17, 36, 37, 52, 54, and 106.

A. Frome, G. Corrado, J. Shlens, S. Bengio, J. Dean, M. Ranzato, and T. Mikolov (2013). DeViSE: A Deep Visual-Semantic Embedding Model, in *Advances In Neural Information Processing Systems, nips 2013*. Cited on pages 14, 15, 16, 19, 20, 104, and 153.

Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong (2012). Attribute Learning for Understanding Unstructured Social Activity, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on pages 15 and 17.

Y. Fu, T. M. Hospedales, T. Xiang, and S. Gong (2013). Learning Multi-modal Latent Attributes, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. PP(99). Cited on pages 14, 15, 17, 20, 104, 106, 113, 114, 116, and 153.

E. Gabrilovich and S. Markovitch (2007). Computing Semantic Relatedness using Wikipedia-based Explicit Semantic Analysis, in *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI) 2007*. Cited on pages 27, 40, and 57.

D. Gehrig, H. Kuehne, A. Woerner, and T. Schultz (2009). HMM-based human motion recognition with optical flow data, in *Humanoid Robots 2009*. Cited on page 70.

R. Girshick, J. Donahue, T. Darrell, and J. Malik (2013). Rich feature hierarchies for accurate object detection and semantic segmentation, Technical report, arXiv:1311.2524. Cited on page 21.

A. M. Glenberg (2002). Grounding language in action, *Psychonomic Bulletin & Review*. Cited on pages 27 and 120.

G. Griffin, A. Holub, and P. Perona (2007). Caltech-256 Object Category Dataset, Technical report 7694, Caltech. Cited on page 58.

G. Griffin and P. Perona (2008). Learning and using taxonomies for fast visual categorization, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 52 and 53.

S. Guadarrama, N. Krishnamoorthy, G. Malkarnenkar, R. Mooney, T. Darrell, and K. Saenko (2013a). YouTube2Text: Recognizing and Describing Arbitrary Activities Using Semantic Hierarchies and Zero-Shoot Recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 14, 26, 30, 32, 33, 136, 139, 153, and 158.

S. Guadarrama, L. Riano, D. Golland, D. Goehring, Y. Jia, D. Klein, P. Abbeel, and T. Darrell (2013b). Grounding Spatial Relations for Human-Robot Interaction, in *International Conference on Intelligent Robots and Systems 2013*. Cited on pages 27 and 160.

A. Gupta and L. S. Davis (2007). Objects in Action: An Approach for Combining Action Understanding and Object Perception, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. Cited on page 90.

A. Gupta, P. Srinivasan, J. B. Shi, and L. Davis (2009). Understanding videos, constructing plots learning a visually grounded storyline model from annotated videos, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 22, 24, 30, 121, 136, and 139.

S. Gupta and R. J. Mooney (2010). Using Closed Captions as Supervision for Video Activity Recognition, in *Proceedings of the Twenty-Fourth AAAI Conference on Artificial Intelligence (AAAI-2010) 2010*. Cited on pages 26 and 121.

A. Haghighi, P. Liang, T. Berg-Kirkpatrick, and D. Klein (2008). Learning Bilingual Lexicons from Monolingual Corpora, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2008*. Cited on page 159.

P. Hanckmann, K. Schutte, and G. J. Burghouts (2012). Automated Textual Descriptions for a Wide Range of Video Events with 48 Human Actions, in *Trends and Topics in Computer Vision : ECCV 2012 Workshops 2012*. Cited on pages 29, 136, and 139.

C. Hardmeier, S. Stymne, J. Tiedemann, and J. Nivre (2013). Docent: A Document-Level Decoder for Phrase-Based Statistical Machine Translation, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2013*. Cited on page 158.

D. R. Hardoon, S. Szedmák, and J. Shawe-Taylor (2004). Canonical Correlation Analysis: An Overview with Application to Learning Methods, *Neural Computation*, vol. 16(12), pp. 2639–2664. Cited on page 24.

A. Hauptmann, M. Christel, and R. Yan (2008). Video Retrieval Based on Semantic Concepts, *Proc. IEEE*, vol. 96(4). Cited on page 91.

T. J. Hazen, F. Richardson, and A. Margolis (2007). Topic identification from audio recordings using word and phone recognition lattices, in *IEEE Workshop on Automatic Speech Recognition & Understanding (ASRU) 2007*. Cited on page 23.

S. R. Howell, D. Jankowicz, and S. Becker (2005). A model of grounded language acquisition: Sensorimotor features improve lexical and grammatical learning, *JML*. Cited on pages 27 and 120.

E. H. Huang, R. Socher, C. D. Manning, and A. Y. Ng (2012). Improving word representations via global context and multiple word prototypes, in *Proceedings of the 50th Annual Meeting of the Association for Computational Linguistics: Long Papers-Volume 1 2012*. Cited on page 15.

H. Jégou, M. Douze, C. Schmid, and P. Pérez (2010). Aggregating local descriptors into a compact image representation, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 59.

H. Jhuang, J. Gall, S. Zuffi, C. Schmid, and M. J. Black (2013). Towards understanding action recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 22, 24, and 75.

T. Joachims (1999). Transductive inference for text classification using support vector machines, in *Proceedings of the International Conference on Machine Learning (ICML) 1999*. Cited on page 17.

P. Kankuekul, A. Kawewong, S. Tangruamsub, and O. Hasegawa (2012). Online Incremental Attribute-based Zero-shot Learning, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 107.

M. Kemmler, E. Rodner, and J. Denzler (2010). One-Class Classification with Gaussian Processes, in *ACCV (2) 2010*. Cited on page 16.

C. Kemp, J. B. Tenenbaum, T. L. Griffiths, T. Yamada, and N. Ueda (2006). Learning Systems of Concepts with an Infinite Relational Model, in *AAAI 2006*. Cited on page 39.

M. U. G. Khan and Y. Gotoh (2012). Describing video contents in natural language, in *Proceedings of the Workshop on Innovative Hybrid Approaches to the Processing of Textual Data 2012*. Cited on page 29.

