A Hierarchical Representation for Human Activity Recognition with Noisy Labels

Hu, N.; Englebienne, G.; Lou, Z.; Kröse, B.

DOI
10.1109/IROS.2015.7353719

Publication date
2015

Document Version
Author accepted manuscript

Published in
IROS Hamburg 2015 conference digest

A Hierarchical Representation for Human Activity Recognition with Noisy Labels

Ninghang Hu[1,2], Gwenn Englebienne[2], Zhongyu Lou[2] and Ben Kröse[2,3]

***Abstract*— Human activity recognition is an essential task for robots to effectively and efficiently interact with the end users. Many machine learning approaches for activity recognition systems have been proposed recently. Most of these methods are built upon a strong assumption that the labels in the training data are noise-free, which is often not realistic. In this paper, we incorporate the uncertainty of labels into a max-margin learning algorithm, and the algorithm allows the labels to deviate over iterations in order to find a better solution. This is incorporated with a hierarchical approach where we jointly estimate activities at two different levels of granularity. The model is tested on two datasets, *i.e.,* the CAD-120 dataset and the Accompany dataset, and the proposed model shows outperforming results over the state-of-the-art methods.**

I. Introduction

Activity recognition system is an important component for robots. For example, a robot can give suggestions of a daily routine to an elderly person by monitoring her activities [1]. A robot can also provide accurate assistance to the user by observing the activity that the user is currently performing [2]. Although many prior works have focused on this topic, recognizing complex activities remains a challenging task that still needs to be solved. In this paper, we propose a hierarchical model that models activities with different complexity, and the model is incorporated with a max-margin learning algorithm that allows labels to have uncertainty.

In our approach we refer to the low-level representation as *actions*, which are defined as the atomic movements of a person that relate to at most one object in the environment, *i.e.,* reaching, placing, opening, and closing. Most of these actions are completed in a relatively short period of time. In contrast, *activities* refer to a complete sequence that is composed of different *actions*. The experiments on the CAD-120 dataset demonstrate that it is beneficial to add a low-level representation of the activities [3].

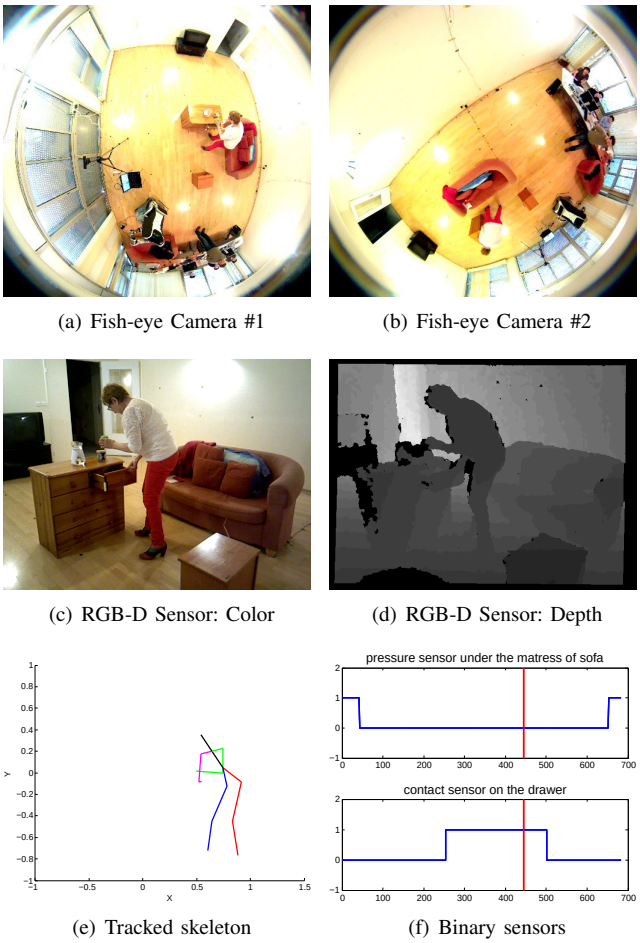
A first characteristic of our approach is the way we use the labels in the training. The approach is based upon our earlier work [4] on recognizing actions, where we used a layer of latent variables to improve the model expressiveness. However, in this paper, we build a hierarchical approach where the labels are jointly estimated by considering the interaction between activities and actions.

The research is funded by the European project ACCOMPANY (grant agreement No. 287624) and the European project MONARCH (grant agreement No. 601033).

[1] N. Hu is with Electrical Engineering and Computer Sciences (EECS), University of California, Berkeley, USA nh@berkeley.edu

[2] G. Englebienne, Z. Lou and B. Kröse are with Intelligent System Lab Amsterdam, University of Amsterdam, 1098XH Amsterdam, The Netherlands {g.englebienne,z.lou,b.j.a.krose}@uva.nl

[3] B. Kröse is also with the Amsterdam University of Applied Science



(a) Fish-eye Camera #1 (b) Fish-eye Camera #2

(c) RGB-D Sensor: Color (d) RGB-D Sensor: Depth

(e) Tracked skeleton (f) Binary sensors

Fig. 1. Data that are collected in the Accompany dataset. The dataset consists of images from the fish-eye images ((a), (b)) captured by the Care-O-bot 3 [5] RGB-D images ((d) and (c)), and (f) binary data from the ambient sensors. (e) Human skeletons are extracted based on the RGB-D data.

