Remember and Transfer what you have Learned – Recognizing Composite Activities based on Activity Spotting

Ulf Blanke
Computer Science, TU Darmstadt
blanke@cs.tu-darmstadt.de

Bernt Schiele
MPI Informatics, Saarbrücken
schiele@mpi-inf.mpg.de

Abstract

Activity recognition approaches have shown to enable good performance for a wide variety of applications. Most approaches rely on machine learning techniques requiring significant amounts of training data for each application. Consequently they have to be retrained for each new application limiting the real-world applicability of today's activity recognition methods. This paper explores the possibility to transfer learned knowledge from one application to others thereby significantly reducing the required training data for new applications. To achieve this transferability the paper proposes a new layered activity recognition approach that lends itself to transfer knowledge across applications. Besides allowing to transfer knowledge across applications this layered approach also shows improved recognition performance both of composite activities as well as of activity events.

1 Introduction

Human activity recognition using wearable sensors has been an active research area due to its importance for context-aware-systems. Substantial progress has been achieved for application scenarios [2, 8, 21] spanning across different areas, such as the entertainment, industrial or healthcare domains. Some devices are already commercially available, e.g., watches that log the wearer's motion over weeks to infer fitness levels or the Nintendo Wii gesture controllers.

While the recognition of low level activities, such as walking, standing or sitting yields impressive results, e.g., [15], recognition of more complex or composed activities, such as cooking or cleaning, is far less researched, e.g., [1, 12], and consequently still an open research question. Probably the single most important difficulty to recognize complex activities is the inherent variability in executing such activities by different people and with different durations. As most of today's activity recognition methods rely on machine learning techniques this variability requires prohibitive amounts of training data for each novel application.

The starting point of this paper is therefore to take a fresh look at the problem and to explicitly design a novel activity recognition architecture that is suited to transfer acquired knowledge from one application to another. The main goal is thereby to reduce the amount of required training data to a minimum while allowing to reliably recognize complex activities. The basic idea is illustrated in Fig. 1. Let us

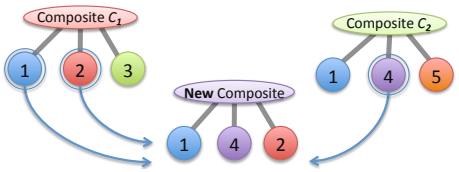

Figure 1. Transferring knowledge of activity events to construct new composite activities

assume we want to learn the model of a new composite activity (C_{new}) that shares low-level activities ('1', '2', '4') with two previously learned composite activities (C_1, C_2). Rather than to learn C_{new} from scratch requiring substantial amounts of training data we can transfer the low-level activity recognizers from C_1 and C_2 to train (using few training samples) or even construct (using prior knowledge) a model for C_{new}. The underlying hierarchical model that enables such transfer of knowledge is in line with research in psychology and linguistics that show evidence that there is good agreement on how humans perceive complex activities [6, 20]. Accordingly we borrow the term *partonomy* [31], describing "part-of-whole" relationships between activities and composites of activities. We explicitly use the term *activity events* for underlying activities. For high level activities composed of such underlying events, we use the term *composite activity*.

The main contributions are threefold. First, we propose a new partonomy-based approach that lends itself for knowledge transfer thereby reducing the required amounts of training data for new complex activities and applications. To this end, we propose a multi-layered discriminative model using a combination of joint boosting [26] and conditional random fields [14] for detecting composite activities from spotted activity events (Sec. 3). Second, we investigate the possibility to transfer activity events across datasets

to recognize new composite activities (Sec. 5). Third, we show that the partonomy-based approach improves both the recognition of composite activities as well as the recognition of activity events compared to a direct approach without hierarchical structure (Sec. 4 and 5).

2 Related Work

Activity recognition received much attention in recent years aiming at a variety of different activities. Impressive performance has been reported for activities such as standing, sitting, walking [15] or gestures [18, 30]. Recently, also the recognition of complex activities [1, 12, 24] has been explored. Such activities consist of various underlying activities and usually span over larger timescale. To capture such complex activities, temporal probabilistic models are applied, including hidden Markov models (HMMs) [18, 23] and conditional random fields (CRFs). In [11] a framework is presented for recognizing concurrent and interleaved activities based on skip-chain CRFs. In [17] hierarchical CRFs are applied to model activities and significant places of individuals using GPS and high level context information. Within robotics a performance comparison between CRFs and HMMs is conducted [27], concluding that CRFs can yield better recognition. This concords to [14, 29]. In [4] recognition of complex activities is approached top-down to identify specific low level activities that discriminate complex high level activities. Interestingly, their findings match the observation that humans themselves can identify activities by their distinctive parts without recovering the part structure itself [7, 25, 31]. While in computer vision the potential of hierarchies [19] or partonomies [10] for object detection is being well explored, it has not yet been addressed widely for activity recognition.

