A Database for Fine Grained Activity Detection of Cooking Activities

Marcus Rohrbach   Sikandar Amin   Mykhaylo Andriluka   Bernt Schiele
Max Planck Institute for Informatics, Saarbrücken, Germany

Abstract

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

1. Introduction

Human activity recognition has gained a lot of interest due to its potential in a wide range of applications such as human-computer interaction, smart homes, elderly/child care, or surveillance. At the same time, activity recognition still is in its infancy due to the many challenges involved: large variety of activities, limited observability, complex human motions and interactions, large intra-class variability vs. small inter-class variability, etc. Many approaches have been researched ranging from low level image and video features [6, 15, 38], over semantic human pose detection [33], to temporal activity models [12, 22, 32].

While impressive progress has been made, we argue that the community is still addressing only part of the overall activity recognition problem. When analyzing current benchmark databases, we identified three main limiting factors. First, many activities considered so far are rather coarse-grained, i.e. mostly full-body activities e.g. *jumping* or *waving*. This appears rather untypical for many application domains where we want to differentiate between more fine-grained activities, e.g. *cut* (Figure 1a) and *peel* (Figure 1f). Second, while datasets with large numbers of activities exist, the typical inter-class variability is high. This seems rather unrealistic for many applications such as surveillance or elderly care where we need to differentiate between highly similar activities. And third, many databases address the problem of activity *classification* only without looking into the more challenging and clearly more realistic problem of activity detection in a continuous data stream. Notable exceptions exist (see Sec. 2) even though these have other limitations such as small number of classes.

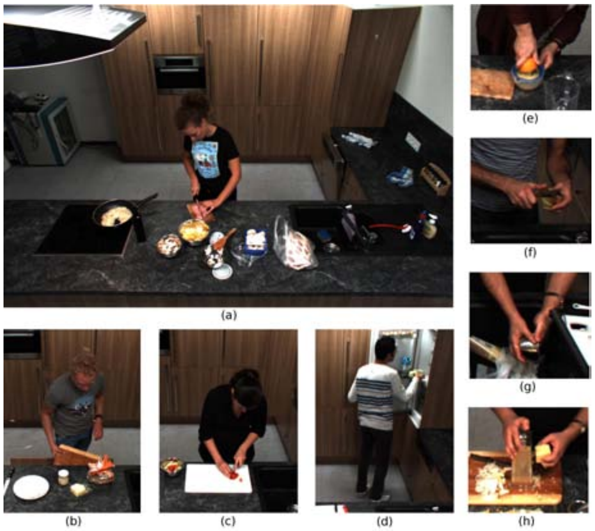

Figure 1. Fine grained cooking activities. (a) Full scene of *cut slices*, and crops of (b) *take out from drawer*, (c) *cut dice*, (d) *take out from fridge*, (e) *squeeze*, (f) *peel*, (g) *wash object*, (h) *grate*

This paper therefore proposes a new activity dataset that aims to address the above three shortcomings. More specifically we propose a dataset that contains 65 activities that are for the most part fine-grained, where the inter-class variability is low, and that are recorded continuously so that we can evaluate both classification and detection performance. More specifically, we consider the domain of recognizing cooking activities where it is important to recognize small differences in activities as shown in Figure 1, e.g. between *cut* (Figure 1a,c) and *peel* (Figure 1f), or at an even finer scale between *cut slices* (1a) and *cut dice* (1c).

Our contribution is twofold: First, we introduce a novel dataset which distinguishes 65 fine-grained activities. We propose a classification and detection challenge together with appropriate evaluation criteria. The dataset includes high resolution image and video sequences (jpg/avi), ac-

978-1-4673-1228-8/12/$31.00 ©2012 IEEE

Dataset	cls, det	classes	clips: videos	sub- jects	# frames	reso- lution
Full body pose datasets						
KTH [31]	cls	6	2,391	25	≈200,000	160x120
USC gestures [21]	cls	6	400	4		740x480
MSR action [43]	cls,det	3	63	10		320x240
Movie datasets						
Hollywood2 [19]	cls	12	1,707:69			
UCF50[1]	cls	50	>5,000			
HMDB51 [13]	cls	51	6,766			height:240
Coffee and Cigarettes [17]	det	2	264:11			
High Five [24]	cls,det	4	300:23			
Surveillance datasets						
PETS 2007 [11]	det	3	10		32,107	768x576
UT interaction [28]	cls,det	6	120	6		
VIRAT [23]	det	23	17			1920x1080
Assisted daily living datasets						
TUM Kitchen [36]	det	10	20	4	36,666	384x288
CMU-MMAC [14]	cls,det	>130		26		1024x768
URADL [20]	cls	17	150:30	5	$\leq$ 50,000	1280x720
Our database	cls,det	65	5,609:44	12	881,755	1624x1224