A second issue concerns the labels of the actions and activities in the dataset. The uncertainty of the labels in the CAD120 (also for TUM-Kitchen [6] and CAD-60 [7]) are unknown. We know that labeling data is a very subjective work and sometimes the annotators may make mistakes which add noise into the labels. Treating these noisy labels as the *ground truth* is very harmful for most learning methods. To deal with that, Hu et al. proposed to use the soft labeling [8], a method that models each label as a multinomial distribution rather than deterministic. They assume noise is located during the transition of *action*, and the soft label encodes the action both before and after the transition segment. In this

paper, in contrast, we treat all of the labels as noisy data, and we add small probabilities to the *incorrect* labels (*i.e.,* labels that are different from the annotation we have), so that the model can converge to a better representation of the actions. Moreover, when learning model parameters, our method treats the action nodes as latent variables so that they evolve over iterations.

The third issue concerns the dataset. The CAD-120 is collected with a single modality sensor (*i.e.,* an RGB-D camera), and actors who performed in the dataset are about the same age. For realistic scenarios, however, it is crucial to combine different sensors in order to increase the robustness of the systems, and the target users should not be limited to a certain age group. In this paper, we introduce the Accompany dataset (Fig. 1), which consists of data that are collected by multiple types of sensors, *i.e.,* two ceiling-mounted fish-eye cameras, one RGB-D sensor, one magnetic contact sensor, and one pressure sensor. The Accompany dataset contains four subjects, including one elderly person, performing daily activities in an apartment. Labels are annotated by multiple human annotators, and the uncertainty can be easily generated from their voting.

In summary, the contributions of this paper are two-fold. Firstly, we propose a hierarchical approach that jointly models activities and actions, and the model is incorporated with soft labeling. This allows the learning algorithm to take labeling uncertainty into consideration, so that the labels can converge to a better representation. Secondly, we present the Accompany dataset, a challenging dataset where activities are performed by people with different age group. The labels are annotated by multiple persons and from that we can obtain the labeling uncertainly. The data are collected with multimodality sensors, which enable the research of data fusion.

II. Related Works

In this section, we review the recent works on human activity recognition in robot related scenarios along with the datasets that have been used for evaluation.

The early approaches recognize human activities directly based on the video data, *e.g.,* Qian et al. [9] extract shape features based on the color images, and the activities are classified with a multi-class Support Vector Machine (SVM). Kasteren et al. [10] use a collection of binary sensors for estimating a sequence of activities. Recently, many approaches show that the activity recognition performance can be improved by adding a layer of low-level representation of the activities. Sung et al. [7] use the skeleton joints of a person for activity classification. The skeleton joints are detected by OpenNI using a RGB-D sensor. They proposed a hierarchical approach that models activities with two layers. The upper layer is a sequence of actions and the temporal dynamics of the sequence are captured by the model parameters. The lower layer consists of a sequence of latent variables for capturing the sub-level representation of the activities. Instead of treating actions as latent variables, Koppula et al. [3] explicitly defined the action layer by hand labeling. In addition, the objects are associated with several affordance labels that characterize properties of the objects, such as whether objects can be reached, moved, or placed. They model the human-object interaction using Conditional Random Fields (CRFs). The actions and objects are represented as separate nodes, and these are connected with undirected edges in the graph. Similar to [7], Hu et al. [4] also added a layer of latent variables, but their latent variables are used to describe the types of actions, and there is only one latent node associated with each video segment. The model parameters are learned with the latent Structured SVM [11], which is a generalized version of SVM that allows having structured outputs. These three approaches focus on modeling actions only, and the activities are predicted only after the action sequences are obtained. Similar to the idea of modeling human-object interaction [3], we propose an approach that models the interaction between actions and activities in a hierarchical framework, where the action labels and the activity labels are jointly estimated. To model the labeling uncertainty, our approach adopts the method of soft labeling [8], where we add a small portion of noise to the labels and each label is converted to a multinomial distribution over all possible labels. This helps with the situation where the labels provided by the dataset contain small errors. The proposed model combines the soft labeling approach with the hierarchical approach, and the model shows outperforming results over the state-of-the-art approaches.