M. U. G. Khan, L. Zhang, and Y. Gotoh (2011). Human Focused Video Description, in *ICCV Workshops 2011*. Cited on pages 29, 30, 136, 139, and 158.

A. Kilgarriff and G. Grefenstette (2003). Introduction to the Special Issue on the Web as Corpus, *Computational Linguistics*, vol. 29, pp. 333–347. Cited on page 18.

O. Kliper-Gross, T. Hassner, and L. Wolf (2012). The Action Similarity Labeling Challenge, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34(3), pp. 615–621. Cited on pages 72 and 73.

P. Koehn (2010). *Statistical Machine Translation*, Cambridge University Press. Cited on pages 2, 136, 137, and 151.

P. Koehn, H. Hoang, A. Birch, C. Callison-Burch, M. Federico, N. Bertoldi, B. Cowan, W. Shen, C. Moran, R. Zens, C. Dyer, O. Bojar, A. Constantin, and E. Herbst (2007). Moses: Open Source Toolkit for Statistical Machine Translation, in *Annual Meeting of the Association for Computational Linguistics: Human Language Technologies (demo) 2007*. Cited on pages 138, 141, and 142.

A. Kojima, T. Tamura, and K. Fukunaga (2002). Natural Language Description of Human Activities from Video Images Based on Concept Hierarchy of Actions, *International Journal of Computer Vision (IJCV)*. Cited on pages 29, 136, and 139.

A. Kovashka and K. Grauman (2010). Learning a Hierarchy of Discriminative Space-Time Neighborhood Features for Human Action Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 88 and 90.

N. Krishnamoorthy, G. Malkarnenkar, R. J. Mooney, K. Saenko, and S. Guadarrama (2013). Generating Natural-Language Video Descriptions Using Text-Mined Knowledge, in *AAAI 2013*. Cited on pages 26, 30, and 159.

A. Krizhevsky, I. Sutskever, and G. E. Hinton (2012). ImageNet Classification with Deep Convolutional Neural Networks, in *Advances in Neural Information Processing Systems (NIPS) 2012*. Cited on pages 21, 59, 151, 157, and 160.

H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre (2011). HMDB: A Large Video Database for Human Motion Recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 22, 72, and 73.

G. Kulkarni, V. Premraj, S. Dhar, S. Li, Y. Choi, A. C. Berg, and T. L. Berg (2011). Baby talk: Understanding and generating simple image descriptions, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 28, 30, 31, 33, 136, 139, 140, 143, 144, 146, 147, 149, and 155.

N. Kumar, A. C. Berg, P. N. Belhumeur, and S. K. Nayar (2009). Attribute and Simile Classifiers for Face Verification, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on page 37.

P. Kuznetsova, V. Ordonez, A. Berg, T. Berg, and Y. Choi (2013). Generalizing Image Captions for Image-Text Parallel Corpus, in *The 51st Annual Meeting of the Association for Computational Linguistics - Short Papers 2013*. Cited on pages 25, 32, and 33.

P. Kuznetsova, V. Ordonez, A. C. Berg, T. L. Berg, and Y. Choi (2012). Collective Generation of Natural Image Descriptions, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2012*. Cited on pages 31, 136, 139, 144, and 153.

C. Lampert, H. Nickisch, and S. 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 (CVPR) 2009*. Cited on pages 14, 16, 19, 20, 27, 36, 37, 38, 39, 41, 45, 47, 52, 53, 55, 56, 64, 88, 90, 93, 153, 156, and 165.

C. Lampert, H. Nickisch, and S. Harmeling (2013). Attribute-Based Classification for Zero-Shot Learning of Object Categories, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. PP(99). Cited on pages 14, 53, 104, 106, 107, 108, 110, 111, 113, and 114.

I. Laptev (2005). On Space-Time Interest Points, in *International Journal of Computer Vision (IJCV) 2005*. Cited on pages 21, 70, 74, and 90.

I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld (2008). Learning realistic human actions from movies, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 21, 26, 30, 71, 74, 75, 80, and 91.

I. Laptev and P. Pérez (2007). Retrieving actions in movies, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. Cited on pages 72, 73, and 159.

H. Larochelle, D. Erhan, and Y. Bengio (2008). Zero-data Learning of New Tasks, in *AAAI 2008*. Cited on page 14.

S. Lazebnik, C. Schmid, and J. Ponce (2006). Beyond Bags of Features: Spatial Pyramid Matching for Recognizing Natural Scene Categories, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2006*. Cited on page 59.

H. S. Le, A. Allauzen, and F. Yvon (2012). Continuous Space Translation Models with Neural Networks, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics 2012*. Cited on page 159.

Q. V. Le, W. Y. Zou, S. Y. Yeung, and A. Y. Ng (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis, in *Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on 2011*. Cited on pages 21, 24, and 157.

C. W. Leong and R. Mihalcea (2011). Going Beyond Text: A Hybrid Image-Text Approach for Measuring Word Relatedness, in *Proceedings of the International Joint Conference on Natural Language Processing (IJCNLP) 2011*. Cited on pages 27, 32, and 156.

F. F. Li, R. Fergus, and P. Perona (2006). One-Shot Learning of Object Categories, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 28(4), pp. 594–611. Cited on pages 17, 37, and 53.

L.-J. Li and F.-F. Li (2007). What, where and who? Classifying events by scene and object recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. Cited on page 20.

L. J. Li, R. Socher, and L. Fei-Fei (2009). Towards total scene understanding: Classification, annotation and segmentation in an automatic framework, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on page 19.

L.-J. Li, H. Su, Y. Lim, and L. Fei-Fei (2010a). Objects as Attributes for Scene Classification, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on page 90.

L.-J. Li, H. Su, E. P. Xing, and F.-F. Li (2010b). Object Bank: A High-Level Image Representation for Scene Classification & Semantic Feature Sparsification, in *Advances in Neural Information Processing Systems (NIPS) 2010*. Cited on page 30.

L.-J. Li, C. Wang, Y. Lim, D. Blei, and L. Fei-Fei (2010c). Building and Using a Semantivisual Image Hierarchy, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 19.