Table 1. Activity recognition datasets: We list if datasets allow for classification (cls), detection (det); number of activity classes; number of clips extracted from full videos (only one listed if identical), number of subjects, total number of frames, and resolution of videos. We leave fields blank if unknown or not applicable.

tivity class and time interval annotations, and precomputed mid level representations in the form of precomputed pose estimates and video features. We also provide an annotated body pose training and test set. This allows to work on the raw data but also on higher level modeling of activities. Second, we evaluate several video descriptor and activity recognition approaches. On the one hand we employ a state-of-the-art holistic activity descriptor based on dense trajectories [38] using a trajectory description, HOG and HOF [16], and MBH [7]. On the other hand we propose two approaches based on body pose tracks, motivated from work in the sensor-based activity recognition community [44]. From the experimental results we can conclude that fine grained activity recognition is clearly beyond the current state-of-the-art and that further research is required to address this more realistic and challenging setting.

2. Related work

We first discuss related datasets for activity recognition, and then related approaches to the ones benchmarked on our dataset. [1] gives an extensive survey of the field.

Activity Datasets Even when excluding single image action datasets such as [8], the number of proposed activity datasets is quite large ([2] lists over 30 datasets). Here, we focus on the most important ones with respect to database size, usage, and similarity to our proposed dataset (see Tab. 1). We distinguish four broad categories of datasets: full body pose, movie, surveillance, and assisted daily living datasets – our dataset falls in the last category.

The full body pose datasets are defined by actors performing full body actions. KTH [31], USC gestures [21], and similar datasets [33] require classifying simple full body and mainly repetitive activities. The MSR actions [43] pose a detection challenge limited to three classes. In contrast to these full body pose datasets, our dataset contains many 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 [18, 22, 25] datasets focus on sport activities. Kuehne *et al.*'s evaluation suggests that these activities can already be discriminated by static joint locations alone [13]. Hollywood2 [19] and HMDB51 [13] have very diverse activities. Especially HMDB51 [13] is an effort to provide a large scale database of 51 activities while reducing database bias. Although it includes similar, fine-grained activities, such as *shoot bow* and *shoot gun* or *smile* and *laugh*, most classes have a large inter-class variability and the videos are low-resolution. Although our dataset is easier in respect to camera motion and background it is challenging with respect to a smaller inter-class variability.

The Coffee and Cigarettes [17], High Five [24] are different to the other movie datasets by promoting activity detection rather than classification. This is clearly a more challenging problem as one not only has to classify a pre-segmented video but also to detect (or localize) an activity in a continuous video. As these datasets have a maximum of four classes, our dataset goes beyond these by distinguishing a large number of classes.

The third category of datasets is targeted towards surveillance. The PETS [11] or SDHA2010[2] workshop datasets contain real world situations form surveillance cameras in shops, subway stations, or airports. They are challenging as they contain multiple people with high occlusion. The UT interaction [28] requires to distinguish 6 different two-people interaction activities, such as *punch* or *shake hands*. The VIRAT [23] dataset is a recent attempt to provide a large scale dataset with 23 activities on nearly 30 hours of video. Although the video is high-resolution people are only of 20 to 180 pixel height. Overall the surveillance activities are very different to ours which are challenging with respect to fine-grained body-pose motion.

For the domain of *Assisted daily living (ADL) datasets*, which also includes our dataset, only recently datasets have been proposed in the vision community. The University of Rochester Activities of Daily Living Dataset (URADL) [20] 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.

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

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

In the TUM Kitchen dataset [36] all subjects perform the same high level activity (*setting a table*) and rather similar actions with limited variation. [14, 26] are 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 [14] all subjects prepare the identical five dishes with very similar ingredients and tools. In contrast to this our dataset contains 14 diverse dishes, where each subject uses different ingredients and tools in each dish.