There are a few public datasets available for evaluating the activity recognition systems. The TUM-Kitchen dataset [6] is recorded in a home-care scenario where subjects perform a few daily activities in a kitchen. The kitchen is equipped a set of ambient sensors (*i.e.,* multiple RFID tag reader and magnetic sensors) and four static overhead cameras. The full body joints are tracked with a motion capturing system [12]. The CAD-120 dataset [3] contains data that are captured with an RGB-D sensor. Based on the depth and color images of the RGB-D sensor, both human skeletons joints and objects are detected and tracked. The dataset consists of two layers of activities, and a number of recent approaches have been evaluated on this dataset. Labels in both TUM-Kitchen and CAD-120 are manually labeled, which may contain errors. We collect a new dataset that uses multiple types of sensors, including overhead cameras, binary sensors, and RGB-D sensor. The dataset is annotated by multiple human annotators, and these annotations can be transformed into labeling uncertainty for learning.

III. Model Formulation

The goal of our task is to recognize human activities, denoted as a random variable y, based on the data that are collected from different sensors (denoted as x). The sensory data changes over time and they can be represented as a spatial-temporal volume, where each slide of the volume contains the sensor measurements at a certain time. Because these measurements are usually sampled at a very high frequency, data are almost identical within a short interval

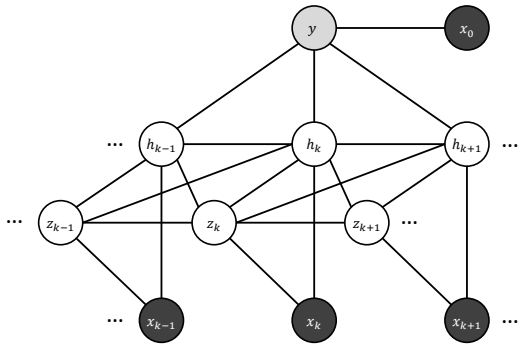


Fig. 2. The graphical representation of our model. y is the target variable we need to estimate, *i.e.,* the activity. h and z are two types of latent variables. h refers to the actions and z refers to the types of actions. x represents the observed data. x_0 represents the global features extracted based on the whole video. The "global" features x_0 is defined in contrast to the features that are "locally" extracted from one segment (*i.e.,* x_k).

of time. Therefore we temporally partition the input volume into a sequence of segments. Each segment is a concatenation of all the data, *i.e.,* x_k that are measured within the time segment k. Let $\phi(x_k)$ be a function that maps the raw input of each segment into a feature space.

In order to increase the model expressiveness, we add two types of latent variables to the model, *i.e.,* h and z. Each segment is associated with the latent state pair (h_k, z_k).

Fig. 2 shows how these variables are related. Formally, let us define the undirected graph as $\mathcal{G} = (\mathcal{V}, \mathcal{E})$, where a node corresponds to the random variables, and an edge measures the compatibility between the nodes. Specifically, Fig. 2 shows three types of nodes. The un-shaded nodes are the random variables that are unconstrained. The nodes in grey represent the target variables that we would like to estimate. The target node is observed during training but unknown during testing. The black nodes represent the observed data, which are fixed during both training and testing.

Given the graph $\mathcal{G}$, we can write the following function by summing over potentials of all the cliques in the graph. Note that each clique is a fully connected sub-graph.

$$\Psi(x,h,z,y) = \psi_1(x_0,y) + \sum_{k=1}^{K} \psi_2(x_k,h_k,z_k) \\ \sum_{k=2}^{K} \psi_3(h_{k-1},h_k,z_{k-1},z_k) + \sum_{k=2}^{K} \psi_4(h_{k-1},h_k,y) \tag{1}$$

where $\psi_i(\cdot), i \in \{1,2,3,4\}$ are potential functions in terms of the random variables h, z, and y, and it measures the compatibility of nodes that fall in the same clique. The detail definition of the potential functions is as follows.

The first potential term in (1) measures the score of a special clique that only contains two nodes, and it favors a certain configuration of the activity when observing the global features x_0.

$$\psi_1(x_0,y) = x_0^\intercal \cdot w(y) + b_1(y) \tag{2}$$

where both w and b are model parameters. $w(y)$ is a sub-set of parameters that are indexed by y, and $b_1(y)$ is a scalar that models the prior of seeing the activity y.

The second potential term ψ_2 models the score of the cliques (x,h,z). For segment k, the second potential is computed as

$$\psi_2(x_k,h_k,z_k) = \phi^\intercal(x_k) \cdot w(x_k,h_k,z_k) + b_2(h_k,z_k) \tag{3}$$

Similar to the first potential, $w(x_k,h_k,z_k)$ selects the parameters that correspond to the joint state (x_k,h_k,z_k), and b_2 is the bias term of this linear function.

The third potential term encompasses the latent states of two contiguous segments, and it measures the score of transitioning between two segments states

$$\psi_3(h_{k-1},h_k,z_{k-1},z_k) = w(h_{k-1},h_k,z_{k-1},z_k) \tag{4}$$

The fourth potential measures the score of the latent states and the activity.

$$\psi_4(h_{k-1},h_k,y) = w(h_{k-1},h_k,y) \tag{5}$$

By summing these potentials over time, we obtain the potential function of the entire graph $\Psi(x,h,z,y)$. For clearance, the bias parameters b_1 and b_2 can be merged with w, and now we can see that $\Psi(x,h,z,y)$ is a linear combination of feature functions.