More recently, the problem of knowledge transfer has been identified as an important challenge. In [28] knowledge about activities within two houses is transferred to a third house by linking commonly used meta features. In [32] activities from one domain are used to help learning in a second, related domain. To link the domains a similarity measurement is derived from the web using an activity taxonomy of both domains. In contrast to their work, focusing on taxonomies of activities, we focus on the partonomy of activities. While *dish-washing* is in a "kind-of"-relationship to *cleaning indoor*, i.e., characterized by a taxonomy, *fill water into basin, scrub dishes, dry dishes* are in a "part-of" relationship of *dish-washing*, creating a partonomy. Bobick [5] classifies motion into a similar partonomy to support the problem of motion-understanding. Different levels of motion are introduced, containing different levels of knowledge.

Within this paper we focus on transferable activity events that constitute such part-of-relationships and can be reused across different composite activities. [22] advocates to transfer prior knowledge of higher level context, while requiring training of lowlevel models only but without providing convincing empirical evidence. In contrast, we propose a layered approach and also experimentally show the potential of transferring activity events to learn new composited activities with minimal training.

3 Partonomy-based Activity Recognition

As outlined before our goal is to recognize activities, that are composed of underlying activity events. Within this scenario we address the following research questions:

1. Does the partonomy-approach improve the recognition of composite activities?
2. Can we transfer knowledge about activity events to learn and recognize new composite activities with minimal training?
3. By learning composites, can we use this knowledge to improve recognition of underlying activity events?

To address these questions, we propose a *recursive bottom-up* and *top-down* multilayer approach. In the first layer L_1, we perform *activity spotting*, to capture *activity events*. Activity events usually appear sporadically within a large background class. This complicates their detection, as these events are easily confused with background events.

For subsequent composite activity layers L_n ($n \geq 2$), we create a partonomy of L_{n-1}-events to infer composite L_n-activities (*bottom-up*). Note that L_n-activities can be recursively composed by L_{n-1}-composite activities. Here we assume knowledge of relevant events of the partonomy. Furthermore, we use information (in a *top-down* fashion) of L_n-composites to gain certainty about underlying L_{n-1}-events and to improve their recognition. As mentioned before, common models to capture temporal sequences are hidden Markov models (HMMs) or conditional random fields (CRFs). As such, they are ideal to model the partonomy of composite activities. In contrast to HMMs, CRFs impose weaker independence assumptions between observations, allowing to directly model temporal dependencies between events for a composite. In this work, spotted activity events and their corresponding probability are fed into a CRF, generating the posterior of the composite activity.

The following Sec. 3.1 describes the activity spotting approach and Sec. 3.2 the composite activity recognition. Please note, that we omit time indices for better readability and denote vectors in bold font.

3.1 L_1-Activity Event Spotting

For L_1-activity event spotting we use a three-step method similar to [33]: segmentation, feature calculation and classification.

Segmentation. To reveal potentially interesting fragments in a continuous sensor stream, we apply a segmentation

technique replacing the standard (fixed) sliding window approach. By assuming low-movement moments at the start and end of interactions, segments of interest are created using such points. This segmentation technique is based on a human body model, which can be inferred by 5 inertial measurement units. Note that we create an over-segmentation containing overlapping segments with 100% recall [33].

Features. Given the sensor data, the segments and the body model from above, we then calculate features. We use common features, such as mean and variance [3, 16, 13], and the body model features from [33]. Primitives of activities, such as moving the arms up or down, bending over, push-or-pull the hands are calculated. In total we gain a feature dimension of about 1700.

Classification. L_1-activity events are spotted using joint boosting [26]. Its feature selection mechanism is useful to select relevant features, which gains computational efficiency for detection. For each segment s we calculate confidences H_s^c for each activity event class c and normalize them by $U_s^c = \frac{exp\{H_s^c\}}{\sum_i exp\{H_s^i\}}$. Next, we pass a feature vector $\mathbf{x}_s = [t, \mathbf{U}_s]^T$ consisting of a segment's (L_1-event's) central time t and the normalized confidences to the L_2-composite activity model.

3.2 L_n-composite Activity Recognition

Given a set of spotted L_1-activity events $\mathbf{x}_s$ in the datastream, we can combine these to composed L_2-activities. Moreover, we define a *recursive* model, enabling the construction of new L_n-composites using underlying L_{n-1}-composites as events. Then $\mathbf{x}_s$ corresponds to posteriors of L_{n-1}-composites and their central time. At this point, one has to be aware of the uncertainty when detecting L_{n-1}-activity events: Errors occur in terms of false detections (*false positives*) or missed events (*false negatives*). We apply a sampling-technique and forward sampled events to a temporal probabilistic model. By selecting relevant activity events and their combination we learn and infer composite activities. Let us now look at the algorithm in more detail.

Sampling. As the exact composition of activity events is priorly unknown due to imperfect recognition, all possible combinations of relevant activity events have to be sampled. To reduce computation time we perform sub-sampling by using a sliding window in event-space. Note the difference to a sliding window in timeframe-space. As mentioned earlier, we can also miss events. To handle this type of error, we 'hallucinate' potentially missed events by assigning a low default probability.