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

Holistic approaches for activity recognition Most approaches for human activity recognition in video focus on using holistic video features, some use the human body pose as a basis. To create a discriminative feature representation of a video many approaches first detect space-time interest points [6, 15] or sample them densely [39] and then extract diverse descriptors in the image-time volume, such as histograms of oriented gradients (HOG) and flow (HOF) [16] or local trinary patterns [42]. Messing *et al.* [20] found improved performance by tracking Harris3D interest points [15]. The second of the two benchmark approaches we evaluate (see Sec. 4.2), is based on this idea: Wang *et al.* [38] track dense feature points and extract strong video features (HOG, HOF, MBH [7]) around these tracks. They report state-of-the art results on KTH [31], UCF YouTube [18], Hollywood2 [19], and UCF sports [25]. Other directions include template based approaches [25] or segmenting the space-temporal data and constructing a graph from this [5]. Another direction is to detect activities with a body-worn camera [34].

Body pose for activity recognition Many human activities such as *sitting*, *standing*, and *running* are defined in terms of body poses and their motion. However, with a few exceptions [10, 33], there exist little work on visual activity recognition based on articulated pose estimation. Pose-based activity recognition appears to work particularly well for images with little clutter and fully visible people as in the gesture dataset from [33]. Estimates of people poses were also used as auxiliary information for activity recognition in single images [40]. However, these systems have not shown to be effective in complex dynamic scenes with frequent occlusions, truncation and complex poses. So far, action recognition in such scenes was addressed only by holistic feature-based methods such as [16] due to the difficulty of reliable pose estimation in the complex real-world conditions.

Sung *et al.* [35] use depth information from a Kinect to estimate pose and distinguish 12 activities. However, in an initial test we found that the Kinect sensor has difficulties to capture fine grained activities due to limited resolution.

3. Fine grained human activities database

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

3.1. Database recording

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

In order to get a variation in activities we always told a participant beforehand to prepare a certain dish (e.g. *salad*), including a set of ingredients (*cucumber, tomatoes, cheese*) and potential tools (*grater*) to use. Instructions were given verbally and frequently participants diverted from the instructions by changing tools, and/or ingredients adding to the variability of the activities. Prior to recording participants were shown our kitchen and places of the required tools and ingredients to feel at home. During the recording participants could ask questions in case of problems and some listened to music. We always start the recording prior to the participant entering the kitchen and end it once the participant declares to be finished, i.e. we do not include the final cleaning process. There was a variety of 14 dishes, namely *sandwich, salad, fried potatoes, potato pancake, omelet, soup, pizza, casserole, mashed potato, snack plate, cake, fruit salad, cold drink*, and *hot drink*. Within these dishes each person used different ingredients resulting in very dissimilar videos, e.g. some participants cooked a packet soup while others prepared it from scratch. Dish preparation time varies from 3 to 41 minutes. For statistics on the activities see Table 5. Most participants were university students from different disciplines recruited by e-mail and publicly posted flyers and paid; cooking experience ranging from beginner cookers to amateur chefs.

We recorded in our kitchen (see Figure 1(a)) with a 4D View Solutions system using a Point Grey Grashopper camera with 1624x1224 pixel resolution at 29.4fps and global shutter. The camera is attached to the ceiling, recording a person working at the counter from the front. We provide the sequences as single frames (jpg with compression set to 75) and as video streams (compressed weakly with mpeg4v2 at a bitrate of 2500).

Method	Torso	Head	upper arm r	 l	lower arm r	 l	All
Original models							
CPS [29]	**67.1**	0.0	53.4	48.6	**47.3**	37.0	42.2
FMP [41]	63.9	**72.1**	**60.2**	**59.6**	42.1	**46.7**	**57.4**
PS [3]	58.0	45.5	50.5	57.2	43.3	38.8	48.9
Trained on our data							
FMP [41]	79.6	67.7	60.7	60.8	50.1	50.3	61.5
PS [3]	**80.1**	**80.0**	**67.8**	**69.6**	48.9	49.6	66.0
FPS (our model)	78.5	79.4	61.9	64.1	**62.4**	**61.0**	**67.9**

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

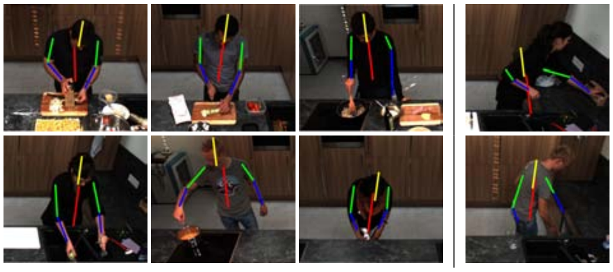

Figure 2. Examples of correctly estimated 2D upper body poses (left) and typical failure cases (right).