IV. Inference and Learning

Now that the graphical structure of the model are known and we know how these random variable are related to each other in terms of the model parameters w, in this section, we will explain how to use such a graphical structure for predicting activities as well as how to learn the model parameters.

A. Inference

The goal of our task is to estimate the correct activity label based on the observed data. This is usually referred as an inference problem. In particular, we would like to solve the following equation:

$$(y^*,h^*,z^*) = \underset{h,z,y}{\operatorname{argmax}}\, \Psi(x,h,z,y) \tag{6}$$

Generally, solving (6) is an NP-hard problem that requires the evaluation of the objective function over an exponential number of state sequences. Considering the structure of cliques, however, we can efficiently decode the activity labels using the following equation:

$$V_k(y,h_k,z_k) = \psi_2(x_k,h_k,z_k) + \max_{(h_{k-1},z_{k-1})} \{\psi_3(h_{k-1},h_k,z_{k-1},z_k) + \psi_4(h_{k-1},h_k,y) + V_{k-1}(y,h_{k-1},z_{k-1})\} \quad (7)$$

where K is the number of segments in the video. The process is initialized with

$$V_1(y,h_1,z_1) = \psi_2(x_1,h_1,z_1) \quad (8)$$

The above function is evaluated iteratively across the entire sequence, and the optimal assignment of the last segment K can be computed as

$$y^*,h_K^*,z_K^* = \underset{y,h_K,z_K}{\operatorname{argmax}} V_K(y,y_K,z_K) + \boldsymbol{w}_5(y)\cdot\Phi(\boldsymbol{x}_0) \quad (9)$$

Knowing the optimal assignment at K, we can track back the best assignment in the previous time step $K-1$. The process continues until all h_k^* and z_k^* have been assigned, *i.e.,* the inference problem in (6) is solved.

B. Max-Margin Learning

Let $(x^{(1)},\pi^{(1)},y^{(1)}),\ldots,(x^{(N)},\pi^{(N)},y^{(N)})$ be a set of labeled training examples, where x is a volume of observed data, y is the activity label, and π corresponds to the soft labels. The soft labels satisfy the following constraint

$$\sum_a \pi_k(a) = 1 \quad (10)$$

$$\forall j:\ \pi_j(a) \geq 0 \quad (11)$$

where the value of $\pi_k(a)$ refers to the problem of segment k being labeled with activity a. Intuitively, we can consider π_k as a multinomial distribution over all possible labels, and it can be obtained through multiple human annotators. To learn our parameters, we apply a Structured-SVM framework [13]. The objective of learning is to optimize the following term with respect to w:

$$\min_w \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N} \Delta(\pi^{(i)},\hat{h},y^{(i)},\hat{y}) \right\} \quad (12)$$

where $\|w\|^2$ is a regularization term and $\Delta(\cdot)$ is a loss function that measures the distance between the prediction of the model and the ground truth labels. Specifically, the loss is computed in the following form:

$$\Delta(\pi^{(i)},\hat{h},y^{(i)},\hat{y}) = \lambda\mathbb{1}(y^{(i)}\neq\hat{y}) + \frac{1}{K}\sum_{k=1}^{K} 1-\pi_k^{(i)}(\hat{h}) \quad (13)$$

where $\mathbb{1}(\cdot)$ is a indicator function and that returns one when the condition is satisfied. λ is a weighting scalar that balances between the loss of activity labels and the loss of low-level labels.

Following [11], we introduce slack variables ξ into the objective function and use the Margin Rescaling Surrogate to reform (12) into a optimization problem that subjects to a finite set of linear inequality constraints

$$\min_{w,\xi} \left\{ \frac{1}{2}\|w\|^2 + C\sum_{i=1}^{N}\xi_i \right\} \quad (14)$$
$$s.t.\ \forall i \in \{1,2,\ldots,N\}:$$
$$\Psi(x^{(i)},h^*,z^*,y^{(i)}) - \Psi(x^{(i)},\hat{h},\hat{z},\hat{y}) \geq \Delta(\pi^{(i)},\hat{h},y^{(i)},\hat{y}) - \xi_i$$

where

$$h^*,z^* = \underset{h,z}{\operatorname{argmax}}\ \Psi(x^{(i)},h,z,y^{(i)}) \quad (15)$$

and

$$\hat{h},\hat{z} = \underset{h,z}{\operatorname{argmax}}\ \Delta(\pi^{(i)},h,y^{(i)},y) + \Psi(x^{(i)},h,z,y) \quad (16)$$

Note that both (15) and (16) can be solved efficiently using a variation of (6). The objective function of (14) is minimized using the Concave Convex Procedure (CCCP) [14]. This is an iterative algorithm that infers the latent variables h and z, and then optimize the model parameters using a linear-kernel Structured SVM. This process is repeated until the convergence.