Features for Compositions. To enable L_n-composite activity recognition, we exploit knowledge about L_{n-1}-activity events, probability and the temporal distances between these events. A Gaussian distance model defined by mean and variance of event distances, maps the timeframe distance between two events to a closed interval between 0 and 1. We calculate the Gaussian from label distances on the training data. Note that we use regularization due to few amounts of training samples.

Composition modeling. As mentioned above, we choose CRFs to model L_n-composite activities by using the following information: *event class* and *probability of event*, *order of events* and *distance between events*. In contrast to HMMs, CRFs are able to directly model dependencies between events, i.e., their temporal distance, using potentials between connected nodes.

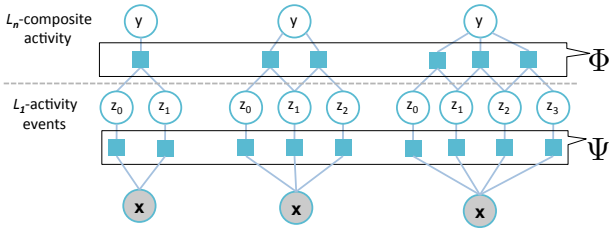

Figure 2. Factorgraph of CRFs for composite activities with 2, 3 or 4 activity events.

For each composite activity m an individual CRF is created. Each CRF consists of a single composite node with binary random variable y and e event nodes, with discrete random variables $\mathbf{z} = [z_1, .., z_e]$ and $z_i \in Z = \{1, .., e\}$ (see Fig. 2). The input feature is denoted by $\mathbf{x} = [\mathbf{x}_1, .., \mathbf{x}_e]$. The probability of the model for one composite is given by

$$P_m(y, \mathbf{z}|\mathbf{x}, \mathbf{v}, \mathbf{w}) = \frac{1}{\mathcal{Z}} \cdot \prod_i \Psi_i(z_i, \mathbf{x}_i, \mathbf{w}_i) \cdot \prod_{i,j} \Phi_{ij}(z_i, z_j, y, \mathbf{x}_i, \mathbf{x}_j, \mathbf{v}_{ij}) \quad (1)$$

for each model m. $\mathcal{Z}$ corresponds to the normalizing partition function. $\mathbf{v}, \mathbf{w}$ parameterize logistic regression functions used for the unary and pairwise potentials. For each model m specific weights $\mathbf{v}$ and $\mathbf{w}$ are used. Due to notational simplicity we omit the index m for the weights.

3.2.1 Unary Potentials

To consider the uncertainty of detected L_{n-1}-events when modeling L_n-composite activities, we set the unary potential functions for each node i as follows:

$$\Psi_i(z_i = c, \mathbf{x}_i, \mathbf{w}_i^c) = exp\{[\mathbf{U}_i^T, 1]\ \mathbf{w}_i^c\}, \text{where } c \in Z \quad (2)$$

Here $\mathbf{U}_i$ correspond to the class confidences from step 1 in Sec. 3.1. $\mathbf{w}$ weights the event confidences by their importance. Note that we pad an additional constant to model the hyperplane offset.

3.2.2 Pairwise Potentials

Intuitively, the pairwise potentials correspond to temporal and sequential dependencies between L_{n-1}-activity events. As can be seen in Fig. 2 our model contains ternary cliques, connecting pairwise sequential L_{n-1}-event nodes i and j to the L_n-composite activity node. The equation for the potential function is given by

$$\Phi_{ij}(z_i, z_j, y = c_3, \mathbf{x}_i, \mathbf{x}_j, \mathbf{v}_{ij}) = \quad (3)$$
$$exp\{[f_1(\mathbf{x}_i, \mathbf{x}_j), f_2(\mathbf{x}_i, \mathbf{x}_j), 1]\ \mathbf{v}_{ij}^{c_3}\}$$

with binary label c_3. For the pairwise potentials we calculate the two features functions f_1 and f_2. First, we inject temporal constraints between activity events into our model:

$$f_1(\mathbf{x}_i, \mathbf{x}_j) = \mathcal{N}(t_i - t_j; \mu, \sigma) \quad (4)$$

A Gaussian model $\mathcal{N}$ is learned on the distance of each sequential pair of activity events on the training data. A second feature function models the sequence of co-occurring activity events:

$$f_2(\mathbf{x}_i, \mathbf{x}_j) = |\mathbf{U}_i - \mathbf{U}_j| \quad (5)$$

by subtracting neighboring features. Again we pad a constant representing the offset by the hyperplane to the feature vector in Eq. 3.