3.2. Database annotations

Activities were annotated with a two-stage revision phase by 6 people with start and end frame as well as the activity categories (see Tab. 5) using the annotation tool Advene [4]. The dataset contains a total of 5,609 annotations of 65 activity categories. This includes a background activity for the detection task which is generated automatically for all intervals without any other manual annotation for at least 30 frames (1 second), e.g. because the person is not (yet) in the scene or doing an unusual activity which is not annotated.

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

We also provide intermediate representations of holistic video descriptors, human pose detections, tracks, and features defined on the body pose (Sec. 4). We hope this will foster research at different levels of activity recognition.

4. Approaches

To better understand the state-of-the-art for the challenging task of fine-grained activity recognition we benchmark two approaches on our new dataset. The first (Sec. 4.1) uses features derived from an upper body model motivated by the intuition that human body configurations and human body motion should provide strong cues for activity recognition in general but particularly for fine-grained activity recognition. The second (Sec. 4.2) is a state-of-the-art method [38] that has shown promising results on various datasets.

4.1. Pose-based approach

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

Although pose-based activity recognition approaches were shown to be effective using inertial sensors [44], they have not been evaluated when the poses are estimated from monocular images. Inspired by [44] we build on a similar feature set, computing it from the temporal sequence of 2D body configurations. In the following we first evaluate the state-of-the-art in 2D pose estimation in the context of our dataset. We then introduce our pose-based activity recognition approach that builds on the best performing method.

2D human pose estimation In order to identify the best 2D pose estimation approach we use our 2D body joint annotations (see Sec. 3.2). We compare the performance of three recently proposed methods: the cascaded pictorial structures (CPS) [29], the flexible mixture of parts model (FMP) [41] and the implementation of pictorial structures model (PS) of [3]. Notice that these methods are designed for generic 2D pose estimation. In particular they do not rely on background subtraction or strong assumptions on the appearance of body limbs (*e.g.* skin color).

For evaluating these methods we adopt the PCP measure (percentage of correct parts) proposed in [10] that computes the percentage of body parts correctly localized by the pose estimation method. A body part is considered to be localized correctly if the predicted endpoints of the part are within half of the part length from their ground-truth positions. We first compare the implementations and pre-trained models made publicly available by the authors. Results are shown in the upper part of Tab. 2. The FMP model performs best, likely due to its ability to handle foreshortening of the body parts that occurs frequently in our data.

To push the performance further we retrain the two best performing models (FPM and PS) on our training set, which results in improvements from 57.4 to 61.5 PCP for the FMP model and from 48.9 to 66 PCP for the PS model (Tab. 2, last column). While demonstrating best results, the PS model is still defined in terms of rigid body parts, which is suboptimal for our task. In order to address that we define a flexible variant of the PS model (FPS) that instead of 6 parts used in the original model, consists of 10 parts corresponding to head, torso, as well as left and right shoul-

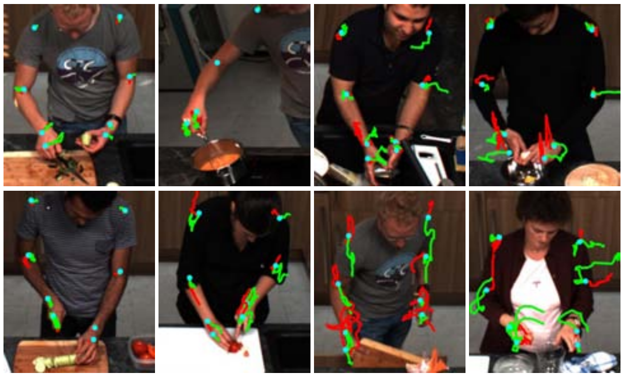

Figure 3. Sample tracks for different activities. *Backward tracks in green*, *forward tracks in red* and *initial pose in cyan*. First row, (left to right): *peel*, *stir*, *wash objects*, *open egg*, Second row (left to right): *cut slices*, *cut dice*, *take out from drawer*, *open egg*,

der, elbow, wrist and hand. While overall the extended FPS model improves over PS model only by 1.9 PCP (66.0 PCP for PS models vs. 67.9 PCP for FPS), it improves the detection of lower arms by more than 11 PCP which are most important for fined-grained activity recognition. Based on this comparison we rely on the FPS model in the subsequent steps of our pose-based activity recognition pipeline. Figure 2 visualizes several examples of the estimated poses for FPS. Notice that we can correctly estimate poses for a variety of activities and body configurations while maintaining precise localization of the body joints.