C. Initialize Latent Variables

Before the learning algorithm starts, we need to initialize both of the latent variables, *i.e.,* h and z. Since h has its associated soft labels, we initialize the state of h by assigning the action labels that have the largest probability. In the case of draw, the labels are randomly picked. To initialize z, we apply K-means clustering based on the observations x, and assign latent states according to their cluster assignments.

V. Experiments and Results

Our model was evaluated on two datasets. We used the benchmark CAD-120 dataset to compare the performance of our new model with the state-of-the-art and with our previous work. Secondly we used the dataset that was collected in the EU project Accompany[1], *i.e.,* the *Accompany* dataset. This dataset was chosen because it was annotated by multiple annotators, resulting in uncertainty in the labels, and also because it contains data from multiple modalities.

In this section, we first describe the benchmark dataset CAD-120, and the performance of our model. Then we describe details of the Accompany dataset, and demonstrate the performance of the proposed model on the Accompany dataset.

A. Experiments on CAD-120 Dataset

The CAD-120 dataset [3] consists of 120 RGB-D videos that are performed by 4 subjects. The dataset consists of 10 *activities* (*i.e., making cereal, taking medicine, stacking objects, unstacking objects, microwaving food, picking objects, cleaning objects, taking food, arranging objects, having a meal*) and 10 *actions* (*i.e., reaching, moving, pouring, eating, drinking, opening, placing, closing, scrubbing, null*). Both the *activity* labels and the *action* labels have been manually

[1] http://accompanyproject.eu/

annotated. Both objects and human skeleton positions are tracked in the dataset.

The CAD-120 is very challenging dataset because the large variation of actions that are performed by the subjects, *e.g.,* The subjects include both left and right handed person, and they grasp objects in opposite ways. Even the same subject can perform the same action many very different ways, depending on the context of objects. *E.g.,* both *opening a bottle* and *opening a microwave* are the action *opening*, however, they may look quite different from each other in the input videos.

To make a fair comparison with the state-of-the-art approaches, we use the same training/testing splits as [3] and [4], *i.e.,* we train the model based on videos of 3 subjects and the model is tested on a *new* subject. The same set of input features is extracted. The model is evaluated with 4-fold cross-validation. Each cross-validation is restarted for 3 times in order to test the stability. We report the accuracy, recall, precision, and F-score of the test performance, and the standard errors are also included.

B. Results

Table I reports the results of the CAD-120 dataset under the ground truth segmentation. *Single Layer* refers to the approach where the activities labels are directly inferred based on the observations, and there is no intermediate action layer embedded in the model. We show that the single layer approach is significantly outperformed by the hierarchical approaches, suggesting the benefits of adding the low-level representation into the model. Our model (row 5 and 6) is evaluated under two labeling configuration. *Model Ori.* indicates the model where the original labels of the CAD-120 dataset are used, *i.e.,* hard labeling. In contrast, *model Ori.* refers to our soft labeling method, where we add noise to the data. We show that our best model is obtained by the soft labeling method, *i.e., our model (Soft)*, achieving 96.4 for precision and 95.0 for recall. Both our models show outperforming results over the state-of-the-art approach [15]. Notably, the precision is improved by 1.4 percentage points and the recall is improved by 1.7 percentage points.

The confusion matrix of our best model (*soft*) is shown in Fig. 3. We can see high responses on the diagonal of the matrix, indicating most of the activities are correctly classified. The most difficult ones are among "arranging objects", "stacking objects", and "unstacking objects". This is because these activities are all related to loops of *reaching*, *moving*, and *placing* a number objects, therefore they are similar in the low-level representation.

C. Experiments with Accompany Dataset

The Accompany dataset[2] consists of 4 subjects performing 8 daily activities in a living room. The image data contains 44 videos with in total 16531 image frames. Four subjects were asked to complete the activities based on their personal preferences, and there was no constraint in the way how the activities should be performed, *i.e.,* it is up to the subjects in which order the actions are performed. This makes a large variation in performing these activities, and the order of actions to perform the same activity is different among the subjects. The activities are *Taking Medicine*, *Drinking Water*, *Preparing Flowers*, *Placing Flowers*, *Reading Books*, *Watching TV*, *Having Cookie*, and *Making Phone Call*. The action labels are *reaching*, *moving*, *opening*, *closing pouring*,*eating*, *drinking*, *placing*, *holding*, *cutting*, *idle*. These labels are independently annotated by multiple annotators. With the labeling of multiple annotators, we can generate the soft labels, which incorporate the uncertainty of labeling. The same evaluation procedure was used as CAD-120.

1) Data Acquisition: The dataset contains 44 videos in total, and each video contains data that are recorded by the sensing system. The sensing system contains a list of sensors and cameras which can be used for recognizing human activities. Specifically, we have one RGB-D sensor (ASUS Xtion Pro Live) mounted on the head of the Care-O-bot 3, an additional RGB-D sensor for tracking human skeleton joints, two fish-eye cameras mounted on the ceiling to monitor the entire room, one pressure sensor under the

[2]The dataset will be publicly available at http://ninghanghu.eu/accompany_dataset.html.

