# Recognizing Fine-Grained and Composite Activities Using Hand-Centric Features and Script Data

**Marcus Rohrbach**[1,2] · **Anna Rohrbach**[1] · **Michaela Regneri**[3,6] · **Sikandar Amin**[1,4] · **Mykhaylo Andriluka**[1,5] · **Manfred Pinkal**[3] · **Bernt Schiele**[1]

Received: 19 June 2014 / Accepted: 13 July 2015 / Published online: 22 August 2015

**Abstract** Activity recognition has shown impressive progress in recent years. However, the challenges of detecting fine-grained activities and understanding how they are combined into composite activities have been largely overlooked. In this work we approach both tasks and present a dataset which provides detailed annotations to address them. The first challenge is to detect fine-grained activities, which are defined by low inter-class variability and are typically characterized by fine-grained body motions. We explore how human pose and hands can help to approach this challenge by comparing two pose-based and two hand-centric features with state-of-the-art holistic features. To attack the second challenge, recognizing composite activities, we leverage the fact that these activities are compositional and that the essential components of the activities can be obtained from textual descriptions or scripts. We show the benefits of our hand-centric approach for fine-grained activity classification and detection. For composite activity recognition we find that decomposition into attributes allows sharing information across composites and is essential to attack this hard task. Using script data we can recognize novel composites without having training data for them.

**Keywords** Activity recognition · Fine-grained recognition · Script data · Hand detection

## 1 Introduction

Human activity recognition in video is a fundamental problem in computer vision. State-of-the-art methods (e.g. Tang et al. 2012; Wang et al. 2013b; Wang and Schmid 2013; Karpathy et al. 2014) 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).

While impressive progress has been made, we argue that most works are addressing only a part of the overall activity recognition challenge. Many application scenarios, such as human–robot interaction or elderly care require to understand complex activities (e.g. *does the person prepare food?*), consisting of multiple fine-grained activities and object manipulations (e.g. *is it fried and what is in it?*). Frequently it is important to recognize both, the individual steps and the high level composite activities, e.g. as we have shown for the task of video description (Rohrbach et al. 2014). Consequently we approach both problems in this work: recognizing fine-grained activities and recognizing composite activities. *Fine-grained activities* are defined as a set of activities which are visually very similar, i.e. have a low inter-class variability. *Composite activities* are activities which can be temporally decomposed into multiple shorter activities, i.e. they consist of multiple steps. We note that both the terms are not exclusive, i.e. composite activities can also

Communicated by Ivan Laptev, Josef Sivic, and Deva Ramanan.

✉ Marcus Rohrbach
rohrbach@mpi-inf.mpg.de; rohrbach@berkeley.edu

[1] Max Planck Institute for Informatics, Saarbrücken, Germany

[2] UC Berkeley EECS and ICSI, Berkeley, CA, USA

[3] Department of Computational Linguistics and Phonetics, Saarland University, Saarbrücken, Germany

[4] Department of Informatics, Technische Universität München, München, Germany

[5] Stanford University, Stanford, USA

[6] SPIEGEL-Verlag, IT Department, Saarland University, Hamburg, Germany

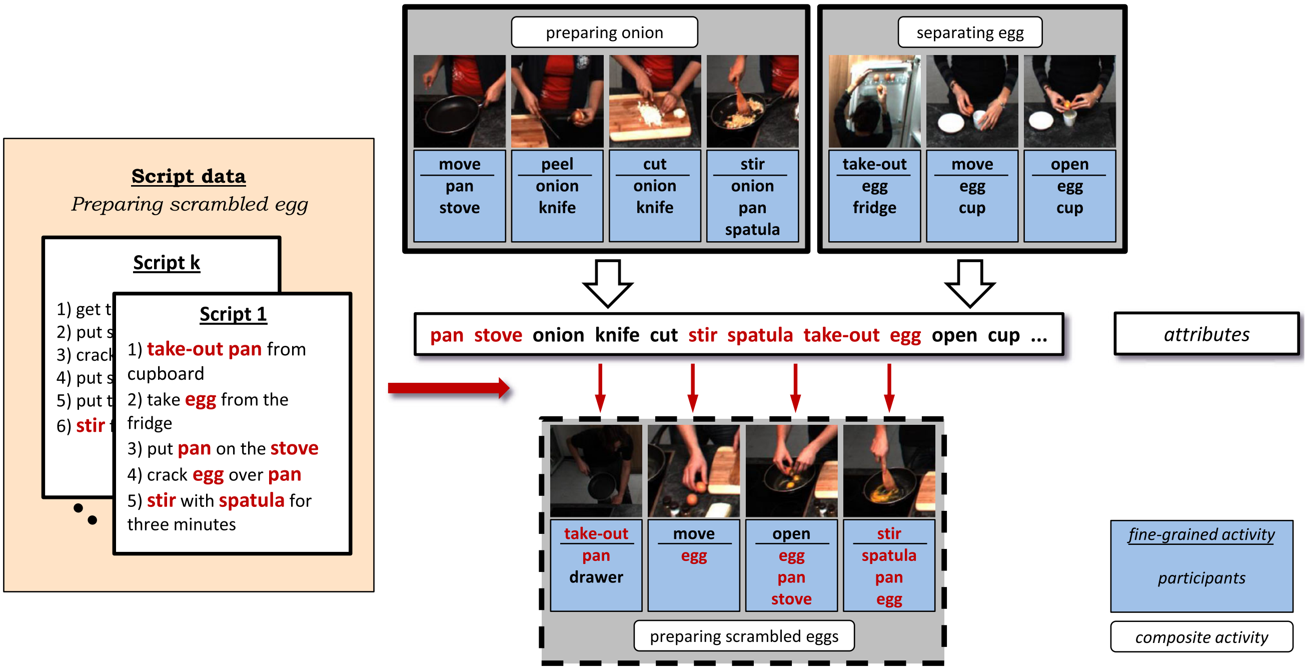


**Fig. 1** Sharing or transferring attributes of composite activities using script data. Composite activities (*gray boxes*) are composed of activities and their participants (*light-blue boxes*), modeled as attributes. These attributes can be transferred to unseen composite activities (*dashed-line box*) with the help of script data which allows estimating the relevant attributes (*red*). Our activities have the additional challenge of being fine-grained, we thus refer to them as fine-grained activities (Color figure online)

be fine-grained. In fact some of our composites are very similar. However, in our work we consider composite activities which consist of fine-grained activities.

When surveying the field we also noticed a lack of datasets allowing to pursue the challenges of fine-grained and composite activity recognition. Specifically this is reflected in the following limiting factors of current benchmark databases. First, while datasets with large numbers of activities exist, the typical inter-class variability is high. This seems rather unrealistic for many domains such as surveillance or elderly care where we need to differentiate between consequentially different but visually similar activities e.g. *hug someone* versus *hold someone* or *throw in garbage* versus *put in drawer*. Second, the activities considered so far are full-body activities, e.g. *jumping* or *running*. This appears rather untypical for many applications where we want to differentiate between more small motion and frequently hand centric activities. Consider e.g. the *cutting activity* in domains such *cooking* (see Fig. 1), *handicraft work* or *surgeries*, as well as different *repairing* activities in the domain of *house keeping* or *machine maintenance* with subtle difference in motion and low inter-class variability. As a third limitation we found that many available databases contain videos of few second length and focus on simple basic-level activities such as *walking* or *drinking*. In contrast, the recognition of longer-term, complex, and composite activities such as *assembling furniture*, *food preparation*, or *surgeries* have been rarely addressed in computer vision. Notable exceptions exist (see Sect. 2) even though these have other limiting factors such as small number of classes.

In this work, which is an extension of our original publications (Rohrbach et al. 2012a, b), we recorded, annotated, and publicly released a large-scale dataset in a kitchen scenario which addresses the discussed limitations. This allows us to work on the challenges of fine-grained and composite activity recognition as follows.

Recognizing fine-grained activities is challenging due to their low inter-class variability. In contrast to fine-grained object recognition challenges where the same object category typically is also visually consistent, activities of the same category are frequently very diverse, i.e. have a high intra-class variability. Consider e.g. the activities *peeling*, which can be very different depending of the participating object: *peeling a carrot* versus *peeling a pineapple*. At the same time, we have to handle small differences between categories, i.e. low inter-class variability, consider e.g. *mix* versus *stir* or *slice* versus *cut dice*. This typically requires to understand the difference between fine-grained body motions. To approach both of these challenges we propose to focus on body pose and hands. As can be seen in Figs. 1 and 2 many fine-grained activities, especially in our kitchen scenario, are hand-centric. Here it is not only important to understand the activity but also the participating object, e.g. *open egg* versus *open tin*. We thus propose to focus on the hand regions for extracting visual features. However, hand detection is a challenging problem in itself in real-world scenarios due to

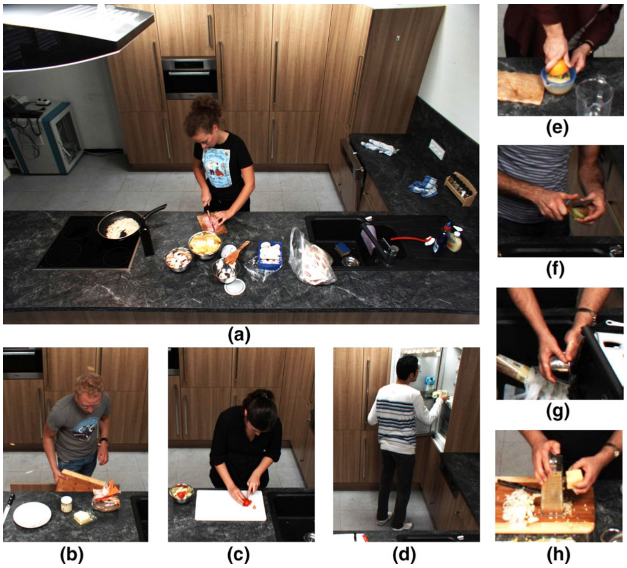


**Fig. 2** Single frames from the dataset depicting fine-grained cooking activities and diverse sets of tools and ingredients (participants). **a** Full scene of *slicing* in the composite activity *omelet*, and crops of **b** *take out*, **c** *dicing*, **d** *take out*, **e** *squeeze*, **f** *peel*, **g** *wash*, **h** *grate* (Color figure online)

a large variability in shape and frequent partial occlusions (Mittal et al. 2011; Gkioxari et al. 2013). To get reliable hand detections, we integrate a hand detector into an articulated pose estimation. Consequently we use the hand position to extract color Sift and Dense Trajectories (Wang et al. 2013a) and learn detectors for fine-grained activities and their participating objects. Recently, Jhuang et al. (2013) showed that exploiting body pose in form of body joints can be beneficial for full-body activities. We explore two approaches based on body pose tracks, motivated from work in the sensor-based activity recognition community (Zinnen et al. 2009).

For recognizing composite activities, state-of-the-art methods, which build on discriminative learning from low-level activity features, experience scalability issues due to the typically highly diverse composite activities and little training data. A promising approach towards scaling 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 building on an attribute-based representation, with attributes denoting the fine-grained activities and the participating objects. For example in Fig. 1 the composite activity *preparing scrambled egg* shares the attributes *stir* and *spatula* with the composite activity *preparing onion* and the attributes *open* and *egg* with the composite activity *separating egg*. Instead of learning a holistic 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. 2013) and have also been applied to activity recognition (Liu et al. 2011). A major challenge to recognize everyday activities is that these composite activities can often be performed in a wide variety of ways, and it is practically infeasible to create a visually annotated training set with all possible alternatives. Instead, we collect a large number of textual descriptions (scripts) for a composite activity to compute the association strength between attributes and composite activities. Using this script data we can not only handle the inherent variation of composites but also recognize unseen composite activities. As illustrated in Fig. 1, the attributes in red are determined to be important for *preparing scrambled eggs* using script data and can be transferred from known composites such as *separating egg* and *preparing onion*.

Our main contributions are as follows. First, we propose several hand- and pose-based activity recognition approaches to recognize fine-grained activities and their object participants. We benchmark them together with state-of-the-art activity recognition features on our dataset. Second, we contribute an attribute-based approach which shares knowledge across composite activities and exploits textual script data to handle their large variability and allows transfer to unseen composite activities. Third, we recorded and annotated a video dataset called *MPII Cooking 2* . It provides challenges for classification and detection of fine-grained activities and their participants, human pose estimation, and composite activity recognition (optionally) using script data. In addition to activity recognition, which is the focus of this work, the dataset is also being used for 3D human pose estimation (Amin et al. 2013), multi-frame pose estimation (Cherian et al. 2014), grounding semantic similarities of natural language sentences in video (Regneri et al. 2013), and for generating natural language descriptions (Rohrbach et al. 2013b, 2014).

The remaining article is structured as follows. We first make an extensive review of related datasets, activity recognition approaches, and the use of text data for visual recognition in Sect. 2. Then we introduce our MPII Cooking 2 dataset in Sect. 3 which we benchmark in the subsequent sections. In Sect. 4 we make a quantitative comparison of our pose-recognition and hand detection with related work on the pose challenge of our dataset. Using the pose-estimation and hand detections we define several visual features and discuss fine-grained activity detection in Sect. 5. In Sect. 6 we present our approach to combine the fine-grained activities to composite activities and integrate script data. In Sect. 7 we evaluate fine-grained and composite activity recognition and then we conclude with the most important findings and directions for future work in Sect. 8.

## 2 Related Work

We first present an overview of the different video activity recognition datasets (Sect. 2.1) and then review recent

**Table 1** Overview of activity recognition datasets

| Dataset | cls, det | Classes | Clips/videos | Subjects | # Frames | Resolution |
|---|---|---|---|---|---|---|
| *Full body pose datasets* | | | | | | |
| KTH (Schuldt et al. 2004) | cls | 6 | 2391 | 25 | ≈200,000 | 160 × 120 |
| USC gestures (Natarajan and Nevatia 2008) | cls | 6 | 400 | 4 | | 740 × 480 |
| MSR action (Yuan et al. 2009) | cls, det | 3 | 63 | 10 | | 320 × 240 |
| *Movie and web video datasets* | | | | | | |
| Hollywood2 (Marszalek et al. 2009) | cls | 12 | 1707/69 | | | |
| UCF 101 (Soomro et al. 2012) | cls | 101 | 13,320 | | ≈2,400,000 | 320 × 240 |
| Sports-1M (Karpathy et al. 2014) | cls | 487 | 1.1 mil | | | |
| HMDB51 (Kuehne et al. 2011) | cls | 51 | 6766 | | | Height: 240 |
| ASLAN (Kliper-Gross et al. 2012) | cls | 432 | 3631/1571 | | | |
| Coffee and Cigarettes (Laptev and Pérez 2007) | det | 2 | 264/11 | | | |
| High Five (Patron-Perez et al. 2010) | cls, det | 4 | 300/23 | | | |
| MPII Movie Description (Rohrbach et al. 2015) | cls, det | | 68,327/94 | | | 1920 × 1080 |
| *Surveillance datasets* | | | | | | |
| PETS 2007 (Ferryman 2007) | det | 3 | 10 | | 32,107 | 768 × 576 |
| UT interaction (Ryoo and Aggarwal 2009) | cls, det | 6 | 120 | 6 | | |
| VIRAT (Oh et al. 2011) | det | 23 | 17 | | 1920 × 1080 | |
| *Assisted daily living datasets* | | | | | | |
| TUM Kitchen (Tenorth et al. 2009) | det | 10 | 20/4 | | 36,666 | 384 × 288 |
| CMU-MMAC (De la Torre et al. 2009) | cls, det | >130 | 26 | | | 1024 × 768 |
| URADL (Messing et al. 2009) | cls | 17 | 150/30 | 5 | ≤ 50,000 | 1280 × 720 |
| MPII Cooking 2 (our dataset) | cls, det | 67/ 59 | 14,105/273 | 30 | 2,881,616 | 1624 × 1224 |

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

approaches to activity recognition (Sect. 2.2), putting a focus on works which use human pose as a cue. Next we discuss works which use textual information for improved recognition of activities (Sect. 2.3). We conclude by relating them to our work (Sect. 2.4).