To extract the trajectories of body joints, an option is to extend our pose estimation to the temporal domain. However, temporal coupling of joint positions significantly complicates inference and approaches of this kind have only recently begun to be explored in the literature [30]. Moreover, our dataset consists of over 800,000 frames and to deal with this sheer complexity of estimating human poses for this dataset we choose a different avenue which relies on search space reduction [10] and tracking. To that end we first estimate poses over a sparse set of frames (every 10-th frame in our evaluation) and then track over a fixed temporal neighborhood of 50 frames forward and backward. For tracking we match SIFT features for each joint separately across consecutive frames. To discard outliers we find the largest group of features with coherent motion and update the joint position based on the motion of this group. In order to reduce the search space further we use a person detector [9] and estimate the pose of the person within the detected region with 50% border around.

This approach combines the generic appearance model learned at training time with the specific appearance (SIFT) features computed at test time. When initialized with successful pose estimates it provides reliable tracks of joints in the local temporal neighborhood (see Figure 3).

Body model and FFT features Given the body joint trajectories we compute two different feature representations: Manually defined statistics over the body model trajectories, which we refer to as *body model features* (BM) and Fourier transform features (FFT) from [44] which have shown effective for recognizing activities from body worn wearable sensors.

For the BM features we compute the *velocity* of all joints (similar to gradient calculation in the image domain) which we bin in a 8-bin histogram according to its direction, weighted by the speed (in pixels/frame). This is similar to the approach in [20] which additionally bins the velocity's magnitude. We repeat this by computing *acceleration* of each joint. Additionally we compute *distances* between the right and corresponding left joints as well as between all 4 joints on each body half. For each distance trajectory we compute statistics (mean, median, standard deviation, minimum, and maximum) as well as a rate of change histogram, similar to velocity. Last, we compute the angle trajectories at all inner joints (wrists, elbows, shoulders) and use the statistics (mean etc.) of the angle and angle speed trajectories. This totals to 556 dimensions.

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

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

4.2. Holistic approach

Most approaches for activity recognition are based on a bag-of-words representations. We pick a recently suggested approach [38] which extracts histograms of oriented gradients (HOG), flow (HOF) [16], and motion boundary histograms (MBH) [7] 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 [38] we generate a codebook for each of the four features of 4,000 words using k-means from over a million sampled features.

4.3. Activity classification and detection

We train classifiers on the feature representation described in the previous section given the ground truth intervals and labels. We train one-vs-all SVMs using mean SGD [27] with a χ^2 kernel approximation [37]. While we use ground truth intervals for computing classification results we use a sliding window approach to find the correct interval of a detection. To efficiently compute features of a sliding window we build an integral histogram over the histogram of the codebook features. We use non maximum suppression over different window lengths and start with the maximum score and remove all overlapping windows. In the detection experiments we use a minimum window size of 30 with a step size of 6 frames; we increase window and step size by a factor of $\sqrt{2}$ until we reach a window size of 1800 frames (about 1 minute). Although this will still not cover all possible frame configurations, we found it to be a good trade-off between performance and computational costs.

5. Evaluation

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

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

5.1. Classification results

Tab. 3 summarizes the classification results. The first section of the table shows results for the approaches based on the articulated pose model (see Sec. 4.1), while the second section shows results of the state-of-the-art holistic dense trajectories [38] feature representation (see Sec. 4.2). Overall we achieve a mean multi-class recall or accuracy (Tab. 3, second last column), between 21.8% and 45.1% which should be compared to chance level of 1.6% for the 64 classes (we exclude the background class for classification).

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

Table 3. Classification results, in % (see Sec. 5.1)

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

Next we compare the results of the holistic approaches (Sec. 2, Tab. 3) based on dense trajectories [38]. Trajectory has the lowest performance with 35.4% Pr, 33.3% Rc, and 42.0% AP. In line with results reported by [38] for other datasets HOG, HOF, and motion boundary histograms (MBH) improve over this performance. MBH achieves 44.6% Pr, 40.5% Rc, and 52.0% AP. Combining all holistic approaches again significantly improves performance by more than 4% to 49.4% Pr, 44.8% Rc and 59.2% AP.