C. Liang, C. Xu, J. Cheng, W. Min, and H. Lu (2013). Script-to-Movie: A Computational Framework for Story Movie Composition, *Multimedia, IEEE Transactions on*, vol. 15(2), pp. 401–414. Cited on page 3.

D. Lin (1998). An Information-Theoretic Definition of Similarity, in *Proceedings of the International Conference on Machine Learning (ICML) 1998*. Cited on pages 18, 27, and 40.

H.-T. Lin, C.-J. Lin, and R. C. Weng (2007). A note on Platt's probabilistic outputs for support vector machines, *Machine Learning*. Cited on pages 31 and 112.

Y. Lin, F. Lv, S. Zhu, M. Yang, T. Cour, K. Yu, L. Cao, and T. Huang (2011). Large-scale image classification: fast feature extraction and SVM training, in *Proceedings of the*

IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011. Cited on pages 59 and 60.

J. Liu, B. Kuipers, and S. Savarese (2011). Recognizing Human Actions by Attributes, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 88, 90, 91, and 106.

J. G. Liu, J. B. Luo, and M. Shah (2009). Recognizing realistic actions from videos 'in the wild', in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 21, 73, 74, and 88.

A. Lopez (2008). Statistical machine translation, *ACM*. Cited on page 138.

U. Luxburg (2007). A tutorial on spectral clustering, *Stat Comput*, vol. 17(4), pp. 395–416. Cited on page 106.

I. Maglogiannis, D. Vouyioukas, and C. Aggelopoulos (2009). Face detection and recognition of natural human emotion using Markov random fields, *Personal and Ubiquitous Computing*, vol. 13(1), pp. 95–101. Cited on page 29.

M. Maier, U. V. Luxburg, and M. Hein (2008). Influence of graph construction on graph-based clustering measures, in *Advances in Neural Information Processing Systems (NIPS) 2008*. Cited on pages 107 and 109.

M. Marszalek, I. Laptev, and C. Schmid (2009). Actions in context, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 21, 72, 73, 74, and 88.

M. Marszalek and C. Schmid (2007). Semantic Hierarchies for Visual Object Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2007*. Cited on pages 37, 52, and 53.

S. Mathe, A. Fazly, S. Dickinson, and S. Stevenson (2008). Learning the abstract motion semantics of verbs from captioned videos, in *SLAM08 2008*. Cited on pages 28, 33, and 121.

C. Matuszek, D. Fox, and K. Koscher (2010). Following directions using statistical machine translation, in *Proceedings of the 5th ACM/IEEE International Conference on Human Robot Interaction, HRI 2010, Osaka, Japan, March 2-5, 2010 2010*. Cited on page 138.

K. McRae, G. S. Cree, M. S. Seidenberg, and C. McNorgan (2005). Semantic feature production norms for a large set of living and nonliving things, *Behavior Research Methods*, vol. 37(4), pp. 547–559. Cited on page 27.

T. Mensink, J. Verbeek, F. Perronnin, and G. Csurka (2012). Metric Learning for Large Scale Image Classification: Generalizing to New Classes at Near-Zero Cost, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on pages 14, 17, 19, 20, 104, 106, and 107.

R. Messing, C. Pal, and H. Kautz (2009). Activity recognition using the velocity histories of tracked keypoints, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on pages 21, 72, 74, 79, and 88.

T. Mikolov, I. Sutskever, K. Chen, G. S. Corrado, and J. Dean (2013). Distributed Representations of Words and Phrases and their Compositionality, in *Advances in Neural Information Processing Systems (NIPS) 2013*. Cited on page 15.

J. Mitchell and M. Lapata (2008). Vector-based Models of Semantic Composition, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2008*. Cited on page 129.

M. Mitchell, J. Dodge, A. Goyal, K. Yamaguchi, K. Stratos, X. Han, A. Mensch, A. C. Berg, T. L. Berg, and H. D. III (2012). Midge: Generating Image Descriptions From Computer Vision Detections, in *Proceedings of the Conference of the European Chapter of the Association for Computational Linguistics 2012*. Cited on pages 31, 33, 136, 139, and 146.

Y. Moses, S. Ullman, and S. Edelman (1996). Generalization to novel images in upright and inverted faces, *Perception*, vol. 25, pp. 443–461. Cited on page 104.

T. S. Motwani and R. J. Mooney (2012). Improving Video Activity Recognition using Object Recognition and Text Mining, in *Proceedings of the 20th European Conference on Artificial Intelligence (ECAI-2012) 2012*. Cited on pages 23 and 121.

A. Mykhaylo, R. Stefan, and S. Bernt (2011). Discriminative Appearance Models for Pictorial Structures, *International Journal of Computer Vision (IJCV)*. Cited on page 154.

H. H. Nagel (1988). From Image Sequences Towards Conceptual Descriptions, *Image and Vision Computing*, vol. 6, pp. 59–74. Cited on page 29.

P. Natarajan and R. Nevatia (2008). View and scale invariant action recognition using multiview shape-flow models, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 72 and 73.

B. Neumann and H.-J. Novak (1983). Event Models for Recognition and Natural Language Description of Events in Real-World Image Sequences, in *Proceedings of the Eighth International Joint Conference on Artificial Intelligence (IJCAI-83) 1983*. Cited on page 29.

A. Y. Ng, M. I. Jordan, and Y. Weiss (2002). On spectral clustering: Analysis and an algorithm, in *Advances in Neural Information Processing Systems (NIPS) 2002*. Cited on page 106.

B. Ni, Y. Song, and M. Zhao (2011). YouTubeEvent: On large-scale video event classification, in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops) 2011*. Cited on pages 22 and 23.

J. Niebles, C.-W. Chen, and L. Fei-Fei (2010). Modeling Temporal Structure of Decomposable Motion Segments for Activity Classification, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on pages 22, 70, 73, 88, and 90.

M.-E. Nilsback and A. Zisserman (2008). Automated flower classification over a large number of classes, in *Sixth Indian Conference on Computer Vision, Graphics & Image Processing (ICVGIP) 2008*. Cited on page 19.