### 2.1 Activity Datasets

Even when excluding single image action datasets such as the Stanford-40 Action Dataset (Yao et al. 2011b) or the Pascal Action Classification Challenge (Everingham et al. 2011), the number of proposed activity datasets is quite large (Chaquet et al. (2013) survey 68 datasets). Here, we focus on the most important ones with respect to database size, usage, and similarity to our proposed dataset (see Table 1). We distinguish four broad categories of datasets: full body pose, movie and web, surveillance, and assisted daily living datasets—our dataset falls in the last category.

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[1] and similar datasets (Liu et al. 2009; Niebles et al. 2010; Rodriguez et al. 2008) 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). UCF50 has been extended to UCF 101 (Soomro et al. 2012), significantly increasing the number of categories to 101 and including 2.4 million frames at a rather low resolution of 320 × 240. The Sports-1M dataset exceeds all datasets with respect to number of clips (1.1 million) and categories (487 different sports), which are, however, only weakly labeled. 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 the database

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

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. ASLAN (Kliper-Gross et al. 2012) focuses on a larger number of activities but with little training data per category. The task is to identify similar videos rather than categorising them. A significantly larger video collection is evaluated during the TRECVID challenge (Over et al. 2012). The 2012 challenge consisted of 291 h of short videos from the Internet Archive (archive.org) and more than 4000 h of multi-media (audio and video) data. The challenge covers 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 recent MPII Movie Description dataset (Rohrbach et al. 2015) does not label clips with labels but with natural sentences which are sourced from movie scripts and audio descriptions for the blind.

The third category of datasets is targeted towards surveillance. The PETS (Ferryman 2007) or SDHA2010[2] workshop datasets contain real world situations from 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 h of video. Although the video is high-resolution people are only of 20 to 180 pixel height. Overall the surveillance activities are very different to ours which are challenging with respect to fine-grained hand motion.

Next we discuss the domain of *Assisted daily living (ADL) datasets*, which also includes our dataset. 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 composite 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 59 diverse dishes, where each subject uses different ingredients and tools in each dish. The authors also record an egocentric view. Similarly to (Farhadi et al. 2010; 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 on the one hand a detection challenge of fine-grained activities and on the other hand a recognition challenge of highly variable composite activities.

### 2.2 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.

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 histograms of oriented 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 Trajectories approach from Wang et al. (2013a) uses this idea: it tracks dense feature points and extracts strong video features around these tracks, namely HOG, HOF, and Motion Boundary Histograms (MBH, Dalal et al. 2006). They report state-of-the art results on several datasets including KTH (Schuldt et al. 2004), UCF YouTube (Liu et al. 2009), Hollywood2 (Marszalek et al. 2009), and HMDB51 (Kuehne et al. 2011). Recently, Wang and Schmid (2013) improved their approach by removing background flow and by ensuring that detected humans do not contribute to the background motion estimation. Additionally

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

they replace the BoW encoding with Fisher vectors. The computational effort of this approach can be significantly reduced by replacing dense flow with motion information from video compression Kantorov and Laptev (2014). As alternative to manually defined activity features, Taylor et al. (2010), Baccouche et al. (2011), Le et al. (2011), and Ji et al. (2013) use deep learning with convolutional neural networks to learn an activity feature representation. So far these approaches cannot reach the manually defined Dense Trajectories even when learning on a database of over a 1 million videos (Karpathy et al. 2014).

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). Earlier work in human pose based activity recognition employed motion capture systems using physical on-body markers to reliably capture human poses, e.g. (Campbell and Bobick 1995). Such an approach is impractical for recording realistic data. Recently a number of hand and pose-centric approaches have been proposed for activity recognition for more realistic video recordings (Fathi et al. 2011; Packer et al. 2012; Yao et al. 2011a; Sung et al. 2011; Raptis and Sigal 2013; Jhuang et al. 2013) as well as in static images (Yang et al. 2011; Yao and Li 2012). Packer et al. demonstrate impressive results in recognition of kitchen activities using body poses recovered from depth images. Fathi et al. (2011) propose a hand-centric approach for learning effective models of activities from egocentric video by observing regularities in hand-object interactions. Hand poses have been shown to facilitate extraction of appearance features for activity recognition in static images (Karlinsky et al. 2010). Pose-based models are effective for activity recognition when body poses can be estimated reliably, as e.g. in depth images (Packer et al. 2012; Sung et al. 2011). Mittal et al. (2011) and Gkioxari et al. (2013) aim for specialized representations for hands, but do not apply them to pose estimation or activity recognition. 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 Trajectories features (Wang et al. 2013a); this is also true for estimated pose using (Yang and Ramanan 2013) but only on a subset where the full body is visible.

Although several interesting techniques have been proposed to model the temporal structure of videos, they typically perform only below or on par with bag-of-word based approaches: A simple temporal structure is encoded in the template-based Action MACH from Rodriguez et al. (2008), Brendel and Todorovic (2011) model temporal and spatial structure by segmenting the space-temporal volume, and Niebles et al. (2010) model activities as a temporal composition of primitive actions and discriminatively learn such models. 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. An AND/OR graph structure can be used to combine different features at its nodes (Tang et al. 2013) or model co-occurring and consecutive actions (Gupta et al. 2009). Recently Pirsiavash and Ramanan (2014) have shown how to efficiently parse activity videos with segmental grammars.

### 2.3 Natural Language Text for Activity Recognition

Natural language descriptions have shown beneficial for image segmentation (Socher and Fei-Fei 2010) or recognizing object categories (Wang et al. 2009b; Elhoseiny et al. 2013). Similar to our work, Elhoseiny et al. 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. Semantic similarity from linguistic resources has also been used to allow zero-shot recognition in images via attributes and direct similarity (Rohrbach et al. 2010) and by learning an embedding into a linguistic word vector space (Socher et al. 2013; Frome et al. 2013). Additionally to transferring knowledge one can exploit the unlabeled instances to improve recognition, assuming a transductive setting. For this, Fu et al. (2013) exploit the test-data distribution by performing a single round of self-training by averaging over the k-nearest neighbors.

Teo et al. (2012) improve activity recognition by adding object detectors, which are selected based on the linguistic co-occurrence statistics in the newswire Gigaword Corpus. A similar idea is pursued by Motwani and Mooney (2012), who mine and cluster verbs from descriptions of the video snippets in the MSVD dataset (Chen and Dolan 2011). Zhang et al. (2011) show that tf∗idf can identify the most relevant terms in text descriptions collected for seven video scenes allowing to yields close to perfect (98 %) recognition accuracy on their dataset. Ramanathan et al. (2013) jointly recognize actions and roles in YouTube videos using their captions. They mine a large number of YouTube descriptions and use a topic model to estimate the semantic relatedness between an action/role and a description.

Another line of work focuses on describing videos with natural language descriptions. Recently Guadarrama et al. (2013) generated simple sentences for the Microsoft Video

Description corpus (Chen and Dolan 2011) containing challenging web videos. Das et al. (2013) compose descriptions for kitchen videos of their YouCook dataset showing YouTube cooking videos. Finally, we have shown how to learn a translation model for generating natural sentences on our dataset (Rohrbach et al. 2013b).

### 2.4 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 Trajectories approach from Wang et al. (2013a) as well as two pose-based and two hand centric approaches on our MPII Cooking 2 dataset. 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 the relational and distance features defined on joints by Jhuang et al. Similarly to their work we define relational and distance metrics between joints per frame and over time. However, our activities contain very subtle motions and the people have a very similar pose for most activities, which reduces the benefits of this feature representation. Jhuang et al. examine the advantages of focusing Dense Trajectories (Wang et al. 2013a) on body joints. In our static scene (holistic) Dense Trajectories are already restricted to human body as the features are only extracted on moving points. However, in this work we propose to focus on hands, as they are the main cue for recognizing our fine-grained activities and participating objects.

In Amin et al. (2013) we improve the hand localization by leveraging multiple cameras to handle self-occlusion. In this work we remain monocular and propose to use a specialized hand detector to improve pose estimation and activity recognition.

To improve fine-grained activities and their participating objects we train a classifier on stacked classifier scores from co-occurring activities/objects as well as from temporal context after max pooling. Classifier stacking has previously been explored e.g. in (Ting and Witten 1997; Liu et al. 2012; Sill et al. 2009). Most relevant to our work, Liu et al. (2012) try to optimize the usage of training data and avoid overfitting when learning stacked video classifiers. This could be beneficial when applied to our approach.

In this work 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. Script data allows for zero-shot recognition, which has mainly been used for object recognition, but also for multi-media data by Fu et al. (2013). Fu et al. learn a latent attribute representation on the known classes, but then use manually defined attribute associations to transfer.

While the temporal structure, i.e. temporal ordering, 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 this work we exploit temporal co-occurrence within the same time interval and context of short actions and their participating objects within the entire video using max pooling. For long term composite activities we aggregate its components with max pooling ignoring the 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. Bojanowski et al. (2014) have recently shown the benefit of movie scripts as a weak supervision. They use the ordering constraints provided by the script data to localize the actions and to learn action models.

Finally we shortly summarize how this work extends our original publications (Rohrbach et al. 2012a, b). First, we updated the dataset by correcting and unifying some of the annotations and adding a few more videos. We refer to this new version as MPII Cooking 2. It supersedes both previous datasets, see Table 3. Second, we present hand-centric approaches for fine-grained recognition, namely an integration of pose-estimation and hand detector and Hand centric features for activity recognition (arXiv: Senina et al. 2014). Third, we integrated our Propagated Semantic Transfer (PST) from Rohrbach et al. (2013b) for composite recognition. Fourth, we extended qualitative and quantitative results. Fifth, we extended the discussion of related work. Sixth, we rerun experiments with updated version of Dense Trajectories (Wang and Schmid 2013). And last, we will release the updated version of the dataset, new intermediate features as well as the script data.

## 3 Dataset “MPII Cooking 2”

For our dataset we video-recorded human subjects cooking a diverse set of dishes, e.g. *making pizza* or *preparing cucumber*. The dishes form the *composite activities* and the individual steps taken are the *fine-grained activities*, e.g. *cut*, *pour*, or *spice*. All videos have a composite label and are annotated with time intervals. Each time interval has a fine-grained activity and the participating objects as labels. A subset of frames was annotated with human pose and hands. In the following we provide details and statistics of the dataset, Figs. 1 and 2 show example frames of the dataset.

### 3.1 Dataset Statistics and Versions

We recorded 30 subjects in 273 videos with a total length of more than 27 h or 2,881,616 frames. Each video contains a single subject preparing a certain dish.

The dataset was recorded in two batches. The first part contains few, but very diverse and complex dishes (see upper part of Table 2) and was presented in Rohrbach et al. (2012a). The second part, presented in Rohrbach et al. (2012b), focuses on composite activities and thus contains significantly more dishes/composites which are slightly shorter and simpler, see lower part of Table 2. The second set of composite activities are selected according to our script corpus which we describe below in Sect. 3.4. We ignored some of them which were either too elementary to form a composite activity (e.g. *how to secure a chopping board*), were duplicates with slightly different titles, or because of limited availability of the ingredients (e.g. *butternut squash*).

For this work we corrected and unified some of the annotations and added a few more videos. We refer to this new dataset version as MPII Cooking 2. It supersedes both previous datasets. Table 3 compares the different versions and shows different statistics about them. The table also shows the proposed training/validation/test split, which is selected in a way that for all 31 composite activities in the test set, there are at least 3 training/validation videos and there is no overlap between training, validation, and test subjects. In contrast to the earlier versions we avoid multiple test splits for simpler evaluation and to reduce the computational burden for other researchers evaluating on the dataset.

### 3.2 Dataset Recording and Annotation Protocol

To record realistic behavior we neither asked subjects to perform certain activities nor to follow a certain recipe but we told them only which dish they should prepare. This resulted in a larger variety of how subjects prepared things. This means subjects used different tools for preparation (*knife* or *peeler* for *peeling*), took different steps (e.g. some people cooked the vegetables some did not), and did things in different temporal orders for the same dish (e.g. *washed* the vegetable before or after they *peeled* it). Before the recording the subjects were shown our kitchen and places of tools and ingredients to feel at home. During the recording subjects could ask questions in case of

**Table 2** Composite activities (dishes) of MPII Cooking 2 dataset, composites marked in bold are part of the test split

| | |
|---|---|
| MPII Cooking | Sandwich, salad, fried potatoes, potato pancake, omelet, soup, pizza, casserole, mashed potato, snack plate, cake, fruit salad, cold drink, and hot drink |
| MPII Composites | **Cooking pasta**, juicing {**lime**, **orange**}, making {**coffee**, **hot dog**, **tea**}, pouring beer, preparing {asparagus, **avocado**, **broad beans**, broccoli and cauliflower, **broccoli**, carrots and potatoes, **carrots**, **cauliflower**, **chilli**, **cucumber**, **figs**, **garlic**, **ginger**, **herbs**, **kiwi**, **leeks**, **mango**, **onion**, **orange**, peach, peas, **pepper**, **pineapple**, **plum**, **pomegranate**, **potatoes**, **scrambled eggs**, spinach, spinach and leeks}, **separating egg**, sharpening knives, **slicing loaf of bread**, using {microplane grater, pestle and mortar, speed peeler, **toaster**, tongs}, zesting lemon |

**Table 3** Dataset statistics

| | Videos | Subjects | Categories | | Ground truth | Attribute | Video |
|---|---|---|---|---|---|---|---|
| | | | Composites | Attributes | time intervals | instances | duration (min) |
| MPII Cooking (Rohrbach et al. 2012a) | 44 | 12 | 14 | 218 | 3824 | 15,382 | 3–41 |
| MPII Composites (Rohrbach et al. 2012b) | 212 | 22 | 41 | 218 | 8818 | 33,876 | 1–23 |
| Combined | 256 | 30 | 55 | 218 | 12,642 | 49,258 | 1–41 |
| MPII Cooking 2 | 273 | 30 | 59 | 222 | 14,105 | 54,774 | 1–41 |
| - Training set | 201 | 24 | 58 | 222 | 10,931 | 42,619 | 1–41 |
| - Validation set | 17 | 1 | 17 | 107 | 445 | 1662 | 1–8 |
| - Test set | 42 | 5 | 31 | 169 | 2102 | 8023 | 1–13 |

Note that the train/val/test split do not add up to the full dataset, as some videos of the test subjects are not used as they have less than three train/val videos problems and some listened to music. We always started the recording with an empty and clean kitchen, prior to the subject entering the kitchen and ended it once the subject declared to be finished, i.e. we did not include the final cleaning process. Most subjects were university students from different disciplines recruited by e-mail and publicly posted flyers. Subjects were paid per hour and cooking experience ranged from beginner cookers to amateur chefs.