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

A low-level combination of pose and holistic approaches (Tab. 3, last line) shows slight improvement over the holistic approach (Tab. 3, second last line). We achieve 50.4% multi-class precision, 45.1% multi-class recall (or accuracy), and 57.9% AP (slightly dropped). Although we believe this is an encouraging first result, it shows that fine-grained activity recognition is indeed difficult.

A more detailed class level evaluation based on the con-

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

Table 4. Detection results, in % (see Sec. 5.2)

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

Table 5. Detection results per class in % (see Sec. 5.2). Column 2: total number of annotations; columns 3 to 5: AP for (the combined version of) pose-based, holistic, and pose + holistic approaches; column 6,7: multi-class precision and recall for the Combined pose + holistic approach.

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

Examining the intermediate representation of 2D tracks we found that the tracks for fine-grained activities *peel* vs. *cut slices* (Figure 3, first column) can distinguish fine-grained movements (sideways hand movement vs. vertical movement) highlighting the potential benefit of using body-pose features.

5.2. Detection results

Tab. 4 shows detection results and Tab. 5 results per class of the respective combined approaches. Overall performance ranges from the combined pose-based approaches of 17.7% AP (8.6% Pr, 21.3% Rc) over 44.2% AP (17.7% Pr, 40.3% Rc) for the holistic approaches to 45.0% AP (19.8% Pr, 40.2% Rc) when combining pose-based and holistic. The improvements of the combination, similar to the classification results, underlines the complementary nature of the two approaches. Even though overall the performance for the detection task (Tab. 4) is lower than for classification (Tab. 3) the relative performances are similar: pose-based approaches perform below holistic approaches and combining individual approaches improves performance, respectively. In all cases multi-class precision is significantly lower than recall, indicating a high number of false positives. Frequently short activity fragments score very high within other longer fragments or sometimes one ground truth label is fragmented into several shorter ones. We hope this dataset will provide a base for exploring how to best attack these multi-class activity detection challenges.

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

6. Conclusion

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

We benchmark two approaches on this dataset. The first is based on human body joint trajectories and the second on state-of-the-art holistic features [38]. Combined they achieve 45.1% mean multi-class recall or accuracy and 57.9% mean average precision on the classification task and 45.0% mean average precision on the detection task. Individually the pose-based approach is outperformed by the dense trajectories which can be attributed to limitations of current articulated pose estimation approaches and the sparser and weaker feature representation. Our analysis of the detection task suggests that especially fine-grained activities are very difficult to detect.

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

Acknowledgement: We would like to thank Ulf Blanke for code and discussions on activity recognition.

References

[1] J. Aggarwal and M. Ryoo. Human activity analysis: A review. *ACM Comput. Surv.*, 43, Apr. 2011. 2

[2] M. A. R. Ahad, J. Tan, H. Kim, and S. Ishikawa. Action dataset - a survey. In *SICE*, 2011. 2

[3] M. Andriluka, S. Roth, and B. Schiele. Pictorial structures revisited: People detection and articulated pose estimation. In *CVPR*, 2009. 4

[4] O. Aubert and Y. Prié. Advene: an open-source framework for integrating and visualising audiovisual metadata. In *ACM Multimedia*, 2007. 4

[5] W. Brendel and S. Todorovic. Learning spatiotemporal graphs of human activities. In *ICCV*, 2011. 3

[6] B. Chakraborty, M. B. Holte, T. B. Moeslund, J. Gonzalez, and F. X. Roca. A selective spatio-temporal interest point detector for human action recognition in complex scenes. In *ICCV*, 2011. 1, 3

[7] N. Dalal, B. Triggs, and C. Schmid. Human detection using oriented histograms of flow and appearance. In *ECCV*, 2006. 2, 3, 5

[8] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The PASCAL action classification taster competition, 2011. 2

[9] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *PAMI*, 32, 2010. 5

[10] V. Ferrari, M. Marin, and A. Zisserman. Progressive search space reduction for human pose estimation. In *CVPR 2008*. 3, 4, 5

[11] J. M. Ferryman, editor. *PETS*, 2007. 2

[12] D. Gehrig, H. Kuehne, A. Woerner, and T. Schultz. Hmm-based human motion recognition with optical flow data. In *Humanoid Robots'09*. 1