TABLE I

PERFORMANCE OF ACTIVITY RECOGNITION DURING TESTING ON THE CAD-120 DATASET. THE RESULTS ARE REPORTED IN TERMS OF ACCURACY, PRECISION, RECALL AND F-SCORE. THE STANDARD ERROR IS ALSO REPORTED.

Results with Ground-truth Segmentation				
Methods	Accuracy	Precision	Recall	F1-Score
Single layer	74.2 ± 5.1	78.5 ± 4.7	73.3 ± 5.1	75.8 ± 4.9
Koppula et al. [3]	84.7 ± 2.4	85.3 ± 2.0	84.2 ± 2.5	84.7 ± 2.2
Hu et al. [16]	90.0 ± 2.9	92.8 ± 2.3	89.7 ± 3.0	91.2 ± 2.5
Koppula et al. [15]	93.5 ± 3.0	95.0 ± 2.3	93.3 ± 3.1	94.1 ± 2.6
Our Model (Ori.)	93.6 ± 2.7	95.2 ± 2.0	93.3 ± 2.8	94.2 ± 2.3
Our Model (Soft)	**95.2 ± 2.7**	**96.4 ± 2.1**	**95.0 ± 2.8**	**95.7 ± 2.3**

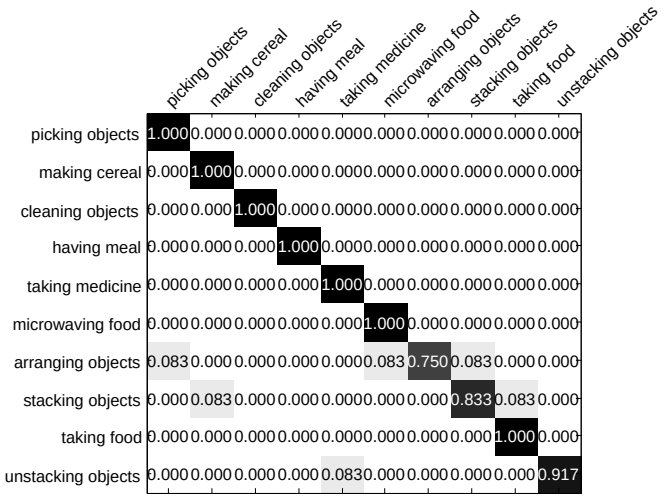


Fig. 3. The confusion matrix for activity classification on the CAD-120 dataset. The ground-truth segmentation is used.

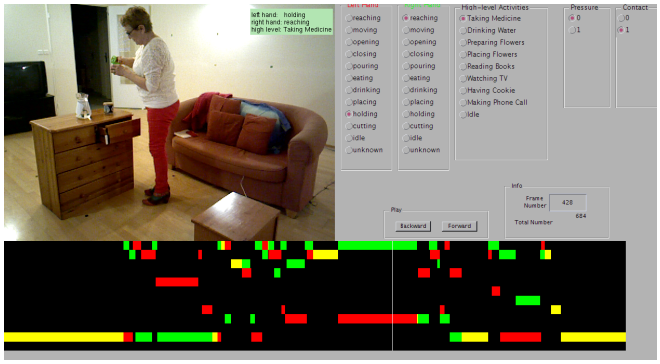


Fig. 4. GUI for Data Labeling

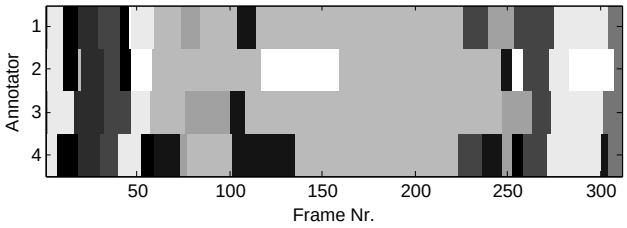


Fig. 5. Visualization of the action labels that are assigned by multiple annotators. Each row corresponds with the labeling results of one annotator. Note that the videos are randomly assigned to the annotators. In this case, there are only 4 annotators working on this part of the video.

sofa, and one contact sensor on the drawer, two laser scanners on the Care-O-bot 3 for the purpose of localization. All the sensors are connected into the central network system. We host the ROS core on the robot. Data from the sensors are converted to time-stamped ROS messages once being recorded. To overcome the limited writing speed of the hard drive, we distribute the data writing tasks to several hard SSDs to facilitate high speed of data writing. The data are recorded as ROS bag files, so they can be played back and visualized using the Rviz component in ROS.

2) Labeling: We invited seven annotators to independently label the activities and actions by watching the videos. None of the annotators has background of related background in activity recognition. The videos are randomly distributed to the annotators. We also added an additional option for the annotator. When action labels are not clear, the annotators are allowed to choose "unknown" to assign empty labels to the associated video frames.