F. J. Och and H. Ney (2003). A Systematic Comparison of Various Statistical Alignment Models, *CL*. Cited on page 142.

S. Oh, A. Hoogs, A. G. A. Perera, N. P. Cuntoor, C.-C. Chen, J. T. Lee, S. Mukherjee, J. K. Aggarwal, H. Lee, L. S. Davis, E. Swears, X. Wang, Q. Ji, K. K. Reddy, M. Shah, C. Vondrick, H. Pirsiavash, D. Ramanan, J. Yuen, A. Torralba, B. Song, A. Fong, A. K. R. Chowdhury, and M. Desai (2011). A large-scale benchmark dataset for event recognition in surveillance video, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 72 and 73.

V. Ordonez, G. Kulkarni, and T. L. Berg (2011). Im2Text: Describing Images Using 1 Million Captioned Photographs, in *Advances in Neural Information Processing Systems (NIPS) 2011*. Cited on pages 25, 31, and 139.

J. Orkin and D. Roy (2009). Automatic learning and generation of social behavior from collective human gameplay, in *Proceedings of AAMAS 2009 2009*. Cited on pages 28 and 122.

D. N. Osherson, J. Stern, O. Wilkie, M. Stob, and E. E. Smith (1991). Default Probability, *Cognitive Science*, vol. 15(2), pp. 251–269. Cited on pages 27 and 39.

P. Over, G. Awad, M. Michel, J. Fiscus, G. Sanders, B. Shaw, A. F. Smeaton, and G. Quéenot (2012). TRECVID 2012 – An Overview of the Goals, Tasks, Data, Evaluation Mechanisms and Metrics, in *Proceedings of TRECVID 2012 2012*. Cited on pages 26 and 73.

M. Palatucci, D. Pomerleau, G. Hinton, and T. Mitchell (2009). Zero-Shot Learning with Semantic Output Codes, in *Advances in Neural Information Processing Systems (NIPS) 2009*. Cited on pages 15, 37, 104, and 106.

S. J. Pan and Q. Yang (2010). A Survey on Transfer Learning, *TKDE*, vol. 22, pp. 1345–59. Cited on page 106.

K. Papineni, S. Roukos, T. Ward, and W. jing Zhu (2002). BLEU: a Method for Automatic Evaluation of Machine Translation, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2002*. Cited on page 144.

D. Parikh and K. Grauman (2011). Relative attributes, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on page 104.

A. Patron-Perez, M. Marszalek, A. Zisserman, and I. D. Reid (2010). High Five: Recognising human interactions in TV shows, in *Proceedings of the British Machine Vision Conference (BMVC) 2010*. Cited on pages 72 and 73.

D. Pecher and R. A. Zwaan (2005). *Grounding cognition: The role of perception and action in memory, language, and thinking*, Cambridge University Press. Cited on page 27.

F. Perronnin, J. Sánchez, and T. Mensink (2010). Improving the Fisher Kernel for Large-Scale Image Classification, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on page 59.

H. Pirsiavash and D. Ramanan (2012). Detecting activities of daily living in first-person camera views, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on page 74.

L. Pishchulin, M. Andriluka, P. Gehler, and B. Schiele (2013). Strong Appearance and Expressive Spatial Models for Human Pose Estimation, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on page 1.

B. T. Polyak and A. B. Juditsky (1992). Acceleration of Stochastic Approximation by Averaging, *SICON*. Cited on page 60.

R. Raina, A. Battle, H. Lee, B. Packer, and A. Ng (2007). Self-Taught Learning: Transfer Learning from Unlabeled Data, in *Proceedings of the International Conference on Machine Learning (ICML) 2007*. Cited on page 104.

V. Ramanathan, P. Liang, and L. Fei-Fei (2013). Video Event Understanding using Natural Language Descriptions, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 23, 149, and 159.

M. Raptis and L. Sigal (2013). Poselet Key-framing: A Model for Human Activity Recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 22 and 75.

C. Rashtchian, P. Young, M. Hodosh, and J. Hockenmaier. (2010). Collecting Image Annotations Using Amazon's Mechanical Turk, in *NAACL HLT 2010 Workshop on Creating Speech and Language Data with Amazon's Mechanical Turk 2010*. Cited on page 25.

H. Reckman, J. Orkin, and D. Roy (2011). Extracting aspects of determiner meaning from dialogue in a virtual world environment, in *Proceedings of CCS 2011 2011*. Cited on pages 28 and 122.

M. Regneri, A. Koller, and M. Pinkal (2010). Learning Script Knowledge with Web Experiments, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2010*. Cited on pages 24, 94, and 158.

M. Regneri, M. Rohrbach, D. Wetzel, S. Thater, B. Schiele, and M. Pinkal (2013). Grounding Action Descriptions in Videos, *Transactions of the Association for Computational Linguistics (TACL)*, vol. 1, pp. 25 – 36. Cited on page 12.

R. Rifkin and A. Klautau (2004). In Defense of One-Vs-All Classication, in *JMLR 2004*. Cited on page 53.

M. D. Rodriguez, J. Ahmed, and M. Shah (2008). Action MACH a spatio-temporal Maximum Average Correlation Height filter for action recognition, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2008*. Cited on pages 21, 22, 73, 74, and 88.

D. Roggen, A. Calatroni, M. Rossi, T. Holleczek, K. Förster, G. Tröster, P. Lukowicz, Bannach, G. Pirkl, A. Ferscha, J. Doppler, C. Holzmann, M. Kurz, G. Holl, R. Chavarriaga, H. Sagha, H. Bayati, M. Creatura, , and J. del R. Millán (2010). Collecting complex activity data sets in highly rich networked sensor environments, in *7th International Confernce on Networked Sensing Systems (INSS) 2010*. Cited on page 74.

M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele (2012a). A database for fine grained activity detection of cooking activities, in *2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 11 and 158.

M. Rohrbach, S. Ebert, and B. Schiele (2013a). Transfer Learning in a Transductive Setting, in *Advances in Neural Information Processing Systems (NIPS) 2013*. Cited on page 12.