Composite activities are annotated on the level of each video. Fine-grained activities were annotated with a two-stage revision phase with start and end frame using the annotation tool Advene (Aubert and Prié 2007). In addition to the activity category each annotation consists of used tools, ingredients, and locations (we refer to them as participants). Composite activities were chosen as described in Sects. 3.1 and 3.4. Activity, tool, ingredient, and location categories were chosen to describe all activities the human subjects were performing. The decision was made after the recording on the base what the human subjects did. With respect to the level of detail, we do not annotate the specific motions (e.g. move arm up or down) but what effect or semantic they have (e.g. open versus close). See Table 7 for the chosen granularity.

We recorded in our kitchen (see Fig. 2a) with a 4D View Solutions system using a Point Grey Grasshopper camera with 1624 × 1224 pixel resolution at 29.4 fps 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 bit-rate of 2500). For most videos we recorded 7 additional camera views on the kitchen, a subset was used and released by Amin et al. (2013). Although they are not used in this work we will make the remaining 7 views available upon publication. All fine-grained and composite activity annotations are also valid for the other cameras as each frame was synchronized across all 8 cameras.

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

The dataset provides furthermore human body pose annotations (see Sect. 3.3), script data (see Sect. 3.4) and there exist textual descriptions in the TACoS (Regneri et al. 2013) and TACoS multi-level corpus (Rohrbach et al. 2014). The descriptions in TACoS describe what happens in a specific video and are temporally aligned to the video, i.e. they provide a textual annotation. In contrast, the scripts used in this work are collected independently of the video and thus contain domain or script knowledge, i.e. what activities and what objects are likely used for a certain dish. As they are not specific to the training videos they allow to transfer and generalize to novel test scenarios.

### 3.3 Pose Challenge

A subset of frames have articulated human pose and hand annotations to learn and evaluate pose estimation approaches and hand detectors. For human pose we annotated the frames with right and left shoulder, elbow, wrist, and hand joints as well as head and torso. We have 2994 frames of 10 subjects for training of pose annotation and an additional of 4250 training images with hand points used for training the hand detector. For testing we sample 1277 frames from all activities with 7 subjects as test set for the pose challenge. All training and test frames are from MPII Cooking (Rohrbach et al. 2012a) and thus avoid an overlap with the test subjects and test composites in MPII Cooking 2.

### 3.4 Mining Script Data for Composite Activities

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 which corresponds to composite activities in our case. Scenarios (e.g. *eating in a restaurant*) are temporally ordered events (*the patron enters restaurant, he takes a seat, he reads the menu, …*) and subjects (*patron, waiter, food, menu, …*). Written event sequences for a scenario can be collected on a large scale using crowd-sourcing (Regneri et al. 2010). We make use of this method to collect scripts for our composite activities and assembling a large number of written sequences for each of those.

We collect natural language sequences similar to Regneri et al. (2010) using Amazon's Mechanical Turk[3]. 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"[4]. All those tasks/scenarios are about processesing ingredients or using certain kitchen tools. In addition to the data we collected in this experiment, we use data from the OMICS corpus (Singh et al. 2002) and Regneri et al. (2010) for 6 kitchen-related composite activities. This results in a corpus with 59 composite activities and 2124 sequences in sum, having a total of 12,958 individual event descriptions. Note that for practical reasons we only recorded videos for 35 of these composite activities as discussed in Sect. 3.1. They are listed in Table 2 under "MPII Composites".

This script corpus provides much more variation than the limited number of video training examples can capture. Of course this also poses a challenge, because we need to

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

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

**Table 4** Three example scripts for the composite activity *preparing cucumber*

| | | |
|---|---|---|
| 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 cutting board. |
| 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. |

overcome the problem of different wordings and coordinated events: Table 4 shows three examples we collected for the composite activity *preparing cucumber*. They differ in verbalization (e.g. *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 noisy but big script corpus.

In Sect. 6.4 we will describe how we extract semantic relatedness from this data.

## 4 Hand Detection and Pose Estimation

One goal of this paper is to investigate the applicability of state-of-the-art pose estimation methods in the context of activity recognition. Therefore, in this section we propose our new pose estimation method based on Andriluka et al. (2011) and benchmark it on our dataset together with state-of-the-art pose estimation methods. Another goal is to demonstrate the importance of hand-based features for recognizing activities and their participants. For this we need to localize hands, which is in itself a challenging task due to partial occlusions, obstruction by manipulated objects, and variability of hand postures. In order to achieve high quality hand localization we leverage two complementary sources of information. We exploit the characteristic appearance of hands in order to train an effective hand detector. We then integrate observations from this detector in our pose estimation approach to take advantage of the context provided by the other body parts. As another finding, we show that localization of all body parts benefits significantly from our specialized hand detector.

In the following we introduce our hand detector (Sect. 4.1) and pose estimation method (Sect. 4.2) as well as how we combine them (Sect. 4.3). In Sect. 4.4 we evaluate our proposed approaches as well as state-of-the-art pose estimation methods on our dataset.

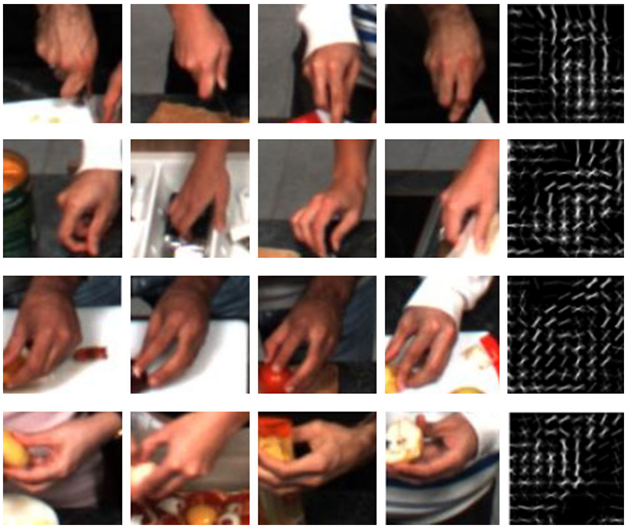


**Fig. 3** Examples of training images assigned to 4 different hand components, each row shows images from one component. Rows 1 and 2 correspond to right hand components, and rows 3 and 4 to left hand components (Color figure online)

### 4.1 Hand Detection Based on Local Appearance

As a basis for our hand detector we rely on the deformable part models (DPM, Felzenszwalb et al. 2010). We discuss several design choices in order to achieve best performance.

#### *4.1.1 Detection of Left and Right Hands*

We aim for a hand detector that can correctly distinguish the left and right hand of a person. The rationale behind this is that for many activities left and right hands have different roles (e.g. for a cutting activity the dominant hand is typically holding a knife while the supporting hand is holding the object that is being cut). Further, we would like to avoid situations when two strong hypotheses for one of the hands are chosen over two hypotheses for both hands. We achieve this by dedicating separate DPM components to left and right hands and jointly training them within the same detector (see examples in Fig. 3). Note that in contrast to the default setting mirroring is switched off in DPM. At test time we pick the best scoring hypothesis among the components corresponding to left and right hands.

#### *4.1.2 Component Initialization*

We capture the variance of hand postures by decomposing the hands' appearance into multiple modes and representing each mode with a specific DPM component. We found that a rather large number of components is necessary to achieve good detection performance. We initialize the components by clustering the HOG descriptors of the training examples using K-means as in Divvala et al. (2012). The detection further improves by first clustering the training examples by hand orientation and then by HOG.

#### *4.1.3 Body Context*

We improve the hand localization by augmenting the hand detector with the context provided by a person detector. We rely on the person detector to constrain the search for hands to the image locations within the extended person bounding box and also constrain the scale of the hands detector to the scale of the person hypothesis.

### 4.2 Pose Estimation

We base our pose estimation approach on the pictorial structures (PS) approach (Fischler and Elschlager 1973; Felzenszwalb and Huttenlocher 2005). In PS the body is represented as a collection of rigid parts linked via a set of pairwise part relationships. Unlike the original model we define a flexible variant of the PS model (FPS) that consists of $N = 10$ parts corresponding to head, torso, as well as left and right shoulders, elbows, wrists and hands. Denoting the configuration of parts as $L = l_1, \ldots, l_N$, and image observations as $D$, the posterior over the part configuration is given by

$$p(L|D) \propto \prod_{(i,j)\in E} p(l_i|l_j) \cdot \prod_{i=1}^{i=N} p(D|l_i), \tag{1}$$

where $E$ is a set of connected part pairs. We build on the publicly available PS implementation from Andriluka et al. (2011). In this model the pairwise connections between parts form a tree structure, which permits efficient and exact inference. The pairwise terms represent the spatial relationships between part positions and are modeled as Gaussians with respect to relative position and orientation of parts. The appearance of individual parts is represented with boosted part detectors and shape context image features. Conceptually the formulation of Andriluka et al. (2011) is similar to flexible mixture of parts model (FMP, Yang and Ramanan 2011). The FMP model represents appearance of each body part with a set of HOG templates. Pairwise terms are adapted depending on the particular template. Parameters of appearance templates and pairwise terms of the FMP model are jointly trained using max-margin objective. The model of Andriluka et al. (2011) relies on a single appearance template for all parts. Parameters of pairwise terms are estimated using maximum likelihood independently from appearance terms. We extend this model by incorporating color features into the part likelihoods by stacking them with shape context features prior to part detector training. We encode the color as a multidimensional histogram in RGB space using 10 bins for each color dimension which results in 1000 dimensional feature vectors. We then concatenate color and shape context features and train boosted part detectors for each part using the combined representation. We use standard AdaBoost for training and rely on the same weak learners as in Andriluka et al. (2011).

### 4.3 Combining Hand Detection and Pose Estimation

We extend the image observations in Eq. 1 with detection hypotheses for left and right hands, which we obtain using the corresponding components of our hand detector. We denote the set of hand hypotheses produced by our hand detector by $H = \{(d_k, s_k)|k = 1, \ldots, K\}$, where $d_k$ is the image position and $s_k$ the detection score. Based on this sparse set of detections we obtain a dense likelihood map for the hand part $l_h$ using a kernel density estimate:

$$p(H|l_h) = \sum_{k=1}^{K} w_k \exp(-\sigma^2 \|d_k - l_h\|^2), \tag{2}$$

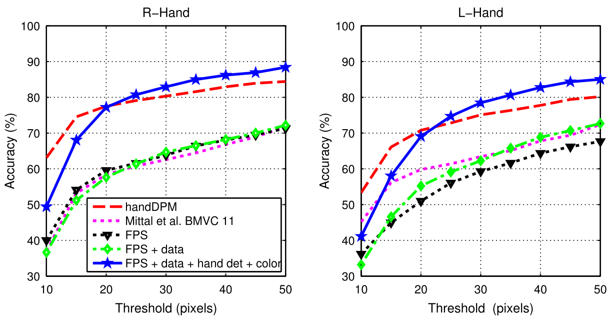
where $w_k = s_k - m$ is a positive weight associated with each hand hypothesis computed by shifting the detection score by the minimal score value $m$. There is no specific upper/lower bound for the scores $s_k$, but since DMP relies on SVM formulation the scores tend to be centered around 0 with confident negative examples having score less than -1. In practice we set $m = -1$ and ignore all detections with a smaller score than $m$.

### 4.4 Evaluation: Pose Estimation and Hand Detection

We first evaluate the results on the upper-body pose estimation task. In order to identify the best 2D pose estimation approach we use our 2D body joint annotations (see Sect. 3.3). For evaluating these methods we adopt the PCP measure (percentage of correct parts) proposed by Ferrari et al. (2008). The results are shown in Fig. 4a. The first three lines compare three state-of-the-art methods: the cascaded

| Method | Torso | Head | upper arm r | upper arm l | lower arm r | lower arm 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 | 78.5 | 79.4 | 61.9 | 64.1 | 62.4 | 61.0 | 67.9 |
| FPS + data | 79.3 | 85.0 | 64.3 | 64.6 | 60.0 | 59.8 | 68.8 |
| FPS + data + hand det | 79.6 | 84.9 | 70.9 | 70.0 | 73.5 | 70.2 | 74.9 |
| FPS + data + color | 80.7 | 85.8 | 69.1 | 67.4 | 69.3 | 65.5 | 73.0 |
| FPS + data + hand det + color | **81.3** | **86.1** | **72.4** | **71.3** | **74.4** | **70.3** | **75.9** |

**(a)**



**(b)**

**Fig. 4** **a** 2D upper body pose estimation results on the "Pose Challenge" of our dataset. The numbers correspond to the "percentage of correct parts" (PCP). **b** Accuracy of different methods for detection of right and left hands for a varying distance (in pixels) from the ground truth position (Color figure online)

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, Andriluka et al. 2011), using their published pose models. Lines 4 and 5 show the models of Yang and Ramanan and Adriluka et al. retrained on our data. Overall the model of Adriluka et al. performs best, achieving 66.0 PCP for all body-parts. We attribute the improvement of PS over FMP to the following. The FMP model encodes different orientation of parts via different appearance templates, whereas the PS model uses a single template that is rotation invariant and is evaluated at all orientations. The FMP model has a larger number of parameters because appearance templates are not shared across different part orientations. A larger number of parameters means that it is easier to overfit the FMP model than the PS model. This could explain the performance differences after retraining on our data. It could also be that finer discretization of body part orientations in the PS model compared to the FMP model is important for good performance. As described above we base our model (FPS) on PS, adding to it flexible part configuration.

The bottom part of the Fig. 4a shows that this as well as our other improvements (more training data comparing to Rohrbach et al. (2012a), color features, and hand detections) in the model each helps to improve performance. Overall, compared to PS, we achieve an improvement from 66.0 to 75.9 PCP and most notably an improvement from 48.9 to 74.4 and from 49.6 to 70.3 for lower arms, which are most important for recognizing hand-centric activities. We also would like to point to the benefit which hand detectors have to pose estimation (compare line 7 vs 8 and 9 vs 10).

Next we discuss the hand detection results. Our final hand detector *handDPM* is based on 32 components with 16 components allocated to each of the hands. The components are initialized by first grouping the training examples of each hand into 4 discrete orientations, and then clustering their HOG descriptors. In the experiments on hand localization we use a metric that reflects the localization accuracy and measures the percentage of hand hypotheses within a given distance from the ground truth. We visualize the results by plotting the localization accuracy for a range of distances.