[13] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: A large video database for human motion recognition. In *ICCV*, 2011. 2

[14] F. D. la Torre, J. Hodgins, J. Montano, S. Valcarcel, R. Forcada, and J. Macey. Guide to the cmu multimodal activity database. Technical Report CMU-RI-TR-08-22, Robotics Institute. 2, 3

[15] I. Laptev. On space-time interest points. In *IJCV*, 2005. 1, 3

[16] I. Laptev, M. Marszalek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *CVPR*, 2008. 2, 3, 5

[17] I. Laptev and P. Pérez. Retrieving actions in movies. In *ICCV*, 2007. 2

[18] J. G. Liu, J. B. Luo, and M. Shah. Recognizing realistic actions from videos 'in the wild'. In *CVPR*, 2009. 2, 3

[19] M. Marszalek, I. Laptev, and C. Schmid. Actions in context. In *CVPR*, 2009. 2, 3

[20] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *ICCV*, 2009. 2, 3, 5

[21] P. Natarajan and R. Nevatia. View and scale invariant action recognition using multiview shape-flow models. In *CVPR*, 2008. 2

[22] J. Niebles, C.-W. Chen, and L. Fei-Fei. Modeling temporal structure of decomposable motion segments for activity classification. In *ECCV 2010*. 1, 2

[23] S. Oh and et al. A large-scale benchmark dataset for event recognition in surveillance video. In *CVPR*, 2011. 2

[24] A. Patron-Perez, M. Marszalek, A. Zisserman, and I. D. Reid. High five: Recognising human interactions in TV shows. In *BMVC*, 2010. 2

[25] M. D. Rodriguez, J. Ahmed, and M. Shah. Action MACH a spatio-temporal maximum average correlation height filter for action recognition. In *CVPR*, 2008. 2, 3

[26] D. Roggen and et al. Collecting complex activity data sets in highly rich networked sensor environments. In *ICNSS*, 2010. 3

[27] M. Rohrbach, M. Stark, and B. Schiele. Evaluating knowledge transfer and zero-shot learning in a large-scale setting. In *CVPR*, 2011. 6

[28] M. S. Ryoo and J. K. Aggarwal. Spatio-temporal relationship match: Video structure comparison for recognition of complex human activities. In *ICCV*, 2009. 2

[29] B. Sapp, A. Toshev, and B. Taskar. Cascaded models for articulated pose estimation. In *ECCV*, 2010. 4

[30] B. Sapp, D. Weiss, and B. Taskar. Parsing human motion with stretchable models. In *CVPR 2011*. 5

[31] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local SVM approach. In *ICPR*, 2004. 2, 3

[32] Z. Sia, M. Peib, B. Yaoa, and S.-C. Zhua. Unsupervised learning of event and-or grammar and semantics from video. In *ICCV*, 2011. 1

[33] V. Singh and R. Nevatia. Action recognition in cluttered dynamic scenes using pose-specific part models. In *ICCV*, 2011. 1, 2, 3

[34] E. H. Spriggs, F. de la Torre, and M. Hebert. Temporal segmentation and activity classification from first-person sensing. In *Egoc.Vis.'09*. 3

[35] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from RGBD images. In *AAAI Workshop*, 2011. 3

[36] M. Tenorth, J. Bandouch, and M. Beetz. The TUM Kitchen Data Set of Everyday Manipulation Activities for Motion Tracking and Action Recognition. In *THEMIS*, 2009. 2, 3

[37] A. Vedaldi and A. Zisserman. Efficient additive kernels via explicit feature maps. In *CVPR*, 2010. 6

[38] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *CVPR*, 2011. 1, 2, 3, 4, 5, 6, 8

[39] H. Wang, M. Ullah, A. Klaser, I. Laptev, and C. Schmid. Evaluation of local spatio-temporal features for action recognition. In *BMVC'09*. 3

[40] W. Yang, Y. Wang, and G. Mori. Recognizing human actions from still images with latent poses. In *CVPR 2010*. 3

[41] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*, 2011. 4

[42] L. Yeffet and L. Wolf. Local trinary patterns for human action recognition. In *ICCV*, 2009. 3

[43] J. S. Yuan, Z. C. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *CVPR*, 2009. 2

[44] A. Zinnen, U. Blanke, and B. Schiele. An analysis of sensor-oriented vs. model-based activity recognition. In *ISWC*, 2009. 2, 4, 5