Inference and Learning. We use loopy belief propagation [9] to infer the marginal probabilities in the nodes. Before describing the learning, let us define training data $\mathcal{X} = \{\mathbf{x}^1, .., \mathbf{x}^N\}$ with labels $\mathcal{Y} = \{y^1, .., y^N\}$ and $\mathcal{Z} = \{\mathbf{z}^1, ..\mathbf{z}^N\}$ with N number of training samples. Given marginal beliefs b, we learn the weights by gradient descent for each class $c \in Z$, binary composite class c_3 and event node i:

$$\frac{\partial log P_m(\mathcal{Y}, \mathcal{Z}|\mathcal{X})}{\partial \mathbf{w}_i^c} = \sum_{n=1}^{N} [\mathbf{U}_i \cdot (\delta(z_i^n = c) - b_i^n(c))] \quad (6)$$

for unary potentials and

$$\frac{\partial log P_m(\mathcal{Y}, \mathcal{Z}|\mathcal{X})}{\partial \mathbf{v}_{ij}^{c_3}} = \sum_{n=1}^{N} ([f_1(\mathbf{x}_i, \mathbf{x}_j), f_2(\mathbf{x}_i, \mathbf{x}_j), 1]^T \cdot (\delta(y^n = c_3) - b^n(c_3))) \quad (7)$$

for the pairwise potentials. As the ratio between positive and negative data (background) is unbalanced we reuse positives samples and regroup them with negative samples, until all samples are considered.

To evaluate our model, we use two datasets called *Bookshelf* and *Mirror*. We first describe the Bookshelf-dataset and show applicability of our approach by comparing to a more standard approach for activity recognition. Here we consider two layers of activities L_1 and L_2. Then we show the ability of knowledge transfer by considering the partonomy of activities on the *Mirror* dataset. To analyze our model with respect to its recursive property we consider three layers of activities L_1, L_2 and L_3.

4 Composite Activity Recognition on the Bookshelf Dataset

Standard approaches for activity recognition make often use of classifiers, which do not consider hierarchical structure, respectively partonomies of composite activities. Often such approaches are combined with a sliding window, for which probabilities of learned activities are estimated.

First, we compare our proposed partonomy-based approach to the *direct* approach using joint boosting and a sliding window as baseline. More specifically, we investigate how both approaches perform when reducing training data for composite activity recognition. We conduct these experiments on a dataset called *Bookshelf*.

L_2-Composite Activity	L_1-Activity Events
Fix Side Parts	Marking–Drill–Screw (Fix Side Parts)
Join Side Parts	Marking–Drill–Screw (Join Side Parts)
Make Back Part	Sawing–Drill–Screw–Hammering
Assemble Box	Marking–Hammering
Hang Up Box	Marking–Drill–Screw–Hang-Up Box
Create Template	2x Marking–Cut Template

Table 1. Six composite activities and the corresponding activity events for *Bookshelf*

4.1 Bookshelf Dataset

As motivated in the introduction we want to advance the state-of-the-art in activity recognition by considering composite physical activities. Recognizing activities within maintenance or construction scenarios is specifically interesting. In order to avoid forgetting steps or to facilitate the reentry into interrupted working cycles context aware systems can be helpful. Therefore we choose a realistic dataset in a workshop scenario, called *Bookshelf*, in which subjects construct a wooden bookshelf (Fig. 3). Typical for

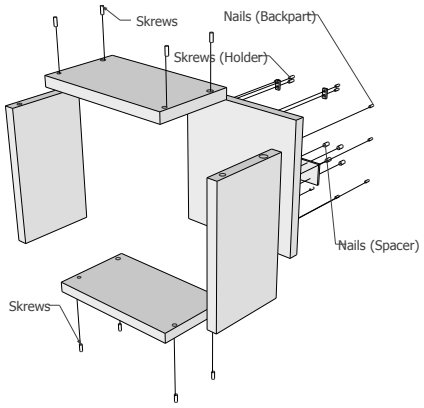

Figure 3. Explosion drawing of Bookshelf

construction tasks, the dataset consists of a variety of L_1-activity events, such as *drilling* or *screwing* with high variability in execution. These are grouped into several steps (L_2-composite activities) that follow a sequential order. Table 1 outlines the composite activities and their corresponding activity events. 10 Subjects (3 female, 7 male, aging from 23 to 37) were asked to build two book boxes. Their body size varies between 5.5 and 6.2 feet (1.68m to 1.89m). 10 hours of data were recorded while 20 book boxes were built.

Setup. The subjects were equiped with 5 Xsens MTx inertial measurement units. The units are mounted at the lower and upper arms and at the top back. Each sensor incorporates 3D-magnetometers, 3D-acceleration and 3D-gyroscopes. A fusion algorithm combines the sensory input

and estimates an absolute 3D-orientation with respect to a global coordinate system. Sampling was set to 50Hz. To allow for fine grained offline annotation video cameras were used to record activities.

Evaluation Procedure. We perform leave-one-out 10-fold cross-validation to evaluate user independent recognition performance. For a reduced training set, we perform repeated leave-one-out random-subsampling cross-validation. We repeat experiments 5 times.

Performance is measured using a one-vs-all classification scheme. Given a detected activity segment (event or composite activity) S and its label A, we use the matching criterion $start(A) \leq center(S) \leq stop(A)$. If the center of S falls in A, then S is considered to be a true positive. We measure performance by precision and recall obtained by thresholding posteriors. Unless otherwise noted, we summarize results by the equal error rate (EER), which equals the break even point of precision and recall.