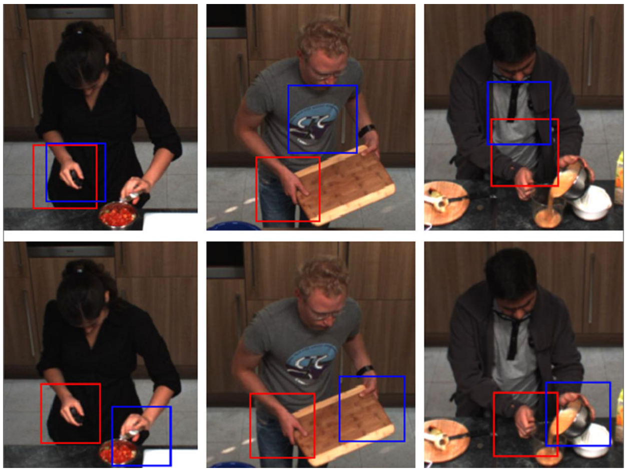


**Fig. 5** Pose helps to resolve failure cases of hand localization (upper row—handDPM, lower row is FPS + data + hand det + color) (Color figure online)

Figure 4b presents the evaluation of the localization accuracy of both hands. We observe that our hand detector (handDPM, red-dashed curve) alone already significantly improves over the proposed FPS approach (black-dotted-triangles). The performance further improves when hand detection hypotheses are integrated within the pose estimation model (blue-solid-stars). However, the improvement is moderate, likely because the pose estimation approach is not optimized specifically for hand detection and has to compromise between localization of hands and other body parts. Some qualitative examples are shown in Fig. 5.

We also compare our hand detector to a state-of-the-art hand detector of Mittal et al. (2011) using the code made publicly available by the authors. We perform the best-case evaluation and assign the hand hypothesis returned by the approach to the closest left and right hand in the ground-truth, as the hand detector does not differentiate between left and right hands. For a fair comparison we also filter the hand detections of Mittal et al. (2011) at irrelevant scales and image locations using body context as explained before. Our detector significantly improves over the hand detector of Mittal et al. (2011), which in addition to hand appearance also relies on color and context features, whereas our hand detector uses hand regions only. Note that there are significant differences between localization accuracy of left and right hands. We attribute this to the fact that the majority of people in our database are right handed. Since people perform many activities with their dominant hand, the pose of the right hand is more likely to be constrained by various activities due to the use of tools such as a knife or peeler. The left hand's pose is far less deterministic and the hand is often occluded behind the counter or while holding various objects.

## 5 Approaches for Fine-Grained Activity Recognition and Detection

In this section we focus on fine-grained activity recognition to approach the challenges typical e.g. for assisted daily living. Along with the activities we want to recognize their participating objects. To better understand the state-of-the-art for this challenging task we benchmark three types of approaches on our new dataset. The first type (Sect. 5.1) uses features derived from upper body model motivated by the intuition that human body configurations and human body motion should provide strong cues for activity recognition. For body pose estimation we rely on our approach described in Sects. 4.2 and 4.3. The second type (Sect. 5.2) are the state-of-the-art Dense Trajectories (Wang et al. 2013a) which have shown promising results on various datasets. It is a holistic approach in a sense that it extracts visual features on the entire frame. As the third type (Sect. 5.3) we present our hand-centric visual features, targeted at recognizing our hand-centric activities and the participating objects which are typically in the hand neighbourhood. For this we propose a hand detector (Sections 4.1, 4.3). Finally, we discuss our approaches to activity classification and detection in Sect. 5.4.

### 5.1 Pose-Based Approach

Pose-based activity recognition approaches were shown to be effective using inertial sensors (Zinnen et al. 2009). Inspired by Zinnen et al. (2009) we build on a similar feature set, computing it from the temporal sequence of 2D body configurations.

We employ a person detector (Felzenszwalb et al. 2010) and estimate the pose of the person within the detected region with 50 % border around. This allows us to reduce the complexity of the pose estimation and simplifies the search to a single scale. To extract the trajectories of body joints we rely 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. This approach combines the generic appearance model learned at training time with the specific appearance (SIFT) features computed at test time.

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

#### *5.1.1 Body Model Features (BM)*

For the BM features we compute the *velocity* of all joints (similar to gradient calculation in the image domain). We bin it in an 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. Similar to the joint trajectories (i.e. trajectories of x,y values) we build corresponding "trajectories" of distance values by stacking the values over temporally adjacent frames. 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.

#### *5.1.2 Fourier Transform Features (FFT)*

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.

#### *5.1.3 Feature Representation*

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 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 3336 for BM and 1536 for FFT.

## 5.2 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, 2013a) 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. (2013a) we generate a codebook for each of the four features of 4000 words using k-means from over a million sampled features.

## 5.3 Hand-Centric Approach

In domains where people mainly perform hand-related activities it seems intuitive to expect that hand regions contain important and relevant information for recognizing those activities and the participating objects. Thus, in addition to using the holistic and pose-based features, we suggest to focus on the hand regions. To obtain the hand locations we rely on our hand detector described in Sect. 4.1 as well as on the pose estimation method with integrated hand candidates (Sect. 4.3). In order to increase the robustness of the method we use both location candidates (provided by the handDPM detector and the final pose model) and sum the obtained features.

### *5.3.1 Hand-Trajectories*

We want to represent different type of information: hand motion, hand shape, and shape variations over time, as well as the appearance of objects manipulated by the hands. We propose to densely sample the neighborhood of each hand and to track those points over time. For tracking and also representing the point trajectories with powerful features we adapt the approach of Wang et al. (2013a). We focus only on densely sampled points around the estimated hand positions instead of sampling the entire video frame. We specify a bounding box around each hand detection and densely sample points inside of it. In our experiment we use 120 × 140 pixels bounding box around hands to include the information about the hands' context. We use 8 pixels grid spacing for points sampling and finally we get 136 interest point tracks for each frame. After extracting the features along computed tracks we create codebooks that contain 4000 words per feature.

### *5.3.2 Hand-cSift*

Color information is another important cue for recognizing activities and even more prominent for recognizing the participating objects. Similar to the previous approach we densely sample the points in the hands' neighborhood and extract color Sift features on 4 channels (RGB + grey). We quantize them in a codebook of size 4000.

## 5.4 Fine-Grained Activity Classification and Detection

### *5.4.1 Activity Classification*

Given a long video we assume that it consists of multiple time intervals. Each such interval $t$ depicts a single fine-grained activity and its participating objects (e.g. *dry, hands, towel*). In the following we refer to both, activities and participants, as activity attributes $a_i$, ($i \in \{1, \ldots, n\}$), i.e. $a_i$ can be any attribute including *cut*, *knife*, or *cucumber*. We train one-vs-all SVM classifiers on the features described in the previous sections given the ground truth intervals and labels. The classifiers provide us with real valued confidence score functions $f_i^{base} : \mathbb{R}^N \mapsto \mathbb{R}$ for attribute $a_i$ and feature vectors of dimension $N$. Combining different features is achieved by concatenating, i.e. stacking, the corresponding feature vectors.

### *5.4.2 Activity Detection*

While we use ground truth intervals for training the activity classifiers, we use a sliding window approach to find the correct interval of 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 min). Although this will still not cover all possible frame configurations, we found it to be a good trade-off between performance and computational costs.

## 6 Modeling Composite Activities

In the previous section we discussed how we recognize fine-grained activities (such as *peeling* or *washing*) and their object participants (such as *grater*, *knife*, or *cucumber*). Now we focus on exploiting the temporal context and on recognizing different composite activities, e.g. *preparing a cucumber* or *cooking pasta*.

For this, we first show how we exploit temporal context and co-occurrence to improve the recognition of fine-grained activities and their object participants (Sect. 6.1). Then, we model composite activities as a flexible combination of attributes, where attributes refer jointly to the fine-grained activities and their object participants (Sect. 6.2). We then show how to use prior knowledge (Sect. 6.3) to improve the recognition of composite activities, overcoming the notorious lack of training data and handling the large variability of composite activities. In Sect. 6.4 we discuss how to mine the semantic relatedness from script data. Finally, in Sect. 6.5 we introduce an automatic approach to temporal video segmentation, which removes the necessity to manually annotate the ground truth intervals in a video.

### 6.1 Recognizing Activity Attributes Using Context and Co-occurrence

For a time interval $t$ we want to classify if a particular fine-grained activity and its participants are present. We refer to activities and participants as activity attributes $a_i$. We distinguish three types of attribute classifiers. The first type of is given by the classifiers introduced in the previous section providing us with confidence score functions $f_i^{base} : \mathbb{R}^N \mapsto \mathbb{R}$ for each attribute $a_i$. Let us denote the score of a given feature vector $x_t$ at time interval $t$ as:

$$s_{i,t} = f_i^{base}(x_t). \tag{3}$$

Together these score constitute a matrix $S$ of dimensions $n \times T$ (# attributes × #timestamps). Based on these scores, we define features for context (in the same video sequence) as well as features for co-occurrence of other attributes (in the same time interval $t$).

Contextual features formalize the intuition that adjacent time frames have strongly related attributes: e.g. if a *cucumber* is *peeled* in one time interval, then *cutting* the *cucumber* is probably also present in the same video sequence. As visualized in Fig. 6a we define a context feature $g_t^{con} : \mathbb{R}^{n \times T} \mapsto \mathbb{R}^n$ at time $t$ by max pooling the scores of each attribute over all time intervals except $t$:

$$g_t^{con}(S) = \max_{u \in \{1,\ldots,T\} \setminus \{t\}} s_u \tag{4}$$

where max is an element-wise operator over all columns $s_u \in \mathbb{R}^n$ of matrix $S$.

Similarly, activity attributes happening at the same time interval $t$ 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 as a feature $g_i^{coocc} : \mathbb{R}^n \mapsto \mathbb{R}^{n-1}$ by stacking all attribute scores at time $t$ excluding $s_{i,t}$:

$$g_i^{coocc}(s_t) = [s_{1,t}; \ldots; s_{i-1,t}; s_{i+1,t}; \ldots; s_{n,t}], \tag{5}$$

where $s_t \in \mathbb{R}^n$ is a column of matrix $S$.

Based on these features we train activity attribute SVM classifiers using the features individually or by stacking them. Specifically we obtain corresponding confidence score functions for context: $f_i^{con} : \mathbb{R}^n \mapsto \mathbb{R}$ and co-occurrence: $f_i^{coocc} : \mathbb{R}^{n-1} \mapsto \mathbb{R}$, where $i$ denotes that a separate function for each attribute $a_i$ is trained. We define corresponding scores as:

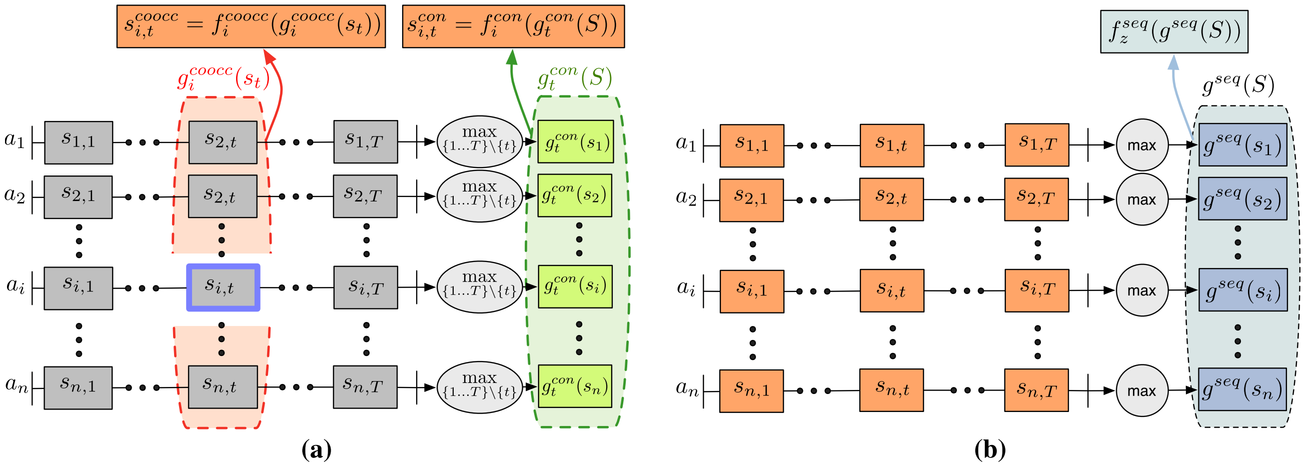


**Fig. 6** Our approach to recognition of attributes (**a**) and composite activities (**b**). **a** Activity attribute recognition using *con*textual and *co-occ*urrence attributes vectors. **b** Composite activity classification using max-pooled activity attributes (Color figure online)

$$s_{i,t}^{con} = f_i^{con}(g_t^{con}(S)) \tag{6}$$

and

$$s_{i,t}^{coocc} = f_i^{coocc}(g_i^{coocc}(s_t)). \tag{7}$$

This formulation can be easily extended to other attribute representations depending on the task and available features.

### 6.2 Composite Activity Classification Using Activity Attributes

We now want to classify composite activities that span an entire video sequence, given attribute classifier scores. We note that we can use any of the scores introduced in the previous section ($s_{i,t}, s_{i,t}^{con}, s_{i,t}^{coocc}$ or their stacked combination). In the following for simplicity we refer to these scores as $s_{i,t}$ and corresponding matrix as $S$. In this approach we rely on the representation that captures likelihoods of the presence or absence of a particular attribute and leave modeling the temporal ordering of attributes for future work. We define a feature for the video sequence as $g^{seq} : \mathbb{R}^{n \times T} \mapsto \mathbb{R}^n$ by max pooling the scores of each attribute over all time intervals (see Fig. 6b):

$$g^{seq}(S) = \max_{t \in \{1,\ldots,T\}} s_t \tag{8}$$

where max is an element-wise operator over all columns $s_t \in \mathbb{R}^n$ of matrix $S$.

To decide on the class $z$ of a sequence $d$ we use the feature $g^{seq}$ and classify it using a nearest neighbor classifier (*NN*) or a one-versus-all SVM given a set of labeled training sequences. The SVM classifier provides us with the following confidence function for all composite classes $z$: $f_z^{seq} : \mathbb{R}^n \mapsto \mathbb{R}$, where the final score is defined as:

$$s_{z,d}^{seq} = f_z^{seq}(g^{seq}(S_d)), \tag{9}$$

where $S_d$ is the score matrix for sequence $d$. The following sections describe alternatives to NN and SVM to incorporate prior knowledge mined from script data.

### 6.3 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 weights $w_{z,i}$, with $Z$ being the number of composite activity classes. The vectors $w_z$ are L1 normalized, i.e. $\sum_{i=1}^{n} w_{z,i} = 1$. Our system extracts those associations from script data (see Sect. 6.4), but the approach generalizes to other arbitrary external knowledge sources. We explore three options to use such information which we detail in the following.

#### 6.3.1 *Script data*

We compute the confidence $f_z^{scriptdata} : \mathbb{R}^n \mapsto \mathbb{R}$ of a sequence being of the composite activity $z$ using the attribute-based feature representation $g^{seq}(S)$ introduced in Eq. (8). Given the weights $w_{z,i}$ we compute a weighted sum:

$$f_z^{scriptdata}(g^{seq}(S)) = \sum_{i=1}^{n} w_{z,i} g_i^{seq}(S). \tag{10}$$

For a specific sequence $d$ with corresponding score matrix $S_d$ we get the following score:

$$s_{z,d}^{scriptdata} = f_z^{scriptdata}(g^{seq}(S_d)). \tag{11}$$

This formulation is similar to the sum formulation we used in Rohrbach et al. (2011) for image recognition with attributes, which itself is an adaption of the direct attribute prediction model introduced by Lampert et al. (2013). Note that the weight matrix retrieved from script data is sparse (most $w_{z,i} = 0$). When mining from other corpora one might need to threshold the weights $w_{z,i}$, setting all others to zero, to achieve good performance as done e.g. in Rohrbach et al. (2011).