M. Rohrbach, W. Qiu, I. Titov, S. Thater, M. Pinkal, and B. Schiele (2013b). Translating Video Content to Natural Language Descriptions, in *IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on pages 12 and 159.

M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele (2012b). Script data for attribute-based recognition of composite activities, in *Computer Vision - ECCV 2012 : 12th European Conference on Computer Vision 2012*. Cited on pages 11 and 116.

M. Rohrbach, M. Stark, and B. Schiele (2011). Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting, in *2011 IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 11 and 115.

M. Rohrbach, M. Stark, G. Szarvas, I. Gurevych, and B. Schiele (2010). What helps Where - and Why? Semantic Relatedness for Knowledge Transfer, in *2010 IEEE Conference on Computer Vision and Pattern Recognition (CVPR). - Pt. 2 2010*. Cited on pages 8, 10, 11, 14, 15, 18, 19, 35, 36, 37, 38, 39, 41, 42, 43, 48, 49, 52, 53, 54, 56, 57, 66, 88, 90, 106, 107, 153, and 163.

M. Rohrbach, M. Stark, G. Szarvas, and B. Schiele (2012c). Combining language sources and robust semantic relatedness for attribute-based knowledge transfer, in *Trends and Topics in Computer Vision : ECCV 2010 Workshops 2012*. Cited on page 11.

O. Russakovsky and L. Fei-Fei (2010). Attribute learning in large-scale datasets, in *Trends and Topics in Computer Vision : ECCV 2010 Workshops 2010*. Cited on pages 18 and 57.

M. S. Ryoo and J. K. Aggarwal (2009). Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on pages 22, 72, and 73.

S. Sadanand and J. J. Corso (2012). Action Bank: A High-Level Representation of Activity in Video, in *Proceedings of IEEE Conference on Computer Vision and Pattern Recognition 2012*. Cited on page 90.

M. A. Sadeghi and A. Farhadi (2011). Recognition using Visual Phrases, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 157 and 160.

K. Saenko, B. Kulis, M. Fritz, and T. Darrell (2010). Adapting Visual Category Models to New Domains, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on page 106.

G. Salton and C. Buckley (1988). Term-weighting approaches in automatic text retrieval, in *Information Processing And Management 1988*. Cited on page 96.

B. Sapp, R. Chaudhry, X. Yu, G. Singh, I. Perera, F. Ferraro, E. Tzoukermann, J. Kosecka, and J. Neumann (2011a). Recognizing Manipulation Actions in Arts and Crafts Shows using Domain Specific Visual and Textual Cues, in *The 3rd International Workshop on Video Event Categorization, Tagging and Retrieval for Real-World Applications (VECTaR, ICCV Workshop) 2011*. Cited on page 26.

B. Sapp, A. Toshev, and B. Taskar (2010). *Cascaded Models for Articulated Pose Estimation*. Cited on page 77.

B. Sapp, D. Weiss, and B. Taskar (2011b). Parsing Human Motion with Stretchable Models, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 78.

R. C. Schank and R. P. Abelson (1977). *Scripts, Plans, Goals and Understanding*. Cited on page 94.

W. Scheirer, A. Rocha, A. Sapkota, and T. Boult (2012). Towards Open Set Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 99(PrePrints), p. 1. Cited on page 16.

M. Schmidt (2013). *UGM: Matlab code for undirected graphical models, di.ens.fr/∼mschmidt/Software/UGM.html*. Cited on page 141.

B. Schölkopf, J. C. Platt, J. Shawe-Taylor, A. J. Smola, and R. C. Williamson (2001). Estimating the Support of a High-Dimensional Distribution, *Neural Computation*, vol. 13(7), pp. 1443–1471. Cited on page 16.

C. Schuldt, I. Laptev, and B. Caputo (2004). Recognizing human actions: a local SVM approach, in *ICPR 2004*. Cited on pages 21, 72, 73, 74, and 88.

H. Schwenk (2012). Continuous Space Translation Models for Phrase-Based Statistical Machine Translation, in *24th International Conference on Computational Linguistics (COLING) 2012*. Cited on page 159.

V. Sharmanska, N. Quadrianto, and C. H. Lampert (2012). Augmented Attribute Representations, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on pages 16, 19, 106, and 107.

J. Shotton, A. W. Fitzgibbon, M. Cook, T. Sharp, M. Finocchio, R. Moore, A. Kipman, and A. Blake (2011). Real-time human pose recognition in parts from single depth images, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 22 and 75.

A. Shrivastava, S. Singh, and A. Gupta (2012). Constrained Semi-Supervised Learning Using Attributes and Comparative Attributes, in *Proceedings of the European Conference on Computer Vision (ECCV) 2012*. Cited on page 107.

Z. Sia, M. Peib, B. Yaoa, and S.-C. Zhua (2011). Unsupervised Learning of Event AND-OR Grammar and Semantics from Video, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on page 70.

C. Silberer and M. Lapata (2012). Grounded Models of Semantic Representation, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2012*. Cited on pages 27, 122, and 133.

P. Simard, B. Victorri, Y. LeCun, and J. S. Denker (1991). Tangent prop-a formalism for specifying selected invariances in an adaptive network, in *Advances in Neural Information Processing Systems (NIPS) 1991*. Cited on page 16.

V. Singh and R. Nevatia (2011). Action Recognition in Cluttered Dynamic Scenes using Pose-Specific Part Models, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2011*. Cited on pages 22, 70, 73, and 75.

J. Sivic, B. C. Russell, A. A. Efros, A. Zisserman, and W. T. Freeman (2005). Discovering Object Categories in Image Collections, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2005*. Cited on page 105.

C. Snoek, M. Worring, J. van Gemert, J.-M. Geusebroek, and A. W. M. Smeulders (2006). The challenge problem for automated detection of 101 semantic concepts in multimedia, in *ACM Multimedia 2006*. Cited on page 91.

R. Socher and L. Fei-Fei (2010). Connecting Modalities: Semi-supervised Segmentation and Annotation of Images Using Unaligned Text Corpora, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 23 and 159.