4.2 Results

Fig. 4 shows the average EER both for the partonomy approach as well as for the direct approach without representing any hierarchical structure. When reducing training data

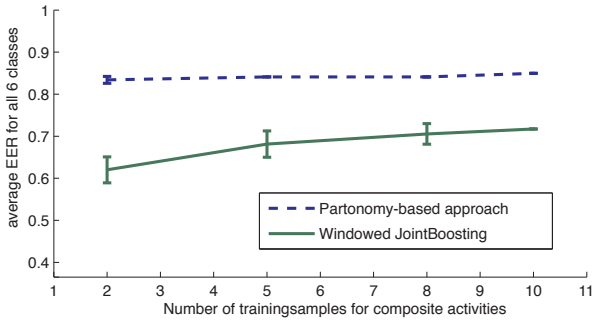

Figure 4. Overall results for L_2-composite activities on *Bookshelf*.

from 10 to 2 subjects for composites, performance remains almost constant for the partonomy-based approach, while the direct approach degrades about 10%. This can be explained by the fact that the direct approach has to learn both the variability of L_1-events as well as of the L_2-composite activities from limited training data resulting in an unrepresentative model. The partonomy model can profit from low variance of composite activities and its prior knowledge about activity events allowing almost constant performance. Moreover, we observe that the window size plays a significant role for the direct approach. As activity duration differs between subjects it is unclear what the optimal size would be. Note again, that our partonomy approach is independent of duration in *timeframes* and completely based on a *sparse* event-space representation, which gains robustness with respect to such time-variability in contrast to sliding window approaches.

Looking at the performance for individual composite activities, we observe that performance degrades marginally when reducing from 10 to 2 training samples for: *making backpart* (95% to 92%), *assembling box* (90% to 88%) and *hanging up box* (100% to 95%). *Create template* yields consistently worse result with 58% EER. It suffers from bad recognition (with respect to false positives) of its underlying L_1-activity events *mark template* (28%), *mark holes* (42%), *cut template* (30%). For *fix side parts* and *join side parts* performance remains constant at 88%, respectively 78%.

Using L_2 topdown-knowledge detection of L_1 activity events can be improved from 70% to 81% EER on average.

We also experimentally reduced the amount of training data for activity events for our partonomy approach. Results dropped significantly and we conclude, that 10 person training is minimal to capture the variability of activity events.

4.3 Discussion

We conclude from the experiments that the partonomy-based approach yields better results than a direct approach that does not consider partonomy of activities. This concords to previous work using layered representations for complex activities [22]. By transferring knowledge about underlying events, training data for composite activities can be reduced with marginal effect on performance. Interestingly, results are still surprisingly good using a direct approach. On this dataset activity events are rarely shared, therefore composite activities themselves are highly discriminative which won't be the case for larger numbers of composite activities. More importantly, while the direct approach yields acceptable results, it is difficult to use its acquired knowledge for unseen and new composite activities. Using a multi-layered approach we can reuse activity events and learn new compositions. Next, we show results of composite activity recognition on the dataset *Mirror* using knowledge transfer of L_1-activity events from the *Bookshelf* dataset.

5 Knowledge Transfer to Mirror-Dataset

Previously we showed that direct and partonomy-based approaches work well. Now, our specific wish is to reuse activity models and to reduce hereby the need of re-training *new* composite activities.

However, when being confronted with new composite activities, direct approaches have to be retrained from scratch. To recognize new composites consisting of shared activity events, knowledge about the underlying event layer is required. In this case partonomy-based approaches can be exploited. By transferring knowledge about L_1-activity-events, our partonomy-model enables us to learn new L_2-, ..., L_n-composite activities.

To investigate the possibilities of transferring knowledge of shared L_1-activity events, a second dataset called *Mirror* is recorded. Fig. 6 shows such a mirror-rack. The mirror is

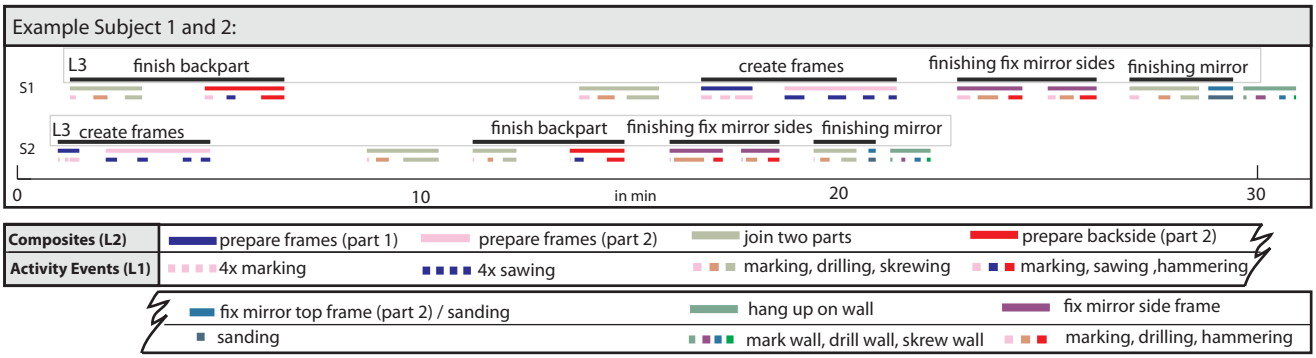

Figure 5. Example for 2 subjects for *Mirror* dataset

framed and equipped with a shelf. To mount it on the wall a (single) holder similar to the *bookshelf* is mounted.