#### 6.3.2 *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 L2 norm for computing the distance of nearest neighbor with the following training class dependent weighted L2 norm. It takes weights of class-attribute associations into account. It is defined between the test attribute vector of unseen class $g^{seq}(S_{test})$ and the training attribute vector $g^{seq}(S_{train}^z)$ of class $z$ as:

$$Dist(S_{test}, S_{train}^z) = \left( \sum_{i=1}^{n} w_{z,i} \left( g_i^{seq}(S_{test}) - g_i^{seq}(S_{train}^z) \right)^2 \right)^{0.5}. \tag{12}$$

To enhance robustness further, we binarize all association weights $w_{z,i}$ by setting all non-zero weights to 1 (and L1-normalize $w_z$). This reduces the distance computation to the relevant attributes, normalized by the total number of relevant attributes.

### 6.3.3 Propagated Semantic Transfer (PST)

As the third approach to integrate external knowledge from script data we use Propagated semantic transfer (PST) which we proposed in Rohrbach et al. (2013a) and summarize shortly in the following. The approach builds on Eq. (10) and uses label propagation to exploit the distances within the unlabeled data, i.e. it assumes a transductive setting where all test data is available when predicting a single test label.

We can incorporate (partially) labeled training data $l_{z,d} \in \{0, 1, \emptyset\}$ for class $z$ and sequence $d$. $\emptyset$ denotes that we do not have a label for this sequence and class. We combine the labels with the predictions in the following way, using only the most reliable predictions $s^{scriptdata}_{z,d}$ (top-$\delta$ fraction) per class $z$:

$$s^{PST}_{z,d} = \begin{cases} \gamma l_{z,d} & \text{if } l_{z,d} \in \{0,1\} \\ (1-\gamma) s^{scriptdata}_{z,d} & \text{if among top-}\delta\text{ fraction} \\ & \text{of predictions for class } z \\ 0 & \text{otherwise.} \end{cases} \tag{13}$$

$\gamma$ provides a weighting between the true labels and the predicted labels. In the zero-shot case we only use predictions and $\gamma = 0$. The parameters $\delta, \gamma \in [0, 1]$ are chosen, similar to the remaining parameters, on the validation set. For zero-shot we use the unlabeled training data as additional data for label propagation.

For computing the distance between the sequences we use the feature representation $g^{seq}(S)$, as for the *NN*-classifier, which is much lower dimensional than the raw video feature representation and provides more reliable distances as we showed in Rohrbach et al. (2013a). We build a k-NN graph by connecting the k closest neighbours. We set the weights of the graph edges between sequences $d$ and $e$ to $exp(-0.5\sigma^{0.5}\|g^{seq}(S_d) - g^{seq}(S_e)\|)$, where $\sigma$ is set to the mean of the distances to the nearest neighbours. We initialize this graph with the scores $s^{PST}_{z,d}$ and propagate them using label propagation from Zhou et al. (2004).

## 6.4 Prior Knowledge from Script Data

We want to quantify what activities and objects typically occur in a composite activity by leveraging the script data we collected (see Sect. 3.4). In order to use prior knowledge from textual script data, we have to match the (controlled) attribute labels from the video annotations to the (freely) written script instances (Sect. 6.4.1). Based on the matched attributes we compute two different word frequency statistics (Sect. 6.4.2).

### 6.4.1 Label Matching

To transfer any kind of knowledge from the script corpus to the attributes in the video annotation, we need to match attribute labels to natural language descriptions. The annotated attribute labels are standard English verbs (for activities, *wash*) and nouns (for participating objects, *carrot*), sometimes with additional particles (*take apart* and *take out*). As the script instances contain freely written natural language sentences, they do not necessarily have any correspondence with the attribute label annotations. We compare two strategies for mapping annotations to script data sentences:

- **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 Statistics Computed on the Script Data

We compute two different association scores between attribute labels $a_i$ and composite activities $z$. For this we concatenate all scripts for a given composite $z$ to a single document $\delta_z$.

- **freq**: word frequency $freq(a_i, \delta_z)$ for each attribute $a_i$ and composite activities $z$.
- **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 $D = \{\delta_1, ..., \delta_z, ..., \delta_m\}$, tf$*$idf for a term or attribute $a_i$ and a document $\delta_z$ is computed as follows:

$$tfidf(a_i, \delta_z) = freq(a_i, \delta_z) * log \frac{|D|}{|\{\delta \in D : a_i \in \delta\}|}, \tag{14}$$

where $\{\delta \in D : a_i \in \delta\}$ is the set of documents containing $a_i$ 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.

We set $w_{z,i} = freq(a_i, \delta_z)$ or $w_{z,i} = tfidf(a_i, \delta_z)$ and L1-normalize all vectors $w_z$. These weights $w_{z,i}$ are then used in Equations (10) and (12) and subsequently also in our PST approach.

### 6.5 Automatic Temporal Segmentation

While we assume a segmented video during training time to learn attribute classifiers as described in Sect. 5.4, we want to segment the video automatically at test time. To avoid noisy and small segments we follow the idea we presented in (Rohrbach et al. 2014), namely we employ agglomerative clustering. We start with uniform intervals of 60 frames and describe each interval with an attribute-classifier score vector. We combine neighbouring intervals based on the cosine similarity of their score vectors and stop when we reach a threshold (found on the validation set). We aim for a segmentation with granularity similar to original manual annotation. After this a separately trained visual background classifier removes irrelevant or noisy segments. In our experiments we show that this leads to composite recognition results, similar to using the ground truth intervals for the attributes.

## 7 Evaluation

In this section we evaluate our approaches to fine-grained and composite activity recognition. We start with the fine-grained activity classification and detection and compare three types of approaches described in Sect. 5, namely pose-based, hand-centric and holistic approaches. Next we evaluate our approaches for composite activity recognition introduced in Sect. 6, evaluating our attributes enhanced with context and co-occurrence, the recognition of composite cooking activities using different levels of supervision, and the zero-shot approach using script data.

### 7.1 Experimental Setup

This section details our experimental setup. We will release evaluation code to reproduce and compare with our results. See Table 3 for the information on our training/validation/test split. We estimate all hyper parameters on the validation set and then retrain the models on the training and validation set with the best parameters.

#### *7.1.1 Experimental Setup Fine-Grained Activity Classification and Detection*

In the fine-grained recognition task we want to distinguish 67 fine-grained activities and 155 participating objects (see Table 7 for the lists of activities and objects). To learn the visual classifiers we use the annotated ground truth intervals provided with the dataset. We train one-vs-all SVMs using mean SGD (Rohrbach et al. 2011) with a $\chi^2$ kernel approximation (Vedaldi and Zisserman 2010). 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 ground-truth. If a second detection fires for one ground-truth label, it is counted as false positive. In the following we report the mean over the average precision (AP) of each class. Combining features is achieved by stacking the bag-of-word histograms.

#### *7.1.2 Experimental Setup Composite Activity Recognition*

For localizing attributes within composite activities we rely on our automatic segmentation (Sect. 6.5). We aim to recognize 31 composite activities (see bold names in Table 2).

We distinguish two cases for training the attributes with respect to composites.

*Attribute training on all composites.* We use all available 218 training + validation videos for training the attribute classifiers. See left half of Tables 8, 9, and 10.

*Attribute training on disjoint composites.* We use all available videos apart from those showing the test composite categories (in total 92 videos). This means that attributes and composites are trained on disjoint sets of composite categories and thus also on disjoint sets of videos. This tests how well novel composite categories can be recognized without additional attribute labels. See right half of Tables 8, 9, and 10.

Next, we have two cases for training the composites.

*With training data for composites.* We train on the 126 training + validation videos whose category is in the set of the 31 test categories. Note that in case of *Attribute training on all composites* the training videos are also part of the attribute training. See top part of Table 9.

*No training data for composites.* Here we do not rely on any training labels for the composite activities. See bottom part of Table 9 and all of Table 10. Combined with *Attribute training on disjoint composites* this is zero-shot recognition.

### 7.2 Fine-Grained Activity Classification and Detection

#### *7.2.1 Activity Classification*

We start with the classification results on fine-grained activities and their participants (Table 5).

The body model features on the joint tracks (BM) achieve a mean average precision (AP) of 18.9 % for activities and 13.8 % for objects. Comparing this to the FFT features, we

**Table 5** Fine-grained activity and object classification results, mean AP in % (see Sect. 7.2 for discussion)

| Approach | Activities | Objects | All |
|---|---|---|---|
| *Pose-based approaches* | | | |
| (1) BM | 18.9 | 13.8 | 15.7 |
| (2) FFT | 19.0 | 16.2 | 17.2 |
| (3) Combined | 24.1 | 19.0 | 20.8 |
| *Hand-centric approaches* | | | |
| (4) Hand-cSift | 23.0 | 23.8 | 23.5 |
| (5) Hand-trajectories | 45.1 | 31.5 | 36.4 |
| (6) Combined | 43.5 | 34.2 | 37.5 |
| *Holistic approach* | | | |
| (7) Dense trajectories | 44.5 | 31.3 | 36.1 |
| *Combinations* | | | |
| (8) Dense Traj,BM,FFT | 43.1 | 30.7 | 35.2 |
| (9) Dense Traj,Hand-Traj | 52.2 | 37.7 | 42.9 |
| (10) Dense Traj,Hand-Traj,-cSift | 51.2 | 39.3 | 43.7 |

observe that FFT performs slightly better, improving over BM the AP by 0.1 and 2.4 % respectively. The combination of BM and FFT features (line 3 in Table 5) yields a significant improvement, reaching AP of 24.1 % for activities and 19.0 % for objects. We attribute this to the complementary information encoded in the features. While BM encodes among others velocity-histograms of the joint-tracks and statistics between tracks of different joints, FFT features encode FFT coefficients of individual joints. Still, this is a relatively low performance. It can be explained, on one hand, by failures of the pose estimation method and, on the other hand, the pose-based features might not contain enough information to successfully distinguish the challenging fine-grained activities and participating objects. Next we look at the performance of our proposed hand-centric features. Color Sift features, densely sampled in the hand neighborhood, allow us to improve the object recognition AP to 23.8 % (Hand-cSift), indicating their better suitability in particular for recognizing objects. Dense Trajectories features computed around hands (denoted as Hand-Trajectories) reach 45.1 and 31.5 % recognition AP for activities and objects, respectively. Combining both features leads to a small disimprovement for activities, however it helps to further improve the object recognition performance to 34.2 %. Overall our hand-centric approach reaches the recognition AP of 37.5 % for activities and objects together. The state-of-the-art holistic approach of Dense Trajectories (Wang et al. 2013a) obtains 44.5 and 31.3 % recognition AP for activities and objects. If compared to our hand-centric features, this is slightly below the Hand-Trajectories, which are restricted to the areas around hands. This supports our hypothesis that the most relevant information for recognizing our fine-grained activities is contained in the hand regions. We also consider several feature combinations (lines 8, 9, 10 in Table 5). Combining Dense Trajectories with the pose-based features does not improve the recognition performance. However, combining them with Hand-Trajectories improves the activity recognition by 7.7 % and object recognition by 6.4 % (line 7 vs 9 in Table 5). Finally, adding the Hand-cSift features allows to reach the impressive 43.7 % recognition AP for activities and objects together.

The detailed comparison of Dense Trajectories, Hand-Trajectories and the final feature-combination (line 10 in Table 5) can be found in Table 7. Hand-Trajectories loose to Dense Trajectories on activities that include "coarser" motion, e.g. *push down*, *hang* or *plug*, and corresponding objects such as *hook* or *teapot*. Note that Hand-Trajectories outperform the Dense Trajectories for 35 activity classes, while in the opposite direction this holds only 25 times (for objects, respectively 65 vs 43 times). This shows again that the hand-centric features consistently outperform the holistic features in both tasks. Some example cases where the hand-centric approach is significantly better, are such activities as *rip open*, *take apart*, and *grate* and such objects as *cauliflower*, *oven*, and *cup*. At the same time the final feature combination (line 10 in Table 5) consistently outperforms both aforementioned features in about 60 % of cases. We demonstrate some qualitative results comparing Dense Trajectories to the final feature combination in Table 11. We also looked closer at the performance of other features. e.g. the combined pose features (line 3 in Table 5) perform well on "coarser", full-body activities, such as *throw in garbage*, *take out*, *move*, while rather poorly on more fine-grained activities. On the other hand the Hand-cSift features are good in recognizing objects with distinct shapes/colors, e.g. *pineapple*, *carrot*, *bowl*, etc.

### *7.2.2 Activity Detection*

Next we look at the detection performance (Table 6), which is inherently more challenging than the classification task. Here the BM features reach 8.3 % overall AP and FFT get 9.3 %. Their combination (line 3 in Table 6) gets 11.4 % overall AP, while Hand-cSift only reaches 10.7 %. Hand-Trajectories alone get 16.6 % AP and combined with Hand-cSift they reach 22.5 %, while the Dense Trajectories get 24.4 % AP. As we can see for this task our hand-centric features perform worse than holistic and even pose-based features (line 3 vs 4 in Table 6). We believe the reason for this is that for correct segmentation of the video into activity intervals we need more holistic information, which the hand-centric features cannot provide, while pose-based and holistic features can capture it better. Similarly, when combining Dense Trajectories with the pose-based features (line 8 in Table 6) we observe a small improvement, supporting our hypothesis that pose indeed helps to capture the detection boundaries. On the

**Table 6** Fine-grained activity and object detection results, mean AP in % (see Sect. 7.2 for discussion)

| Approach | Activities | Objects | All |
|---|---|---|---|
| *Pose-based approaches* | | | |
| (1) BM | 9.7 | 7.6 | 8.3 |
| (2) FFT | 10.5 | 8.7 | 9.3 |
| (3) Combined | 14.3 | 9.8 | 11.4 |
| *Hand-centric approaches* | | | |
| (4) Hand-cSift | 10.5 | 10.9 | 10.7 |
| (5) Hand-trajectories | 21.3 | 14.0 | 16.6 |
| (6) Combined | 26.0 | 20.6 | 22.5 |
| *Holistic approach* | | | |
| (7) Dense trajectories | 29.5 | 21.5 | 24.4 |
| *Combinations* | | | |
| (8) Dense Traj,BM,FFT | 30.7 | 21.5 | 24.8 |
| (9) Dense Traj,Hand-Traj | 34.3 | 25.2 | 28.5 |
| (10) Dense Traj,Hand-Traj,-cSift | 34.5 | 25.3 | 28.6 |

other hand, combining Dense Trajectories with our hand-centric features significantly improves the performance, in particular by 4.7 % for activities and by 3.7 % for objects (line 6 vs 9 in Table 6). Combining the obtained features with the Hand-cSift further improves the results and we reach the 28.6 % overall AP. The improvement obtained after combining holistic and hand-centric features can be explained by the increased classification AP within the obtained intervals. We thus conclude that for activity detection we require holistic information, which can come e.g. from the human pose. Combining the holistic and hand-centric features is still beneficial and significantly improves the performance.