R. Socher, M. Ganjoo, C. D. Manning, and A. Ng (2013). Zero-Shot Learning Through Cross-Modal Transfer, in *Advances in Neural Information Processing Systems 26 2013*. Cited on pages 14, 15, 16, 19, and 153.

E. H. Spriggs, F. de la Torre, and M. Hebert (2009). Temporal segmentation and activity classification from first-person sensing, in *Egoc.Vis. 2009*. Cited on page 75.

M. Stark, M. Goesele, and B. Schiele (2009). A Shape-Based Object Class Model for Knowledge Transfer, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on pages 17, 37, and 156.

S. Stein and S. J. McKenna (2013). Combining Embedded Accelerometers with Computer Vision for Recognizing Food Preparation Activities, in *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp 2013), Zurich, Switzerland 2013*. Cited on page 74.

M. Steyvers (2010). Combining feature norms and text data with topic models, *Acta Psychologica*, vol. 133(3), pp. 234 – 243. Cited on pages 27 and 122.

M. Stikic, D. Larlus, S. Ebert, and B. Schiele (2011). Weakly supervised recognition of daily life activities with wearable sensors, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 33(12), pp. 2521–2537. Cited on page 17.

A. Subramanya, S. Petrov, and F. Pereira (2010). Efficient graph-based semi-supervised learning of structured tagging models, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2010*. Cited on page 17.

J. Sung, C. Ponce, B. Selman, and A. Saxena (2011). Human Activity Detection from RGBD Images, *CoRR*, vol. abs/1107.0169. Cited on pages 22 and 75.

W. Susanto, M. Rohrbach, and B. Schiele (2012). 3D object detection with multiple kinects, in *Computer Vision - ECCV 2012 : Workshops and Demonstrations 2012*. Cited on page 6.

G. Szarvas, T. Zesch, and I. Gurevych (2011). Combining heterogeneous knowledge resources for improved distributional semantic models, in *PCICLing 2011*. Cited on pages 2, 57, and 151.

C. C. Tan, Y.-G. Jiang, and C.-W. Ngo (2011). Towards textually describing complex video contents with audio-visual concept classifiers, in *ACM Multimedia 2011*. Cited on pages 29, 30, 136, 139, and 158.

K. Tang, L. Fei-Fei, and D. Koller (2012). Learning Latent Temporal Structure for Complex Event Detection, in *IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2012*. Cited on pages 22 and 24.

K. Tang, B. Yao, L. Fei-Fei, and D. Koller (2013). Combining the Right Features for Complex Event Recognition, in *International Conference on Computer Vision (ICCV) 2013*. Cited on page 22.

D. M. J. Tax and R. P. W. Duin (2008). Growing a multi-class classifier with a reject option, *Pattern Recognition Letters*, vol. 29(10), pp. 1565–1570. Cited on page 16.

G. W. Taylor, R. Fergus, Y. LeCun, and C. Bregler (2010). Convolutional learning of spatio-temporal features, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*, pp. 140–153, Springer. Cited on pages 21, 157, and 159.

S. Tellex, T. Kollar, S. Dickerson, M. R. Walter, A. G. Banerjee, S. J. Teller, and N. Roy (2011). Understanding Natural Language Commands for Robotic Navigation and Mobile Manipulation, in *Proceedings of the Twenty-Fifth AAAI Conference on Artificial Intelligence, AAAI 2011, San Francisco, California, USA, August 7-11, 2011 2011*. Cited on page 160.

M. Tenorth, J. Bandouch, and M. Beetz (2009). The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition, in *THEMIS 2009*. Cited on pages 72 and 74.

C. L. Teo, Y. Yang, H. Daume, C. Fermuller, and Y. Aloimonos (2012). Towards a Watson that sees: Language-guided action recognition for robots, in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA) 2012*. Cited on page 23.

S. Thater, H. Fürstenau, and M. Pinkal (2011). Word Meaning in Context: A Simple and Effective Vector Model, in *Proceedings of IJCNLP 2011 2011*. Cited on pages 130 and 131.

S. Thrun (1996). Is Learning the n-th Thing Any Easier than Learning the First, in *Advances in Neural Information Processing Systems (NIPS) 1996*. Cited on pages 16, 37, and 104.

A. Torralba, R. Fergus, and W. T. Freeman (2008). 80 million tiny images: a large dataset for non-parametric object and scene recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 58.

A. Torralba, K. Murphy, and W. Freeman (2004). Sharing visual features for multiclass and multiview object detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2004*. Cited on pages 53 and 106.

A. Torralba, K. P. Murphy, W. T. Freeman, and M. A. Rubin (2003). Context-Based Vision System for Place and Object Recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2003*. Cited on page 30.

D. Tran and A. Sorokin (2008). Human Activity Recognition with Metric Learning, in *Proceedings of the European Conference on Computer Vision (ECCV) 2008*. Cited on page 107.

P. D. Turney (2001). Mining the Web for Synonyms: PMI-IR versus LSA on TOEFL, in *European Conference on Machine Learning 2001*. Cited on page 18.

P. D. Turney and P. Pantel (2010). From Frequency to Meaning. Vector Space Models for Semantics, *JAIR*. Cited on page 120.

E. Tzoukermann, J. Neumann, J. Kosecka, C. Fermuller, I. Perera, F. Ferraro, B. Sapp, R. Chaudhry, and G. Singh (2011). Language Models for Semantic Extraction and Filtering in Video Action Recognition, in *AAAI Workshop on Language-Action Tools for Cognitive Artificial Agents 2011*. Cited on pages 18, 19, and 121.

K. E. A. van de Sande, T. Gevers, and C. G. M. Snoek (2010). Evaluating Color Descriptors for Object and Scene Recognition, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*. Cited on page 59.

A. Vedaldi and A. Zisserman (2010). Efficient Additive Kernels via Explicit Feature Maps, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 80, 97, and 112.

L. von Ahn and L. Dabbish (2004). Labeling images with a computer game, in *Proceedings of SIGCHI 2004 2004*. Cited on pages 27 and 121.