The dataset consists of 10 L_1-activity events, which are combined into 6 L_2-composed activities. As we are interested in certain events only, a large null-class is present. Examples for ignored activities include but are not limited to *grabbing tools*, *moving parts*, *fixing parts using clamp* etc. The average ratio between each activity vs. background is 1:280. While the underlying L_1-activity-events (e.g., *drilling*, *screwing*) are similar to building a bookshelf, they constitute a rather different L_2-composition. Table 2 lists

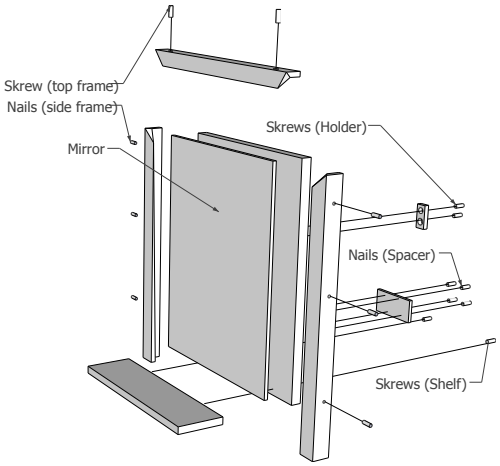

Figure 6. Explosion drawing of Mirror

activities for layer L_1 and L_2 and a third activity layer L_3. In comparison with L_1- and L_2-activities from the Bookshelf (Table 1), one can see differing L_2-composites but using shared L_1-events across both datasets. One L_2-activity only is completely shared (*hang up box/mirror*). Fig. 5 illustrates L_1-, L_2- and L_3-activities for two exemplary subjects. It can be also seen how composite activities differ strongly in length across subjects.

6 subjects (1 female, 5 males aging between 27 and 32) were asked to build such a mirror-rack. Their body size varies between 5.2 and 6.2 feet (1.60m, 1.89m). None of the test subjects has experience in wood engineering

L_2-Composite Activity	**L_1-Activity Events**
Prepare Frames (part 1)	4x Marking
Prepare Frames (part 2)	4x Sawing
Join two Parts	Marking–Drill–Screw
Fix mirror side frame	Marking–Drill–Hammering
Fix mirror top frame (part 2)	*Sanding*
Prepare Backside (part 2)	Marking–Saw–Hammering
Hang Up Box	Mark–Drill–Screw–Hang-Up Box
L_3-Composite Activity	**L_2-Activity Events**
Create frames	Prepare Frames (part 1)–Prepare Frames (part 2)
Finish Backpart	Join two parts–Prepare backside (part 2)
Finishing Mirror sides	2x Fix mirror side frame
Finishing Mirror	Join two parts–sanding

Table 2. 6 L_2-composite activities and the corresponding L_1-activity events. L_3-composite activities and the corresponding L_2-composites for *Mirror*

or is related to activity recognition research. The dataset was recorded one year after the *bookshelf*-experiment, conducted by a different researcher on different subjects. We recorded roughly one subject per day. Each subject was told about the L_2-composites and shown which tools to use and which L_1-steps to take.

5.1 Results

The experimental setup equals the setup from the Bookshelf dataset. We specifically show performance by transferring knowledge of activity events from the Bookshelf to the Mirror dataset. On average we achieve an EER of 63% for L_1-activity events. Fig. 7 shows results on L_1-activity events of the Mirror dataset in more detail. The red line shows the precision recall curve without using knowledge of the partonomy created by L_2-composite activities. For *sawing*, and *drilling* good performance is achieved (96%, 85%). Impressively, *screwing at wall* is immediately recognized with 100% precision and recall. While *screwing, hammering, drill wall* perform above 50% EER, *hang up box, sanding and marking* drop below 50%.

Given L_1 activity events we now model *new* L_2 composites differing from the previous dataset *Bookshelf*. Learning L_2 is done by leave-one-out 6-fold-cross-validation. However, as we have seen in the prior experiment we can reduce the amount of training data to a minimum of two sam-