### 7.3 Context and Co-occurrence for Fine-Grained Activities

While so far we looked at individual fine-grained activities, we now evaluate the benefit from co-occurrence and context as introduced in Sect. 6.1. Table 8 provides the results for recognizing activities and their participants, modeled as attributes. We evaluate in two settings. The left two columns of Table 8 show the results for training on all composites in training set, while the right two columns are trained only on composites absent in test set (Disjoint Composites), i.e. the second is a more challenging problem, as there is less training data and the attributes are tested in a different context (Table 7) . The performance in the first line is equivalent to the results in Table 5. The very left column shows results on Dense Trajectories. More specifically using only temporal context to recognize activity attributes performance drops from 36.1 % AP for the base classifier to 11.1 % 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 (line 4 in Table 8), the performance increases by 2.0 % 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 37.8 and 38.1 %, respectively. A combination of all training modes achieves a performance of 39.3 % AP, improving the base classifier's result by 3.2 %. While results for Dense Trajectories are as expected i.e. adding context and co-occurrence improves performance, the performance drops slightly for the (in general) better performing combined features (second column). However, although the attribute prediction performance drops, we found that for recognizing the composites, context and co-occurrence are still useful.

In the second setting, we restrict the training dataset to composites absent in the test set (right two columns of Table 8), requiring the activity attributes to transfer to different composite activities. When comparing the right two the left columns, we notice a significant performance drop for all classifiers and both features. This decrease can mainly be attributed to the strong reduction of training data to about one third. The base classifier performs best and co-occurrence variants slightly below. Variants including context lead to tremendous performance drops in all combinations because the activity context changes from training to test (having different composite activities).

### 7.4 Composite Cooking Activity Classification

After evaluating attribute recognition performance in Sect. 7.3, we now show the results for recognizing composites as introduced in Sect. 6.2. From the different attribute combination variants we only use the combination of base, context, and co-occurrence (last line in Table 8). Although this is not always the best choice for recognizing attributes we found it to work better or similar to alternatives for composite recognition. The results are shown in Table 9, which, similar to Table 8, shows results for training the attributes on all composites, on the left, and reduced attribute training on non-test composites on the right. In the top section of the table we use training data for the composite cooking activities. In the bottom section of the table we use *no* training data for the composite cooking activities. This is enabled by the use of script data as motivated before. Disregarding the first line which does not use attributes at all and the second line which uses ground truth intervals for attributes, all other lines are based on attributes computed on our automatic temporal segmentation, introduced in Sect. 6.5.

Examining the results in Table 9 we make several interesting observations. First, training composites on attributes of fine-grained activities and objects (line 3 in Table 9) outper-

**Table 7** Fine-grained activities and object classification performance of Dense Trajectories, Hand Trajectories, and their combination including Hand-cSift (line 10 in Table 5) for 67 fine-grained activities and 155 participating objects. AP in %. "-" denotes that the category is not part of the test set and not evaluated

| Activity | Dense Traj | Hand Traj | Combi +cSift | Object | Dense Traj | Hand Traj | Combi +cSift | Object | Dense Traj | Hand Traj | Combi +cSift |
|---|---|---|---|---|---|---|---|---|---|---|---|
| Add | 19.8 | 16.3 | 24.0 | Apple | – | – | – | Mango | 3.8 | 7.0 | 2.5 |
| Arrange | 61.9 | 32.1 | 33.8 | Arils | 19.8 | 57.8 | 12.5 | Masher | – | – | – |
| Change temperature | 69.1 | 78.1 | 75.4 | Asparagus | – | – | – | Measuring-pitcher | 0.7 | 5.0 | 5.3 |
| Chop | 36.6 | 35.4 | 48.3 | Avocado | 2.5 | 4.3 | 3.8 | Measuring-spoon | 34.1 | 12.6 | 7.3 |
| Clean | 32.0 | 33.0 | 33.3 | Bag | – | – | – | milk | 0.4 | 0.4 | 0.4 |
| Close | 76.3 | 68.8 | 77.0 | Baking-paper | – | – | – | Mortar | – | – | – |
| Cut apart | 33.8 | 36.2 | 33.5 | Baking-tray | – | – | – | Mushroom | – | – | – |
| Cut dice | 39.3 | 45.7 | 44.9 | Blender | – | – | – | Net-bag | 0.3 | 0.2 | 0.7 |
| Cut off ends | 21.4 | 52.0 | 31.9 | Bottle | 57.1 | 49.3 | 57.7 | Oil | 52.3 | 47.6 | 55.6 |
| Cut out inside | 2.2 | 0.8 | 2.0 | Bowl | 34.7 | 33.1 | 49.0 | Onion | 19.3 | 20.4 | 22.7 |
| Cut stripes | 12.9 | 13.0 | 15.4 | Box-grater | – | – | – | Orange | 18.4 | 11.1 | 19.3 |
| Cut | 28.3 | 44.9 | 27.2 | Bread | 3.7 | 6.5 | 8.9 | Oregano | – | – | – |
| Dry | 81.9 | 85.1 | 84.5 | Bread-knife | 3.0 | 4.0 | 8.1 | Oven | 30.7 | 73.4 | 89.3 |
| Enter | 100.0 | 100.0 | 100.0 | Broccoli | 2.0 | 2.3 | 5.7 | Paper | – | – | – |
| Fill | 94.3 | 90.8 | 86.2 | Bun | 1.2 | 2.3 | 8.5 | Paper-bag | 20.5 | 10.3 | 33.0 |
| Gather | 25.7 | 23.8 | 35.7 | Bundle | 0.5 | 1.1 | 1.4 | Paper-box | 1.0 | 1.2 | 3.6 |
| Grate | 66.7 | 100.0 | 100.0 | Butter | 6.2 | 1.9 | 9.6 | Parsley | 23.4 | 25.5 | 49.6 |
| Hang | 85.8 | 57.2 | 81.4 | Carafe | 44.4 | 46.7 | 54.4 | Pasta | 26.1 | 16.0 | 40.7 |
| Mix | 10.3 | 5.4 | 52.9 | Carrot | 26.5 | 41.3 | 64.9 | Peach | – | – | – |
| Move | 75.7 | 75.7 | 78.3 | Cauliflower | 29.3 | 68.9 | 73.8 | Pear | – | – | – |
| Open close | 60.8 | 65.7 | 64.7 | Cheese | – | – | – | Peel | 40.3 | 28.6 | 35.2 |
| Open egg | 50.0 | 28.1 | 39.2 | Chefs-knife | 59.9 | 73.3 | 63.1 | Pepper | 3.1 | 14.4 | 6.7 |
| Open tin | – | – | – | Chili | 0.6 | 0.9 | 1.3 | Peppercorn | – | – | – |
| Open | 22.0 | 22.0 | 34.5 | Chive | – | – | – | Pestle | – | – | – |
| Package | 0.4 | 1.6 | 1.8 | Chocolate | – | – | – | Philadelphia | – | – | – |
| Peel | 55.0 | 67.2 | 58.6 | Coffee | 3.3 | 25.0 | 100.0 | Pineapple | 19.5 | 47.0 | 49.7 |
| Plug | 41.6 | 32.6 | 81.0 | Coffee-container | 34.6 | 24.8 | 73.4 | Plastic-bag | 36.4 | 37.7 | 43.6 |
| Pour | 44.8 | 44.9 | 45.1 | Coffee-machine | 34.7 | 65.1 | 91.2 | Plastic-bottle | 4.7 | 2.8 | 9.1 |
| Pull apart | 38.7 | 53.8 | 45.2 | Coffee-powder | 0.5 | 1.3 | 3.0 | Plastic-box | 2.6 | 9.0 | 5.3 |
| Pull up | 79.2 | 21.7 | 75.6 | Colander | 63.4 | 62.2 | 77.9 | Plastic-paper-bag | 0.9 | 14.7 | 19.6 |
| Pull | 1.3 | 9.1 | 1.2 | Cooking-spoon | – | – | – | Plate | 65.7 | 69.2 | 73.9 |
| Puree | – | – | – | Corn | – | – | – | Plum | 0.7 | 2.5 | 1.3 |
| Purge | 0.1 | 0.1 | 0.6 | Counter | 71.8 | 70.3 | 76.5 | Pomegranate | 5.1 | 0.8 | 2.3 |
| Push down | 30.7 | 7.6 | 28.0 | Cream | 0.9 | 0.5 | 1.4 | Pot | 84.3 | 88.0 | 91.1 |
| Put in | 55.5 | 50.8 | 58.0 | Cucumber | 4.3 | 5.2 | 4.1 | Potato | 0.4 | 0.4 | 0.6 |
| Put lid | 87.3 | 85.3 | 90.0 | Cup | 27.0 | 26.7 | 43.6 | Puree | – | – | – |
| Put on | 6.2 | 5.6 | 1.2 | Cupboard | 97.5 | 98.0 | 98.4 | Raspberries | – | – | – |
| Read | 5.1 | 5.4 | 5.6 | Cutting-board | 84.4 | 85.4 | 88.9 | Salad | – | – | – |
| Remove from package | 19.3 | 34.3 | 31.5 | Dough | – | – | – | Salami | – | – | – |
| Rip open | 2.8 | 45.0 | 100.0 | Drawer | 98.2 | 98.4 | 98.5 | Salt | 59.8 | 48.7 | 64.1 |
| Scratch off | 30.7 | 33.1 | 31.9 | Egg | 12.1 | 3.6 | 7.3 | Seed | – | – | – |
| Screw close | 77.3 | 77.5 | 77.5 | Eggshell | 3.5 | 3.6 | 11.2 | Side-peeler | 50.0 | 11.7 | 37.8 |
| Screw open | 78.7 | 69.4 | 79.2 | Electricity-column | 89.3 | 82.3 | 98.1 | Sink | 47.0 | 54.0 | 53.9 |
| Shake | 73.0 | 75.7 | 77.3 | Electricity-plug | 74.3 | 70.6 | 87.7 | Soup | – | – | – |

**Table 7** continued

| Activity | Dense | Hand | Combi | Object | Dense | Hand | Combi | Object | Dense | Hand | Combi |
|---|---|---|---|---|---|---|---|---|---|---|---|
| | Traj | Traj | +cSift | | Traj | Traj | +cSift | | Traj | Traj | +cSift |
| Shape | – | – | – | Fig | 1.0 | 1.0 | 0.9 | Spatula | 72.9 | 76.2 | 78.2 |
| Slice | 47.2 | 71.3 | 57.4 | Filter-basket | 1.3 | 3.4 | 13.1 | Spice | 19.1 | 13.3 | 12.4 |
| Smell | 49.7 | 15.7 | 33.0 | Finger | 18.4 | 15.4 | 8.8 | Spice-holder | 95.6 | 94.4 | 96.3 |
| Spice | 88.6 | 89.0 | 89.2 | Flat-grater | 31.7 | 27.7 | 40.9 | Spice-shaker | 88.3 | 87.3 | 91.5 |
| Spread | 87.1 | 77.1 | 96.7 | Flower-pot | – | – | – | Spinach | – | – | – |
| Squeeze | 90.1 | 92.9 | 91.9 | Food | – | – | – | Sponge | 17.2 | 45.4 | 38.2 |
| Stamp | – | – | – | Fork | 8.7 | 7.5 | 10.5 | Sponge-cloth | 67.1 | 68.1 | 75.0 |
| Stir | 91.2 | 81.9 | 91.7 | Fridge | 100.0 | 99.8 | 100.0 | Spoon | 2.8 | 5.9 | 8.9 |
| Strew | 1.7 | 2.4 | 2.4 | Front-peeler | 21.8 | 6.0 | 17.6 | Squeezer | 52.5 | 67.0 | 59.3 |
| Take apart | 1.6 | 32.1 | 53.3 | Frying-pan | 88.7 | 91.9 | 93.6 | Stone | 0.2 | 0.7 | 0.7 |
| Take lid | 66.2 | 76.8 | 71.7 | Garbage | 13.7 | 17.9 | 27.5 | Stove | 84.4 | 87.2 | 90.4 |
| Take out | 94.1 | 93.9 | 95.1 | Garlic-bulb | 0.3 | 0.6 | 0.8 | Sugar | 22.0 | 24.2 | 29.0 |
| Tap | 3.3 | 4.2 | 6.2 | Garlic-clove | 11.7 | 3.6 | 9.3 | Table-knife | – | – | – |
| Taste | 9.4 | 21.0 | 22.0 | Ginger | 1.9 | 3.3 | 3.6 | Tap | 70.2 | 71.8 | 79.1 |
| Test temperature | 11.3 | 11.8 | 35.1 | Glass | 2.6 | 4.5 | 21.6 | Tea-egg | 37.2 | 28.7 | 36.1 |
| Throw in garbage | 96.7 | 96.0 | 97.1 | Green-beans | 21.1 | 24.6 | 23.2 | Tea-herbs | 60.5 | 55.6 | 91.1 |
| Turn off | 7.4 | 21.1 | 33.0 | Ham | – | – | – | Teapot | 46.4 | 6.7 | 69.1 |
| Turn on | 27.8 | 30.6 | 48.5 | Hand | 95.9 | 95.2 | 96.4 | Teaspoon | 29.2 | 32.4 | 36.5 |
| Turn over | – | – | – | Handle | 100.0 | 9.1 | 100.0 | Tin | – | – | – |
| Unplug | 8.7 | 3.8 | 20.0 | Hook | 95.6 | 71.2 | 98.3 | Tin-opener | – | – | – |
| Wash | 93.4 | 93.9 | 93.7 | Hot-chocolate-powder-bag | – | – | – | Tissue | – | – | – |
| Whip | – | – | – | Hot-dog | 2.1 | 2.7 | 8.8 | Toaster | 1.3 | 8.1 | 6.7 |
| Wring out | 3.3 | 4.5 | 5.3 | Jar | 5.4 | 14.2 | 17.8 | Tomato | – | – | – |
| | | | | Ketchup | 2.0 | 3.1 | 19.6 | Tongs | – | – | – |
| | | | | Kettle-power-base | 14.4 | 9.8 | 41.4 | Top | – | – | – |
| | | | | Kiwi | 1.1 | 2.9 | 1.5 | Towel | 73.2 | 76.9 | 79.2 |
| | | | | Knife | 69.6 | 83.5 | 76.8 | Tube | 1.0 | 9.5 | 10.2 |
| | | | | Knife-sharpener | – | – | – | Water | 55.0 | 46.9 | 57.2 |
| | | | | Kohlrabi | – | – | – | Water-kettle | 40.7 | 25.9 | 53.7 |
| | | | | Ladle | – | – | – | Wire-whisk | – | – | – |
| | | | | Leek | 10.6 | 19.5 | 17.6 | Wrapping-paper | 2.9 | 0.4 | 2.0 |
| | | | | Lemon | – | – | – | Yolk | 0.5 | 0.5 | 0.3 |
| | | | | Lid | 67.1 | 70.8 | 71.8 | Zucchini | – | – | – |
| | | | | Lime | 14.2 | 3.7 | 14.6 | | | | |

forms low-level features (line 1 in Table 9), supporting our claim that for learning composite activities it is important to share information on an intermediate level of attributes.

The second somewhat surprising observation is that recognizing composites based on our segmentation (line 3 in Table 9) outperforms using ground truth segments (line 2 in Table 9). We attribute this to the fact that our segmentation is coarser than the ground truth and that we additionally remove noisy and background segments with a background classifier. This leads to more robust attributes and consequently better composite recognition. This allows to have separate training sets for composites and attributes. This setting is explored in the top right quarter of Table 9. Here the training sequences for attributes are disjoint with the ones for composites, i.e. we do not require the attribute annotataions for the composite training set.