Fig. 4 shows the Graphical User Interface (GUI) of the data annotation software. The user is allowed to choose separate action labels for both left and right hands. To navigate, the user can play the video forward and backward. The annotated activities are visualized under the video. The red color indicates activity from the left hand and green color for the right hand. The yellow color indicates that both hands share the same activity.

The grayscale image in Fig. 5 demonstrates the annotated action labels for a video. The intensity represents different action labels. We can see the assigned labels are quite different across different annotators. We deal with the disagreement by converting them into soft labels.

3) Feature Extraction: The local feature representation consists of three parts. The first part of features is extracted following [7]. We compute the rotation matrix of 9 upper joints (*i.e.,* head, neck, torso, shoulders, elbows, and hands) in the coordinate frame of the torso and head. The 3×3 rotation matrix is converted into 4-dimensional quaternion vector to obtain a more compact representation, giving us 72 features. To capture the posture of the person, *i.e.,* whether the person is standing, sitting, we compute the location of both feet relative to the torso, giving us 6 features. To distinguish whether a person is bending, we compute the angle of the upper body against the vertical plain, and this give us 1 feature. We compute the relative position of both hands with respect to the torso and head, giving us 12 features. The highest and lowest vertical positions of both hands are also computed for the last 6 frames, which gives us 4 features. The second set of features is extracted based on [3]. We compute the distance and displacement that the 9 upper body parts have moved within each segment, which gives us 18 features. Similarly, to capture the dynamics that cross the segments, we compute the distance that the upper joints has moved at the beginning, middle and end of the two consecutive segments (27 features) along with the maximal and minimal distance of all the frames (18 features). The third part encodes data of the binary sensors. We measure the most frequent value along with whether there is a switch in the segment, giving us 4 binary features for the two simple sensors.

The global features are used to capture features of the entire video, and they refer to x_0 in Fig. 2. We compute the total distance that the upper body joints move in each video, and the distances are normalized in order to be invariant to the video length. This gives us 9 features. For binary sensors, we compute the number of signal switches and the proportion of positive signal in the entire video for both sensors, giving us 4 features.

In addition to the skeleton features described above, we also extract the object features that capture the relation between a person and the objects. We refer to [3] for a detailed description of these features.

The final processing of these features is to discretize the data into 10 categories, and convert them using 1-of-N encoding. This pre-processing has two advantages. First, it reduces the effects of minor observation errors. Second, it encodes features into binary values. This is very beneficial for models like SVM, because they are known to be very sensitive to the scale of the data.

4) Results: We evaluate our model on the Accompany dataset to test how the model can be generalized on new data. Since the Accompany Data are labeled with multiple annotators, it is unfeasible to obtain the "ground truth" segmentation. Therefore we apply the uniform segmentation, where the videos are divided into segments with a fixed length (20 frames). Note that here we only use the uniform distribution to generate the baseline results for the

TABLE II

TEST PERFORMANCE OF THE PROPOSED MODEL ON THE ACCOMPANY DATASET. THE RESULTS ARE REPORTED IN TERMS OF ACCURACY, PRECISION, RECALL AND F-SCORE. THE STANDARD ERROR IS ALSO REPORTED.

CAD-120 Dataset with Uniform Segmentation				
Features	Accuracy	Precision	Recall	F1-Score
skeleton only	45.6 ± 6.0	45.8 ± 5.6	46.5 ± 5.9	46.2 ± 5.8
skeleton+object	70.2 ± 5.2	69.3 ± 2.7	67.9 ± 2.0	68.6 ± 2.4
Accompany Dataset with Uniform Segmentation				
skeleton only	45.9 ± 9.6	42.5 ± 9.3	45.3 ± 9.7	43.9 ± 9.5
skeleton+sensor	55.7 ± 8.8	53.3 ± 7.9	52.6 ± 8.2	53.0 ± 8.1

Accompany dataset. But we note that the performance can be improved by using other segmentation methods with heuristics, such as [17].

Table II shows the performance of our model (*soft*) on both datasets. We first evaluate the model only with the skeleton features. Then we add the additional features (binary sensor features for Accompany and object features for CAD-120). When there are only skeleton features, the model produce similar performance on the two datasets. By fusing with the other cues, however, the performance is improved significantly. Notably, the F1-score on CAD-120 has a gain of 22 percentage points after adding object information. In contrast, the F1-score on Accompany is increased by around 9 percentage points after adding the sensory features. We note this can be further improved by incorporating object information into the features.

VI. CONCLUSION

In this paper, we present a hierarchical approach that jointly estimates activities and actions, and the model parameters are learned with a max-margin approach. We assume the provided labels are noisy. By converting them into soft labels, our model shows outperforming results over the state-of-the-art approaches on the CAD-120 dataset.