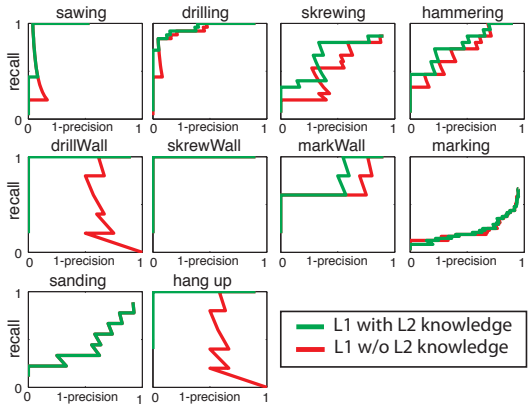

Figure 7. Results for activity events (L1) for *Mirror* dataset

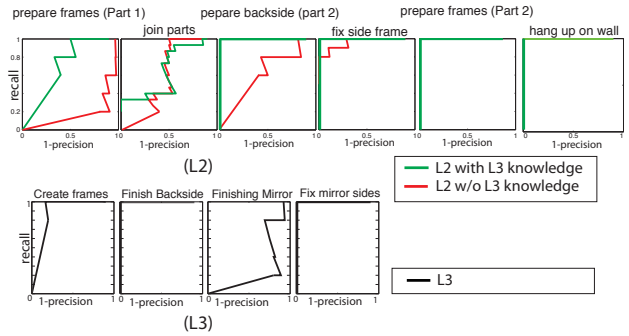

Figure 8. Results for composites L2 and L3 for *Mirror* dataset

ples with marginal decrease of performance. On average L_2 composite activities are recognized with an EER of 71%. Fig. 8 shows results per L_2-composite activity. For three activities (*fix side frame, prepare frames (part 2), hang up on wall*) we achieve almost perfect results. Activities *prepare frames part 1, join parts, prepare backside part 2* are less recognized. An inherent requirement is a reasonable quality of underlying activity events: *prepare frames part 1* uses $4\times$ *marking only*, which is detected with low performance (26% EER). We experienced *join two parts* to be confused with similar composite activities, containing shared activity events (*marking, drilling*) such as *fix mirror side frames*.

As indicated above we consider a third layer L_3. We construct this partonomy by creating recursively a partonomy using L_2-composite activities as events of L_3-activity composites. On average L_3-activities are recognized with an EER of 79%. Fig. 8 shows results on the individual L_3-composite activities. 3 out of 4 activities are perfectly recognized. Only for activity *finishing the mirror*, recognition performance is low. This L_3-composition contains L_2-activities *sanding* and *join two parts*, both not well recognized. As such, their partonomy is not well recognized.

Next we show improved recognition results using *top-down* knowledge from L_2 to L_1-activities, respectively from L_3 to L_2 . For L_1-activity events EER can be improved by 12% to 75% using the L_2 partonomy. As the green line in Fig. 7 shows, we can reduce the amount of false positives for several activities. For *drill wall, hang up box* results improve to 100%. Only for *marking* recognition performance cannot be improved. Note, that *sanding* is not contained in a L_2-composite, but used directly as L_2 activity event in the L_3-composite activity *finishing mirror*.

When incorporating information of the L_3-partonomy in the detection of L_2-activity events, we can improve three L_2-activities. On average performance is improved by 17% to 88% EER. Here *prepare frames (part 1)* profits most of the L_3-partonomy. In combination with perfectly recognized *preparing frames (part 2)* false positives of *preparing frames (part 2)* can be reduced to a large extent.

5.2 Discussion

According to our experiments, composite activity recognition across *different* datasets containing *differing* composites can be efficiently approached by our partonomy-based model. By storing (transferable) knowledge about L_1-activity events we can recognize L_2-, respectively L_3-composite activities with minimal training data. While transfer shows feasible results, difficulties still remain: different tools, users and even screws or nails influence recognition performance. Interestingly, having an imperfect activity event detection, composite activity yield impressive results with increasing performance per layer. Moreover, by recognizing composite activities we can go top down and increase performance of recognizing L_1-activity events significantly. Experimental tests of varying the amount of training data for activity events showed that our model requires a minimal quantity of events in order to recognize composite activities. Large variability of executing activity events demand large amount of training data. However, we showed that we can reuse this data, together with a minimal amount of training data for new composite activities.

6. Conclusion and Future Work

This paper presents a new partonomy-based approach to recognize composite activities. Modeling composite activities as partonomy has several advantages. First, our layered model using CRFs lends itself to transfer knowledge. We showed that transfer of L_1-activity events is possible across different datasets. Hereby, we were able to learn new L_2-composites by minimal learning of event combination. Second, by encapsulating variability of activity events in layer L_1, our model outperforms a direct approach of recognizing L_2-composites, which suffers from variability of underlying L_1 activities. Third, considering a partonomy we can

improve activity events by backprojecting knowledge in a top-down fashion from an L_n-layer to its lower L_{n-1} layer.

In this work, we focused on sequentially ordered events, which frequently appear in a temporally local timespan. In order to be able to capture higher level activities with stronger irregularities with respect to their partonomy, we will extend our model by grammar-oriented descriptions. Consequently, we plan to investigate less structured domains, such as activities of daily living.

Additionally, we will investigate automatic structure learning (i.e., not priorly specifying relationships) and use all layers jointly for training. Constrastly, prior high level knowledge from activity models, e.g., grammatical representation, can be used to recognize complex activity constellations. We would like to analyze the trade off between the amount of training data and high level knowledge that helps learning structure.

7 Achkowledgements

This work was partially funded by the DFG research training group "Topology of Technology" and the DFG project "Methods for Activity Spotting With On-Body Sensors". We acknowledge J. M. Mooij, A. Windsor and A. Globerson for providing their framework [9] implementing inference methods for graphical models. We also thank Christian Wojek and Paul Schnitzspan for sharing their knowledge on conditional random fields.