C. Vondrick, D. Patterson, and D. Ramanan (2012). Efficiently Scaling up Crowdsourced Video Annotation, *International Journal of Computer Vision (IJCV)*. Cited on page 124.

G. Wang and D. Forsyth (2009). Joint Learning of Visual Attributes, Object Classes and Visual Saliency, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on pages 36 and 37.

H. Wang, A. Kläser, C. Schmid, and C. Liu (2013). Dense Trajectories and Motion Boundary Descriptors for Action Recognition, *International Journal of Computer Vision (IJCV)*. Cited on pages 1, 21, 22, 30, 71, 74, 77, 80, 137, 142, 143, 151, and 157.

H. Wang, A. Kläser, C. Schmid, and C.-L. Liu (2011). Action Recognition by Dense Trajectories, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on pages 9, 21, 24, 70, 71, 74, 77, 80, 81, 82, 84, 88, 90, 97, 131, 132, 152, and 154.

H. Wang and C. Schmid (2013). Action Recognition with Improved Trajectories, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2013*. Cited on page 21.

H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid (2009a). Evaluation of local spatio-temporal features for action recognition, in *Proceedings of the British Machine Vision Conference (BMVC) 2009*. Cited on pages 21 and 74.

J. Wang, K. Markert, and M. Everingham (2009b). Learning Models for Object Recognition from Natural Language Descriptions, in *Proceedings of the British Machine Vision Conference (BMVC) 2009*. Cited on pages 18, 19, and 37.

J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, and Y. Gong (2010). Locality-constrained Linear Coding for image classification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on page 59.

M. Weber, M. Welling, and P. Perona (2000). Towards automatic discovery of object categories, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2000*. Cited on page 105.

P. Welinder, S. Branson, T. Mita, C. Wah, F. Schroff, S. Belongie, and P. Perona (2010). Caltech-UCSD birds 200, Technical report, California Institute of Technology. Cited on page 19.

J. Weston, S. Bengio, and N. Usunier (2010). Large scale image annotation: learning to rank with joint word-image embeddings, *Machine Learning*, vol. 81(1), pp. 21–35. Cited on page 15.

Y. W. Wong and R. J. Mooney (2006). Learning for Semantic Parsing with Statistical Machine Translation, in *Proceedings of the Conference of the North American Chapter of the Association for Computational Linguistics 2006*. Cited on pages 138 and 160.

J. Wu, A. Osuntogun, T. Choudhury, M. Philipose, and J. M. Rehg (2007). A Scalable Approach to Activity Recognition based on Object Use, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2007*. Cited on page 90.

T.-F. Wu, C.-J. Lin, and R. C. Weng (2004). Probability Estimates for Multi-class Classification by Pairwise Coupling, *JMLR04*. Cited on page 39.

T. Z. Xi Zhou, Kai Yu and T. S. Huang (2010). Image Classification Using Super-Vector Coding of Local Image Descriptors, in *Proceedings of the European Conference on Computer Vision (ECCV) 2010*. Cited on page 59.

W. Yang, Y. Wang, and G. Mori (2010). Recognizing Human Actions from Still Images with Latent Poses, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2010*. Cited on pages 22 and 75.

Y. Yang and D. Ramanan (2011). Articulated pose estimation with flexible mixtures-of-parts., in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2011*. Cited on page 77.

Y. Yang and D. Ramanan (2013). Articulated Human Detection with Flexible Mixtures of Parts, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 35. Cited on pages 1 and 22.

Y. Yang, C. L. Teo, H. Daumé III, and Y. Aloimonos (2011). Corpus-guided sentence generation of natural images, in *Proceedings of the Conference on Empirical Methods in Natural Language Processing (EMNLP) 2011*. Cited on page 29.

B. Yao, X. Jiang, A. Khosla, A. L. Lin, L. J. Guibas, and L. Fei-Fei (2011). Action Recognition by Learning Bases of Action Attributes and Parts, in *International Conference on Computer Vision (ICCV) 2011*. Cited on pages 71 and 90.

B. Yao and F.-F. Li (2012). Recognizing Human-Object Interactions in Still Images by Modeling the Mutual Context of Objects and Human Poses, *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, vol. 34(9), pp. 1691–1703. Cited on page 20.

B. Yao, X. Yang, L. Lin, M. W. Lee, and S. C. Zhu (2010). I2T: Image Parsing to Text Description, in *CVPRW 2010*. Cited on page 31.

L. Yeffet and L. Wolf (2009). Local Trinary Patterns for human action recognition, in *Proceedings of the IEEE International Conference on Computer Vision (ICCV) 2009*. Cited on pages 21 and 74.

H. Yu and J. M. Siskind (2013). Grounded language learning from videos described with sentences, in *Proceedings of the Annual Meeting of the Association for Computational Linguistics 2013*. Cited on pages 28, 33, and 153.

J. S. Yuan, Z. C. Liu, and Y. Wu (2009). Discriminative subvolume search for efficient action detection, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2009*. Cited on pages 72 and 73.

M. D. Zeiler and R. Fergus (2013). Visualizing and Understanding Convolutional Networks, Technical report, arXiv:1311.2901v1. Cited on page 59.

T. Zesch and I. Gurevych (2010). Wisdom of Crowds versus Wisdom of Linguists - Measuring the Semantic Relatedness of Words, *JNLE*, vol. 16. Cited on page 40.

L. Zhang, M. U. G. Khan, and Y. Gotoh (2011). Video scene classification based on natural language description, in *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops) 2011*. Cited on pages 23 and 24.

R. Zhao, W. Ouyang, and X. Wang (2013). Unsupervised Salience Learning for Person Re-identification, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on page 159.

D. Zhou, O. Bousquet, T. N. Lal, Jason Weston, and B. Schölkopf (2004). Learning with Local and Global Consistency, in *Advances in Neural Information Processing Systems (NIPS) 2004*. Cited on pages 17, 105, 106, 109, 152, and 154.

X. Zhu, Z. Ghahramani, and J. Lafferty (2003). Semi-supervised learning using gaussian fields and harmonic functions, in *Proceedings of the International Conference on Machine Learning (ICML) 2003*. Cited on pages 17, 105, and 106.