A new dataset was collected for evaluating our model. The dataset contains multimodal data, and it is quite challenging because one of four subjects is a real elderly person. The dataset is labeled by multiple annotators, which allows the annotations having a natural conversion to soft labels. The baseline results on the Accompany dataset shows the benefits of fusing different cues for activity recognition, and the performance suggests that the Accompany dataset can be further improved by adding the object features.

REFERENCES

[1] A. N. Aicha, G. Englebienne, and B. Kröse, "How lonely is your grandma?: detecting the visits to assisted living elderly from wireless sensor network data," in *Workshop on Human Factors and Activity Recognition in Healthcare, Wellness and Assisted Living, UbiComp'13*. ACM, 2013, pp. 1285–1294. [Online]. Available: http://dl.acm.org/citation.cfm?id=2497283

[2] F. Amirabdollahian, S. Bedaf, R. Bormann, H. Draper, V. Evers, J. G. Pérez, G. J. Gelderblom, C. G. Ruiz, D. Hewson, N. Hu, and Others, "Assistive technology design and development for acceptable robotics companions for ageing years," *Paladyn, Journal of Behavioral Robotics*, pp. 1–19, 2013.

[3] H. S. Koppula, R. Gupta, and A. Saxena, "Learning Human Activities and Object Affordances from RGB-D Videos," *International Journal of Robotics Research (IJRR)*, vol. 32, no. 8, pp. 951–970, 2013.

[4] N. Hu, G. Englebienne, Z. Lou, and B. Kröse, "Learning Latent Structure for Activity Recognition," in *Proceedings of IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2014, pp. 1048–1053. [Online]. Available: http://dx.doi.org/10.1109/ICRA.2014.6906983

[5] U. Reiser, C. Connette, J. Fischer, J. Kubacki, A. Bubeck, F. Weisshardt, T. Jacobs, C. Parlitz, M. Hagele, and A. Verl, "Care-O-bot 3-Creating a product vision for service robot applications by integrating design and technology," in *Proceedings of the IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS)*. IEEE, 2009, pp. 1992–1998.

[6] M. Tenorth, J. Bandouch, and M. Beetz, "The TUM kitchen data set of everyday manipulation activities for motion tracking and action recognition," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009*, 2009, pp. 1089–1096.

[7] J. Sung, C. Ponce, B. Selman, and A. Saxena, "Human Activity Detection from RGBD Images," in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)*. IEEE, 2011, pp. 842–849. [Online]. Available: http://arxiv.org/abs/1107.0169

[8] N. Hu, Z. Lou, G. Englebienne, and B. Kröse, "Learning to Recognize Human Activities from Soft Labeled Data," in *Robotics: Science and Systems (RSS)*, 2014. [Online]. Available: http://www.roboticsproceedings.org/rss10/p03.html

[9] H. Qian, Y. Mao, W. Xiang, and Z. Wang, "Recognition of human activities using SVM multi-class classifier," *Pattern Recognition Letters*, vol. 31, pp. 100–111, 2010.

[10] T. Van Kasteren, A. Noulas, G. Englebienne, and B. Kröse, "Accurate activity recognition in a home setting," in *Proceedings of the International Conference on Ubiquitous Computing*. ACM, 2008, pp. 1–9.

[11] C. N. Yu and T. Joachims, "Learning structural SVMs with latent variables," in *Proceedings of the International Conference on Pattern Recognition (ICPR)*. ACM, 2009, pp. 1169–1176.

[12] J. Bandouch and M. Beetz, "Tracking humans interacting with the environment using efficient hierarchical sampling and layered observation models," in *2009 IEEE 12th International Conference on Computer Vision Workshops, ICCV Workshops 2009*, 2009, pp. 2040–2047.

[13] I. Tsochantaridis and T. Hofmann, "Support vector machine learning for interdependent and structured output spaces," in *Proceedings of the International Conference on Machine Learning (ICML)*. ACM, 2004, p. 104.

[14] A. Yuille and A. Rangarajan, "The concave-convex procedure (CCCP)," *Advances in neural information processing systems (NIPS)*, vol. 2, pp. 1033–1040, 2002.

[15] H. Koppula and A. Saxena, "Learning Spatio-Temporal Structure from RGB-D Videos for Human Activity Detection and Anticipation," *Proceedings of the International Conference on Machine Learning (ICML)*, 2013.

[16] N. Hu, G. Englebienne, and B. Kröse, "A Two-layered Approach to Recognize High-level Human Activities," in *Proceedings of the IEEE International Symposium on Robot and Human Interactive Communication (ROMAN)*. IEEE, 2014, pp. 243–248. [Online]. Available: http://dx.doi.org/10.1109/ROMAN.2014.6926260

[17] M. Hoai, Z. Lan, and F. D. la Torre, "Joint segmentation and classification of human actions in video," in *Proceedings of the IEEE conference on Computer Vision and Pattern Recognition (CVPR)*, 2011, pp. 3265–3272.