**Table 8** Attribute recognition using context and co-occurrence, mean AP in %. Combi+cSift refers to Dense Traj,Hand-Traj,-cSift, see Sect. 7.3 for discussion

| Attribute training on: | All composites | | Disjoint composites | |
|---|---|---|---|---|
| | Dense Traj | Combi +cSift | Dense Traj | Combi +cSift |
| (1) Base ($s^{base}$) | 36.1 | 43.7 | 33.5 | 35.9 |
| (2) Context only ($s^{con}$) | 11.1 | 12.6 | 6.8 | 8.1 |
| (3) Base + Context | 37.8 | 41.2 | 28.3 | 32.3 |
| (4) Co-occ. only ($s^{coocc}$) | 38.1 | 41.7 | 32.6 | 35.3 |
| (5) Base + Co-occ. | 38.1 | 41.4 | 32.7 | 35.2 |
| (6) Base + Cont. + Co-occ | 39.3 | 41.5 | 30.8 | 32.6 |

**Table 9** Composite cooking activity classification, mean AP in %. Top left quarter: fully supervised, right column: reduced attribute training data, bottom section: no composite cooking activity training data, right bottom quarter: true zero shot. See Sect. 7.4 for discussion

| Attribute training on: | All composites | | Disjoint composites | |
|---|---|---|---|---|
| | Dense Traj | Combi +cSift | Dense Traj | Combi +cSift |
| **With training data for composites** | | | | |
| *Without attributes* | | | | |
| (1) SVM | 39.8 | 41.1 | - | - |
| *Attributes on gt intervals* | | | | |
| (2) SVM | 43.6 | 52.3 | 32.3 | 34.9 |
| *Attributes on automatic segmentation* | | | | |
| (3) SVM | 49.0 | 56.9 | 35.7 | 34.8 |
| (4) NN | 42.1 | 43.3 | 24.7 | 32.7 |
| (5) NN + Script data | 35.0 | 40.4 | 18.0 | 21.9 |
| (6) PST + Script data | 54.5 | 57.4 | 32.2 | 32.5 |
| **No training data for composites** | | | | |
| *Attributes on automatic segmentation* | | | | |
| (7) Script data | 36.7 | 29.9 | 19.6 | 21.9 |
| (8) PST + Script data | 36.6 | 43.8 | 21.1 | 19.3 |

**Table 10** Variants of script knowledge, AP in %. Combi+cSift refers to Dense Traj,Hand-Traj,-cSift. See Sect. 7.4 for discussion

| Attribute training on: | All composites | | Disjoint composites | |
|---|---|---|---|---|
| | Dense Traj | Combi +cSift | Dense Traj | Combi +cSift |
| **No training data for composites** | | | | |
| Script data | | | | |
| (1) freq-literal | 28.2 | 30.5 | 19.8 | 24.1 |
| (2) freq-WN | 25.3 | 28.6 | 17.4 | 20.3 |
| (3) tf∗idf-literal | 35.9 | 31.8 | 20.0 | 23.6 |
| (4) tf∗idf-WN | 36.7 | 29.9 | 19.6 | 21.9 |

Third, the improvements we achieved for fine-grained activities and object recognition by combining hand-centric with holistic features are still evident for composites. The Combination of Dense Trajectoreis, Hand-Trajectories, and Hand-cSift (2nd, 4th column) outperforms in most cases Dense Trajectories only (1st, 3rd column), most notably in the setting “All Composites” for SVM (56.9 % over 49.0 % AP) and PST + Script data (43.8 % over 36.6 % AP).

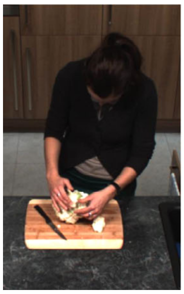
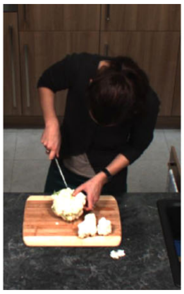
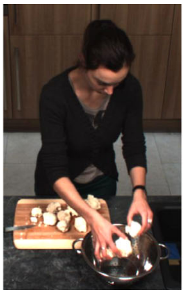
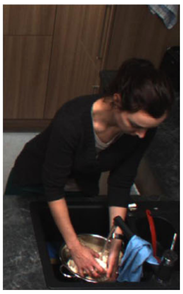
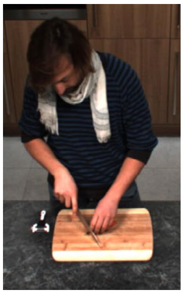
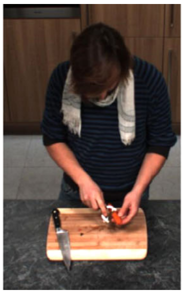
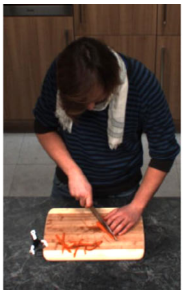
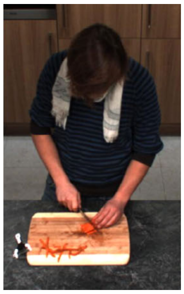
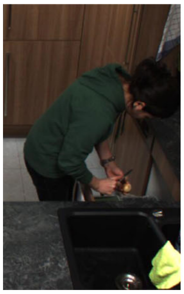
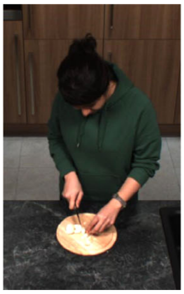
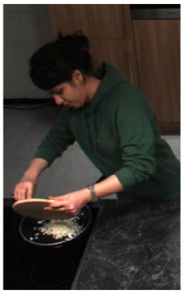
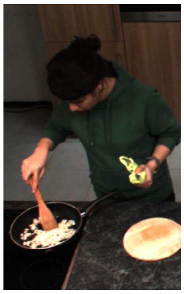
Fourth, using our Propagated Semantic Transfer (PST) approach is in most cases superior to other variants of incorporating script data (NN + Script data/ Script data). Most notably it reaches 57.5 % AP for our combined feature. This is the overall best performance and also outperforms the SVM with 56.6 % AP. PST slightly drops for the last number in table (19.3 %), which we found is due to rather suboptimal parameters selected on the validations set. We note that in the scenario of Disjoint Composites (top right quarter of Table 9) PST + Script data is outperformed by training an SVM. We attribute this to the fact that the attributes are less robust in this scenario (see Table 8) and the SVM can better adjust to that by learning which attributes are reliable and which not. NN and PST are based on distances between attribute score vectors, thus metric learning could be beneficial in these cases.

Fifth, script data does not only allow to achieve the maximum performance but also allows transfer (bottom part of Table 9) achieving in some cases results close to supervised approaches. The bottom right part of the table shows zero-shot recognition. Although here the performance cannot compete with the supervised setting, we like to point out that this is a very challenging scenario, where attributes are trained on different composites, without composite training data, and the video stream has to be segmented automatically.

Sixth, while in Table 9 we always used the variant tf∗idf-WN for Script data, we show different variants of Script data for the case where they are not combined with NN or PST in Table 10. The main observation is that freq-WN performs in all cases worst, most likely the WordNet expansions make the results noisier. While in the first column the tf∗idf-WN works best, there is overall no clear winner. However, when incorporated in PST, it is more important to select appropriate parameters for PST on the validation set rather than selecting the right variant of Script data.

Last, we want to look at an interesting comparison of the first line (SVM without attributes) versus line 8 (PST + Script data), which effectively compares the settings “only composite labels” versus “only attribute labels” (+ Script data). Although the latter does not have any labels for the actual task of composite recognition it either performs close (in case of Dense Trajectories) or slightly better (for combined features). This indicates that our PST + Script data approach is very good in transferring information from the original

**Table 11** Qualitative results for Dense Trajectories and its combination with hand-centric features (line 10 in Table 5) with respect to ground-truth (Color table online)

| | | | | | |
|---|---|---|---|---|---|
| |  |  |  |  | Composites |
| Ground-truth | **cauliflower**, **cutting-board**, **hand**, **pull apart(A)** | **cauliflower**, **cut(A)**, **cutting-board**, **knife** | **add(A)**, **cauliflower**, **colander**, **cutting-board**, **hand** | **cauliflower**, **colander**, **hand**, **wash(A)** | **Preparing cauliflower** |
| Dense Traj | **hand**, **cutting-board**, **pull apart(A)**, onion, peel, cut apart(A) | **knife**, **cutting-board**, cut apart(A), counter, chefs-knife, **cut(A)** | **hand**, **cutting-board**, move(A), counter, bowl, **colander** | **hand**, **wash(A)**, plate, **colander**, onion, peel | Preparing orange |
| Dense Traj, Hand-Traj, -cSift | **hand**, **cutting-board**, cut apart(A), **cauliflower**, onion, **pull apart(A)** | **cauliflower**, cut apart(A), **knife**, chefs-knife, **cutting-board**, **cut(A)** | **hand**, **cutting-board**, move(A), counter, **cauliflower**, **colander** | **hand**, **wash(A)**, bowl, **colander**, **cauliflower**, onion | **Preparing cauliflower** |
| |  |  |  |  | Composites |
| Ground-truth | **carrot**, **chefs-knife**, **cut off ends(A)**, **cutting-board** | **carrot**, **front-peeler**, **peel(A)** | **carrot**, **chefs-knife**, **cut stripes(A)**, **cutting-board** | **carrot**, **chefs-knife**, **cut apart(A)**, **cutting-board** | **Preparing carrot** |
| Dense Traj | **cutting-board**, cut apart(A), **chefs-knife**, **cut off ends(A)**, knife, put on(A) | cutting-board, **peel(A)**, **front-peeler**, chefs-knife, knife, cucumber | **cutting-board**, **chefs-knife**, slice(A), knife, cut apart(A), cucumber | **cutting-board**, **cut apart(A)**, **chefs-knife**, knife, cauliflower, cut off ends(A) | Preparing cucumber |
| Dense Traj, Hand-Traj, -cSift | **cutting-board**, **cut off ends(A)**, **chefs-knife**, cut apart(A), knife, **carrot** | cutting-board, **peel(A)**, **carrot**, chefs-knife, **front-peeler**, cucumber | **cutting-board**, **chefs-knife**, slice(A), knife, **carrot**, cut apart(A) | **cutting-board**, **cut apart(A)**, **chefs-knife**, cut off ends(A), knife, **carrot** | **Preparing carrot** |
| |  |  |  |  | Composites |
| Ground-truth | **knife**, **onion**, **peel(A)** | **chop(A)**, **cutting-board**, **knife**, **onion** | **add(A)**, **cutting-board**, **frying-pan**, **knife**, **onion** | **frying-pan**, **onion**, **spatula**, **stir(A)** | **Preparing onion** |
| Dense Traj | **peel(A)**, hand, **onion**, throw in garbage(A), bowl, front-peeler | **cutting-board**, **knife**, cut dice(A), onion, **chop(A)**, slice(A) | hand, **frying-pan**, **cutting-board**, pot, spatula, **add(A)** | **spatula**, **frying-pan**, **stir(A)**, **onion**, add(A), egg | **Preparing onion** |
| Dense Traj, Hand-Traj, -cSift | **peel(A)**, hand, throw in garbage(A), **onion**, **knife**, peel | **cutting-board**, **knife**, cut dice(A), slice(A), **chop(A)**, chive | hand, **frying-pan**, **add(A)**, pot, spatula, cauliflower | **frying-pan**, **spatula**, **stir(A)**, **onion**, add(A), broccoli | **Preparing onion** |

Top-6 highest scoring attributes (activities and objects) are shown, where (A) denotes activities. Composite activity predictions shown on the right. Correct results marked with bold. Note that many attributes are not correct according to the ground truth but very similar, e.g. we predict *slice* instead of *cut stripes* task it was trained on to another which is very important for adaptation to novel situations, typical for assisted daily living scenarios.

Table 11 provides qualitative results for three composite videos including how they are decomposed into attributes of fine-grained activities and participating objects.

## 8 Conclusion

In this work we address two challenges that have not been widely explored so far, namely fine-grained activity recognition and composite activity recognition. In order to approach these tasks we propose the large activity database MPII Cooking 2. We recorded and annotated 273 videos of more than 27 hours with 30 human subjects performing a large number of realistic cooking activities. Our database is unique with respect to size, length, complexity of the videos, and available annotations (activities, objects, human pose, text descriptions).

To estimate the complexity of fine-grained activity recognition in our database we compare three types of approaches: pose-based, hand-centric, and holistic. We evaluate on a classification and the often neglected detection task. Our results show that for recognizing fine-grained activities and their participating objects it is beneficial to focus on hand regions as the activities are hand-centric and the relevant objects are in the hand neighbourhood.

Composite activities are difficult to recognize because of their inherent variability and the lack of training data for specific composites. We show that attribute-based activity recognition allows recognizing composite activities well. Most notably, we describe 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.

As part of future work we plan to validate our hand-centric approach in other domains and exploit the scripts for composite activity recognition by modeling the temporal structure of the video.

**Acknowledgments** This work was supported by a fellowship within the FITweltweit-Program of the German Academic Exchange Service (DAAD), by the Cluster of Excellence "Multimodal Computing and Interaction" of the German Excellence Initiative and the Max Planck Center for Visual Computing and Communication.

## References

Amin, S., Andriluka, M., Rohrbach, M., & Schiele, B. (2013). Multi-view pictorial structures for 3D human pose estimation. In *Proceedings of the British Machine Vision Conference (BMVC)*. BMVA Press.

Andriluka, M., Roth, S., & Schiele, B. (2009). Pictorial structures revisited: People detection and articulated pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Andriluka, M., Roth, S., & Schiele, B. (2011). Discriminative appearance models for pictorial structures. *International Journal of Computer Vision (IJCV)*, *99*, 259–280.

Aubert, O., & Prié, Y. (2007). Advene: An open-source framework for integrating and visualising audiovisual metadata. In *MM*. ACM.

Baccouche, M., Mamalet, F., Wolf, C., Garcia, C., & Baskurt, A. (2011). Sequential deep learning for human action recognition. In *Human Behavior Understanding* (pp. 29–39). Springer.

Barr, A., & Feigenbaum, E. (1981). *The handbook of artificial intelligence* (Vol. 1). Los Altos: William Kaufman Inc.

Bloem, J., Regneri, M., & Thater, S. (2012). Robust processing of noisy web-collected data. In *KONVENS*.

Bojanowski, P., Lajugie, R., Bach, F., Laptev, I., Ponce, J., Schmid, C., & Sivic, J. (2014). Weakly supervised action labeling in videos under ordering constraints. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Brendel, W., & Todorovic, S. (2011). Learning spatiotemporal graphs of human activities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Campbell, L., & Bobick, A. (1995). Recognition of human body motion using phase space constraints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Chakraborty, B., Holte, M., Moeslund, T., Gonzalez, J., & Roca, X. (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)*.

Chaquet, J., Carmona, E., & Fernández-Caballero, A. (2013). A survey of video datasets for human action and activity recognition. *Computer Vision and Image Understanding*, *117*(6), 633–659.