G. Zimmermann, C. K. Sung, G. Bosch, and J. R. J. Schirra (1987). From Image Sequences to Natural Language: Descriptions of Moving Objects, Technical Report Bericht Nr. 17, Universität des Saarlandes, Informatik, Saarbrücken. Cited on page 29.

A. Zinnen, U. Blanke, and B. Schiele (2009). An Analysis of Sensor-Oriented vs. Model-Based Activity Recognition, in *ISWC 2009*. Cited on pages 24, 71, 77, 79, and 154.

C. L. Zitnick and D. Parikh (2013). Bringing semantics into focus using visual abstraction, in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR) 2013*. Cited on pages 25 and 28.

C. L. Zitnick, D. Parikh, and L. Vanderwende (2013). Learning the Visual Interpretation of Sentences, in *IEEE International Conference on Computer Vision (ICCV), 2013 2013*. Cited on pages 3 and 28.

A. Zweig and D. Weinshall (2007). Exploiting Object Hierarchy: Combining Models from Different Category Levels, in *ICCV 2007*. Cited on pages 37, 52, 53, and 106.

CURRICULUM VITAE

Marcus Rohrbach

Date of birth:	10/03/1983 in Frankfurt, Germany	
Citizenship:	German	
Education:	01/2010 - today	PhD student in computer science, University of Saarland, Germany; supervised by Prof. Dr. Bernt Schiele and Prof. Dr. Manfred Pinkal.
	10/2007 - 12/2009	Master of Science in computer science with a minor in Anglistics, TU Darmstadt (with distinction).
	09/2006 - 04/2007	Visiting student at the University of British Columbia, Canada.
	10/2003 - 09/2006	Bachelor of Science in computer science, TU Darmstadt (very good).
	06/2003	Abitur (high school graduation)
	09/2000 - 06/2001	Scarborough Sixth Form College, England.
Experience:	09/2010 - today	Research assistant with Prof. Dr. Bernt Schiele, Max Planck Institute for Informatics, Saarbrücken, Germany.
	01/2010 - 08/2010	Research assistant with Prof. Dr. Bernt Schiele, TU Darmstadt, Germany.
	08/2008 - 03/2009	Internship at Daimler Research, Ulm, Germany.
Invited Talk:	06/2013	Fine-Grained Visual Categorization Workshop in conjunction with CVPR 2013.
Academic activities:	PC-member	CVPR 2012, 2013, 2014; ECCV 2012; ICCV 2013.
	Reviewer	IEEE Transactions on Pattern Analysis and Machine Intelligence (2011, 2012); ACM Transactions on Interactive Intelligent Systems (2013).

PUBLICATIONS

[12] *Transfer Learning in a Transductive Setting.*
Marcus Rohrbach, Sandra Ebert, and Bernt Schiele.
In Advances in Neural Information Processing Systems 26 (**NIPS**), 2013.

[11] *Translating video content to natural language descriptions.*
Marcus Rohrbach, Wei Qiu, Ivan Titov, Stefan Thater, Manfred Pinkal, and Bernt Schiele.
In International Conference on Computer Vision (**ICCV**), 2013.

[10] *Multi-view Pictorial Structures for 3D Human Pose Estimation.*
Sikandar Amin, Mykhaylo Andriluka, Marcus Rohrbach, and Bernt Schiele.
In British Machine Vision Conference (**BMVC**) 2013.

[9] *Grounding Action Descriptions in Videos.*
Michaela Regneri, Marcus Rohrbach, Dominikus Wetzel, Stefan Thater, Bernt Schiele, and Manfred Pinkal.
Transactions of the Association for Computational Linguistics (**TACL**), 1:25–36, 2013.

[8] *Script data for attribute-based recognition of composite activities.*
Marcus Rohrbach, Michaela Regneri, Mykhaylo Andriluka, Sikandar Amin, Manfred Pinkal, and Bernt Schiele.
Computer Vision - **ECCV** 2012 : 12th European Conference on Computer Vision, 2012.

[7] *3D object detection with multiple kinects.*
Wandi Susanto, Marcus Rohrbach, and Bernt Schiele.
Computer Vision - **ECCV** 2012 : Workshops and Demon-strations, volume 7584 of Lecture Notes in Computer Science, 2012.

[6] *A database for fine grained activity detection of cooking activities.*
Marcus Rohrbach, Sikandar Amin, Mykhaylo Andriluka, and Bernt Schiele.
In 2012 IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2012.

[5] *Evaluating Knowledge Transfer and Zero-Shot Learning in a Large-Scale Setting.*
Marcus Rohrbach, Michael Stark, and Bernt Schiele.
In 2011 IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2011.

[4] *The Benefits of Dense Stereo for Pedestrian Detection.*
Christoph G. Keller, Markus Enzweiler, Marcus Rohrbach, David Fernández and Llorca, Christoph Schnörr, and Dariu M. Gavrila.
IEEE Transactions on Intelligent Transportation Systems (**TITS**), 12:1096–1106,2011.

[3] *Combining Language Sources and Robust Semantic Relatedness for Attribute-Based Knowledge Transfer.*
Marcus Rohrbach, Michael Stark, György Szarvas, Bernt Schiele.
In First International Workshop on Parts and Attributes (**PnA2010**), in conjunction with **ECCV** 2010, published in Trends and Topics in Computer Vision : ECCV 2010 Workshops, Volume 6553 of Lecture Notes in Computer Science, 2012.

[2] *"What Helps Where – And Why? Semantic Relatedness for Knowledge Transfer."*
Marcus Rohrbach, György Szarvas, Iryna Gurevych, Bernt Schiele.
In IEEE Conference on Computer Vision and Pattern Recognition (**CVPR**), 2010.

[1] *High-level fusion of depth and intensity for pedestrian classification.*
Marcus Rohrbach, Markus Enzweiler, and Dariu M. Gavrila.
In Proceedings of the 31st DAGM Symposium on Pattern Recognition (**DAGM**), 2009.