References

[1] O. Amft, C. Lombriser, T. Stiefmeier, and G. Troster. Recognition of user activity sequences using distributed event detection. *EuroSSC*, 2007.

[2] M. Bächlin, K. Förster, and G. Tröster. SwimMaster: a wearable assistant for swimmer. In *Ubicomp*, 2009.

[3] L. Bao and S. S. Intille. Activity recognition from user-annotated acceleration data. *Pervasive 2004*.

[4] U. Blanke and B. Schiele. Daily routine recognition through activity spotting. In *LoCA*, 2009.

[5] A. Bobick. Movement, activity and action: the role of knowledge in the perception of motion. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 352(1358):1257, 1997.

[6] G. H. Bower, J. B. Black, and T. J. Turner. Scripts in memory for text. *Cognitive Psychology*, 11(2):177 – 220, 1979.

[7] H. Bulthoff, S. Edelman, and M. Tarr. How are three-dimensional objects represented in the brain? *Cerebral Cortex*, 5(3):247, 1995.

[8] S. Consolvo, P. Klasnja, D. McDonald, D. Avrahami, J. Froehlich, L. LeGrand, R. Libby, K. Mosher, and J. Landay. Flowers or a robot army?: encouraging awareness & activity with personal, mobile displays. In *Ubicomp*, 2008.

[9] J. M. M. et al. libDAI 0.2.2: A free/open source C++ library for Discrete Approximate Inference. http://www.libdai.org/.

[10] D. Hoiem, A. Efros, and M. Hebert. Putting objects in perspective. *IJCV*, 80(1):3–15, 2008.

[11] D. H. Hu et al. Real world activity recognition with multiple goals. In *UbiComp*, 2008.

[12] T. Huynh, M. Fritz, and B. Schiele. Discovery of activity patterns using topic models. In *UbiComp 2008*.

[13] T. Huynh and B. Schiele. Analyzing features for activity recognition. In *EUSAI05*, Grenoble, France.

[14] J. Lafferty, A. McCallum, and F. Pereira. Conditional random fields: Probabilistic models for segmenting and labeling sequence data. In *Machine Learning*, 2001.

[15] J. Lester, T. Choudhury, and G. Borriello. A practical approach to recognizing physical activities. *Pervasive Computing*, pages 1–16, 2006.

[16] J. Lester, T. Choudhury, N. Kern, G. Borriello, and B. Hannaford. A hybrid discriminative/generative approach for modeling human activities. In *IJCAI*, 2005.

[17] L. Liao, D. Fox, and H. Kautz. Location-based activity recognition. In *NIPS*, 2005.

[18] P. Lukowicz, J. A. Ward, H. Junker, M. Stäger, G. Tröster, A. Atrash, and T. Starner. Recognizing workshop activity using body worn microphones and accelerometers. *Pervasive Computing*, 2004.

[19] K. Murphy, A. Torralba, and W. T. Freeman. Using the forest to see the trees: A graphical model relating features, objects, and scenes. In *NIPS*, 2003.

[20] K. Nelson. Generalized event representations: Basic building blocks of cognitive development. *Advances in Developmental Psychology: Volume 1*, page 131.

[21] N. Oliver and F. Flores-Mangas. HealthGear: a real-time wearable system for monitoring and analyzing physiological signals. In *BSN*, 2006.

[22] N. Oliver, E. Horvitz, and A. Garg. Layered representations for human activity recognition. In *ICMI02*.

[23] T. Starner, J. Weaver, and A. Pentland. Real-time american sign language recognition using desk and wearable computer based video. *PAMI*, 20(12):1371–1375, 1998.

[24] E. Tapia, S. Intille, and K. Larson. Activity recognition in the home using simple and ubiquitous sensors. In *Pervasive04*.

[25] M. Tarr. Rotating objects to recognize them: A case study of the role of mental transformations in the recognition of three-dimensional objects. *Psychonomic Bulletin and Review*, 2:55–82, 1995.

[26] A. Torralba, K. Murphy, and W. Freeman. Sharing visual features for multiclass and multiview object detection. In *CVPR04*.

[27] D. Vail, J. Lafferty, and M. Veloso. Feature selection in conditional random fields for activity recognition. In *IROS07*.

[28] T. van Kasteren, G. Englebienne, and B. Kröse. Transferring Knowledge of Activity Recognition across Sensor Networks. *Pervasive*, 2010.

[29] T. van Kasteren, A. Noulas, G. Englebienne, and B. Kröse. Accurate activity recognition in a home setting. In *UbiComp08*.

[30] T. Westeyn, K. Vadas, X. Bian, T. Starner, and G. Abowd. Recognizing mimicked autistic self-stimulatory behaviors using hmms. In *ISWC05*.

[31] J. Zacks and B. Tversky. Event structure in perception and conception. *Psychological Bulletin*, 127(1):3–21, 2001.

[32] V. Zheng, D. Hu, and Q. Yang. Cross-domain activity recognition. In *Ubicomp*, 2009.

[33] A. Zinnen, C. Wojek, and B. Schiele. Multi activity recognition based on bodymodel-derived primitives. In *LoCA09*.