Chen, D., & Dolan, W. (2011). Collecting highly parallel data for paraphrase evaluation. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Cherian, A., Mairal, J., Alahari, K., & Schmid, C. (2014). Mixing body-part sequences for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Dalal, N., Triggs, B., & Schmid, C. (2006). Human detection using oriented histograms of flow and appearance. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Das, P., Xu, C., Doell, R., & Corso, J. (2013). 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)*.

Divvala, S. K., Efros, A. A., & Hebert, M. (2012). How important are "Deformable Parts" in the Deformable Parts Model? In *Computer Vision–ECCV 2012. Workshops and Demonstrations* (pp. 31–40). Berlin, Heidelberg: Springer.

Elhoseiny, M., Saleh, B., & Elgammal, A. (2013). Write a classifier: Zero-shot learning using purely textual descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Everingham, M., Van Gool, L., Williams, C., Winn, J., & Zisserman, A. (2011). The PASCAL action classification taster competition. *International Journal of Computer Vision*, *88*, 303–338.

Farhadi, A., Endres, I., & Hoiem, D. (2010). Attribute-centric recognition for cross-category generalization. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Fathi, A., Farhadi, A., & Rehg, J. (2011). Understanding egocentric activities. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*. IEEE.

Fellbaum, C. (1998). *WordNet: An electronical lexical database*. Cambridge: The MIT Press.

Felzenszwalb, P., & Huttenlocher, D. (2005). Pictorial structures for object recognition. *International Journal of Computer Vision (IJCV)*, *61*, 55–79.

Felzenszwalb, P., Girshick, R., McAllester, D., & Ramanan, D. (2010). Object detection with discriminatively trained part-based models. *EEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *32*, 1627–1645.

Ferrari, V., Marin, M., & Zisserman, A. (2008). Progressive search space reduction for human pose estimation. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ferryman, J. (Ed.). (2007). *PETS*.

Fischler, M., & Elschlager, R. (1973). The representation and matching of pictorial structures. *IEEE Trans. Comput'73*.

Frome, A., Corrado, G., Shlens, J., Bengio, S., Dean, J., Ranzato, M., & Mikolov, T. (2013). Devise: A deep visual-semantic embedding model. In *Advances in Neural Information Processing Systems (NIPS)*.

Fu, Y., Hospedales, T., Xiang, T., & Gong, S. (2013). Learning multimodal latent attributes. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, (p. 99).

Gkioxari, G., Arbelaez, P., Bourdev, L., & Malik, J. (2013). Articulated pose estimation using discriminative armlet classifiers. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Guadarrama, S., Krishnamoorthy, N., Malkarnenkar, G., Venugopalan, S., Mooney, R., Darrell, T., & Saenko, K. (2013). Youtube2text: Recognizing and describing arbitrary activities using semantic hierarchies and zero-shoot recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Gupta, A., Srinivasan, P., Shi, J., & Davis, L. (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)*.

Jhuang, H., Gall, J., Zuffi, S., Schmid, C., & Black, M. (2013). Towards understanding action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia. IEEE.

Ji, S., Xu, W., Yang, M., & Yu, K. (2013). 3D convolutional neural networks for human action recognition. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *35*(1), 221–231.

Kantorov, V., & Laptev, I. (2014). Efficient feature extraction, encoding and classification for action recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Karlinsky, L., Dinerstein, M., & Ullman, S. (2010). Using body-anchored priors for identifying actions in single images. In *Advances in Neural Information Processing Systems (NIPS)*.

Karpathy, A., Toderici, G., Shetty, S., Leung, T., Sukthankar, R., & Fei-Fei, L. (2014). Large-scale video classification with convolutional neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Kliper-Gross, O., Hassner, T., & Wolf, L. (2012). The action similarity labeling challenge. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *34*(3), 615–621.

Kuehne, H., Jhuang, H., Garrote, E., Poggio, T., & Serre, T. (2011). Hmdb: A large video database for human motion recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

De la Torre, F., Hodgins, J., Montano, J., Valcarcel, S., Forcada, R., & Macey, J. (2009). *Guide to the cmu multimodal activity database*. Technical Report CMU-RI-TR-08-22, Robotics Institute.

Lampert, C., Nickisch, H., & Harmeling, S. (2013). Attribute-based classification for zero-shot learning of object categories. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, (p. 99).

Laptev, I. (2005). On space-time interest points. *International Journal of Computer Vision (IJCV)*, *64*, 107–123.

Laptev, I., & Pérez, P. (2007). Retrieving actions in movies. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Laptev, I., Marszalek, M., Schmid, C., & Rozenfeld, B. (2008). Learning realistic human actions from movies. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Le, Q., Zou, W., Yeung, S., & Ng, A. (2011). Learning hierarchical invariant spatio-temporal features for action recognition with independent subspace analysis. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)* (pp. 3361–3368). IEEE.

Li, L.-J., & Li, F.-F. (2007). What, where and who? classifying events by scene and object recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)* (pp. 1–8). IEEE.

Liu, J., McCloskey, S., & Liu, Y. (2012). Training data recycling for multi-level learning. In *21st International Conference on Pattern Recognition (ICPR), 2012*, (pp. 2314–2318), Nov 2012.

Liu, J., Luo, J., & Shah, M. (2009). Recognizing realistic actions from videos 'in the wild'. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Liu, J., Kuipers, B., & Savarese, S. (2011). Recognizing human actions by attributes. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Marszalek, M., Laptev, I., & Schmid, C. (2009). Actions in context. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, June 2009.

Messing, R., Pal, C., & Kautz, H. (2009). Activity recognition using the velocity histories of tracked keypoints. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Mittal, A., Zisserman, A., & Torr, P. (2011). Hand detection using multiple proposals. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Motwani, T. S., & Mooney, R. J. (2012). Improving video activity recognition using object recognition and text mining. In *ECAI* (pp. 600–605) August 2012.

Natarajan, P., & Nevatia, R. (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)*.

Niebles, J., Chen, C.-W., & Li, F.-F. (2010). Modeling temporal structure of decomposable motion segments for activity classification. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Nilsback, M.-E., & Zisserman, A. (2008). Automated flower classification over a large number of classes. In *ICVGIP* (pp. 722–729). IEEE.

Oh, S., Hoogs, A., Perera, A., Cuntoor, N., Chen, C.-C., Lee, J.T., et al. (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)* (pp. 3153–3160). IEEE.

Over, P., Awad, G., Michel, M., Fiscus, J., Sanders, G., Shaw, B., et al. (2012). Trecvid 2012—An overview of the goals, tasks, data, evaluation mechanisms and metrics. In *Proceedings of TRECVID 2012*. NIST, USA.

Packer, B., Saenko, K., & Koller, D. (2012). A combined pose, object, and feature model for action understanding. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Patron-Perez, A., Marszalek, M., Zisserman, A., & Reid, I.D. (2010). High five: Recognising human interactions in TV shows. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Pirsiavash, H., & Ramanan, D. (2012). Detecting activities of daily living in first-person camera views. In *Proceedings of the IEEE*

*Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Pirsiavash, H., & Ramanan, D. (2014). Parsing videos of actions with segmental grammars. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Ramanathan, V., Liang, P., & Li, F.-F. (2013). Video event understanding using natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Raptis, M., & Sigal, L. (2013). Poselet key-framing: A model for human activity recognition. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Regneri, M., Koller, A., & Pinkal, M. (2010). Learning script knowledge with web experiments. In *Proceedings of the Annual Meeting of the Association for Computational Linguistics (ACL)*.

Regneri, M., Rohrbach, M., Wetzel, D., Thater, S., Schiele, B., & Pinkal, M. (2013). Grounding action descriptions in videos. 1.

Rodriguez, M., Ahmed, J., & Shah, M. (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)*.

Roggen, D., Calatroni, A., Rossi, M., Holleczek, T., Forster, K., Troster, G., et al. (2010). Collecting complex activity data sets in highly rich networked sensor environments. In *INSS*.

Rohrbach, A., Rohrbach, M., Qiu, W., Friedrich, A., Pinkal, M., & Schiele, B. (2014). Coherent multi-sentence video description with variable level of detail. In *German Conference on Pattern Recognition (GCPR)*, September 2014.

Rohrbach, A., Rohrbach, M., Tandon, N., & Schiele, B. (2015). A dataset for movie description. *In Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rohrbach, M., Stark, M., Szarvas, G., Gurevych, I., & Schiele, B. (2010). What helps where–and why? Semantic relatedness for knowledge transfer. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rohrbach, M., Stark, M., & Schiele, B. (2011). Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rohrbach, M., Amin, S., Andriluka, M., & Schiele, B. (2012a). A database for fine grained activity detection of cooking activities. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Rohrbach, M., Regneri, M., Andriluka, M., Amin, S., Pinkal, M., & Schiele, B. (2012b). Script data for attribute-based recognition of composite activities. In *Proceedings of the European Conference on Computer Vision (ECCV)*.

Rohrbach, M., Ebert, S., & Schiele, B. (2013a). Transfer learning in a transductive setting. In *Advances in Neural Information Processing Systems (NIPS)*.

Rohrbach, M., Qiu, W., Titov, I., Thater, S., Pinkal, M., & Schiele, B. (2013b). Translating video content to natural language descriptions. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Ryoo, M., & Aggarwal, J. (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)*.

Salton, G., & Buckley, C. (1988). Term-weighting approaches in automatic text retrieval. In *Information Processing And Management*.

Sapp, B., Toshev, A., & Taskar, B. (2010). Cascaded models for articulated pose estimation.

Schank, R., & Abelson, R. (1977). Scripts, plans, goals and understanding.

Schuldt, C., Laptev, I., & Caputo, B. (2004). Recognizing human actions: A local SVM approach. In *ICPR*.

Senina, A., Rohrbach, M., Qiu, W., Friedrich, A., Amin, S., Andriluka, M., Pinkal, M., & Schiele, B. (2014). Coherent multi-sentence video description with variable level of detail, 03/2014. arXiv:1403.6173.

Shotton, J., Fitzgibbon, A., Cook, M., Sharp, T., Finocchio, M., Moore, R., et al. (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)* (pp. 1297–1304). IEEE.

Sill, J., Takács, G., Mackey, L., & Lin, D. (2009). Feature-weighted linear stacking. arXiv:0911.0460.

Singh, P., Lin, T., Mueller, E., Lim, G., Perkins, T., & Zhu, W. (2002). Open mind common sense: Knowledge acquisition from the general public. In *DOA, CoopIS and ODBASE, 2002*,

Singh, V., & Nevatia, R. (2011). Action recognition in cluttered dynamic scenes using pose-specific part models. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Socher, R., & Li, F.-F. (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)*, San Francisco, CA, June 2010.

Socher, R., Ganjoo, M., Manning, C.D., & Ng, A. (2013). Zero-shot learning through cross-modal transfer. In *Advances in Neural Information Processing Systems (NIPS)* (pp. 935–943).

Soomro, K., Zamir, A.R., & Shah, M. (2012). Ucf101: A dataset of 101 human actions classes from videos in the wild. Technical report, arXiv:1212.0402.

Stein, S., & McKenna, S. (2013). Combining embedded accelerometers with computer vision for recognizing food preparation activities. In *UbiComp*. ACM, September 2013.

Sung, J., Ponce, C., Selman, B., & Saxena, A. (2011). Human activity detection from RGBD images. *CoRR*, abs/1107.0169. informal publication.

Tang, K., Li, F.-F., & Koller, D. (2012). Learning latent temporal structure for complex event detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*, Providence, RI, USA, June 2012.

Tang, K., Yao, B., Li, F.-F., & Koller, D. (2013). Combining the right features for complex event recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Taylor, G.W., Fergus, R., LeCun, Y., & Bregler, C. (2010). Convolutional learning of spatio-temporal features. In *Proceedings of the European Conference on Computer Vision (ECCV)*, (pp. 140–153). Springer.

Tenorth, M., Bandouch, J., & Beetz, M. (2009). The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *THEMIS*.

Teo, C.L., Yang, Y., Daume, H., Fermuller, C., & Aloimonos, Y. (2012). Towards a watson that sees: Language-guided action recognition for robots. In *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (pp. 374–381). IEEE.

Ting, K.M., & Witten, I.H. (1997). Stacked generalization: When does it work? In *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*.

Vedaldi, A., & Zisserman, A. (2010). Efficient additive kernels via explicit feature maps. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, H., & Schmid, C. (2013). Action recognition with improved trajectories. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Sydney, Australia.

Wang, H., Ullah, M., Klaser, A., Laptev, I., & Schmid, C. (2009a). Evaluation of local spatio-temporal features for action recognition. In *Proceedings of the British Machine Vision Conference (BMVC)*.

Wang, H., Kläser, A., Schmid, C., & Liu, C.-L. (2011). Action recognition by dense trajectories. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Wang, H., Kläser, A., Schmid, C., & Liu, C. L. (2013a). Dense trajectories and motion boundary descriptors for action recognition. *International Journal of Computer Vision (IJCV)*, *103*, 60–79.

Wang, J., Markert, K., & Everingham, M. (2009). Learning models for object recognition from natural language descriptions. In Andrea Cavallaro, Simon Prince, Daniel C. Alexander (Eds.), *Proceedings of the British Machine Vision Conference (BMVC)* (pp. 1–11). British Machine Vision Association.

Wang, L., Qiao, Y., & Tang, X. (2013b). Mining motion atoms and phrases for complex action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*.

Welinder, P., Branson, S., Mita, T., Wah, C., Schroff, F., Belongie, S., et al. (2010). Caltech-ucsd birds 200. Technical Report, California Institute of Technology.

Yang, W., Wang, Y., & Mori, G. (2011). Recognizing human actions from still images with latent poses. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Yang, Y., & Ramanan, D. (2011). Articulated pose estimation with flexible mixtures-of-parts. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*. IEEE.

Yang, Y., & Ramanan, D. (2013). Articulated human detection with flexible mixtures of parts. *IEEE Transactions on Pattern Analysis and Machine Intelligence (TPAMI)*, *35*.

Yao, A., Gall, J., Fanelli, G., & Van Gool, L. (2011a). Does human action recognition benefit from pose estimation? In *Proceedings of the British Machine Vision Conference (BMVC)*.

Yao, B., & Li, F.-F. (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)*, *34*(9), 1691–1703.

Yao, B., Jiang, X., Khosla, A., Lin, A.L., Guibas, L.J., & Li, F.-F. (2011b). Action recognition by learning bases of action attributes and parts. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, Barcelona, Spain, November 2011b.

Yeffet, L., & Wolf, L. (2009). Local trinary patterns for human action recognition. In *Proceedings of the IEEE International Conference on Computer Vision (ICCV)*, 29 2009.

Yuan, J., Liu, Z., & Wu, Y. (2009). Discriminative subvolume search for efficient action detection. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition (CVPR)*.

Zhang, L., Khan, M.U.G., & Gotoh, Y. (2011). Video scene classification based on natural language description. In *Proceedings of the IEEE International Conference on Computer Vision Workshops (ICCV Workshops)* (pp. 942–949). IEEE.

Zhou, D., Bousquet, O., Lal, T.N., Weston, J., & Schölkopf, B. (2004). Learning with local and global consistency. In *Advances in Neural Information Processing Systems (NIPS)*.

Zinnen, A., Blanke, U., & Schiele, B. (2009). An analysis of sensor-oriented vs. model-based activity recognition. In *ISWC*.