The DAily Home LIfe Activity Dataset: A High Semantic Activity Dataset for Online Recognition

Geoffrey Vaquette[1,2], Astrid Orcesi[1], Laurent Lucat[1] and Catherine Achard[2]
[1] CEA, LIST, Vision and Content Engineering Laboratory, Point Courrier 173, F-91191 Gif-sur-Yvette, France
[2] Sorbonne University, UPMC Univ Paris 06, CNRS, UMR 7222, ISIR, F-75005, Paris, France

***Abstract*— In this article, we introduce the *DAily Home LIfe Activity (DAHLIA) Dataset*, a new dataset adapted to the context of smart-home or video-assistance. Videos were recorded in realistic conditions, with 3 Kinect^{TM}v2 sensors located as they would be in a real context. The very long-range activities were performed in an unconstrained way (participants received few instructions), and in a continuous (untrimmed) sequence, resulting in long videos (39 min in average per subject). Contrary to previously published databases, in which labeled actions are very short and with low-semantic level, this new database focuses on high-level semantic activities such as "Preparing lunch" or "House Working". As a baseline, we evaluated several metrics on three different algorithms designed for online action recognition or detection.**

I. Introduction

Our work focuses on elderly people activity monitoring and smart home applications and aims to both detect and recognize long daily activities such as "having lunch". These activities consist in numerous actions, making the recognition task challenging. For instance, the "washing dishes" class contains many iterations of "taking object", "washing", "rinsing" or "put object on drain-rack". These sub-actions could also be divided in several gestures as "moving hand forward". The various kitchen configurations, used dishes, persons performing the activities and order in which they perform sub-actions, lead to a very high intra-class variability.

First, we outline three different semantic levels in behaviour analysis:

A *gesture* is a small motion with a very short temporal duration: "approach the hand to the mouth", "advance the leg" or "opens a drawer" for instance. It may not have any semantic meaning.

An *action* is generally composed of several gestures and has an immediate physical effect. The action "drink" for example is composed of the gestures "approach the hand to his mouth" and "move away the hand from his mouth". A typical duration for an action would be around 1 s to 10 s.

An *activity* is a long term behavior as "Having dinner" or "Taking a shower" for instance. It is generally made of several actions ("drinking", "eating", "cutting his meat", "taking the food on his plate"...) themselves composed of several gestures.

Gesture, action or activity recognition domains are currently widely explored. Related works deal with data coming from various sensors as wearable sensors as accelerometers [4], smart watches [2], environmental sensors such as contact sensors (doors, fridge...), proximity sensors to detect people around strategic locations, pressure sensors on the chair to detect sitting people [50], RFID tagged object [34], cameras [50] or Kinect [29].

In smart-home contexts, equipment should be as generic as possible and should be easy to set up. So, fixing sensors on different objects should be avoided as well as wearable sensors and sensors needing complicated calibration. Affordable depth sensors as the Kinect^{TM}suits well for such application since they can be installed as a classical camera would be.

Considering the lack of long range activity dedicated dataset, we captured the *DAily Home LIfe Activity (DAHLIA) Dataset* and make it available to the community[1]. This dataset consist in untrimmed video of high-level activities. Data were recorded with 3 Kinect^{TM}v2 surrounding the scene to deal with environment occlusions and human self-occlusions. We provide four data modalities, namely colour video, depth maps, skeleton body joint locations and body index -a binary mask relative to the detected body.

In this article, we also provide first results on the DAHLIA dataset using three methods namely *Deeply Optimized Hough Transform* [3], [35], Online Efficient Linear Search (ELS) [16] and Max-Subgraph Search [6].

In the following, related work is presented in Section II, we then introduce the DAHLIA dataset along with its specifications and evaluation protocols in Section III. For future comparison, we present a baseline in Section IV and Section V concludes this paper.

II. Related Work

Action analysis is receiving more and more interest in computer science community. Last years, many works were published, aiming to recognize or detect actions from a video stream. To evaluate such methods, various annotated video datasets have been made available to the community. Each proposed dataset was created within a specific context -*i.e.* for a specific application- and can be categorized according to the three following types:

a) Action Recognition, or Action Classification: it consists in finding one label for each pre-segmented video.

b) Action Detection: the goal of such algorithms is to locate occurrences of specific actions in a sequence. While the learning step is often done on segmented and annotated sequences, the testing step consists in detecting action occurrences in an untrimmed sequence.

[1] http://www.kalisteo.eu/en/mobilemii/datasets/index.htm

c) Online Action Recognition: algorithms have to both locate and identify randomly ordered actions in a continuous video stream. They process directly on flow coming from sensors.

This section present some of most common gesture, action or activity analysis datasets.

1) Action recognition oriented datasets: In an early stage, proposed datasets were quite simple *i.e.* with very short video clips, in highly controlled environments and with short and low-diversity gestures [11]. For instance, KTH dataset [28], widely used for action recognition purposes, contains very short black and white videos captured with homogeneous background.

Later, videos were extracted from real streams as television shows, movies or videos from websites [15]. These datasets are more realistic than the previous ones since backgrounds are cluttered and cameras are moving. However, actions occurring in these datasets are still short, and mainly contain only one or a few gestures.

With the emergence of low-cost depth sensors as the Microsoft®Kinect™, several RGB-D datasets have been released. We present some of the most popular ones in the following and suggest interested readers to refer to [49] for a survey.

One of the first published RGB-D action dataset was *MSR-Action3D* [48] which contains 20 actions performed 3 times by 10 subjects. The involved actions are short ones, as "high arm wave", "forward punch", "side kick", "jogging", "tennis serve", etc. The authors provided depth sequences and, later, skeleton data were also published. Wang et al. [37] introduced the *MSRDailyActivity* dataset, captured with a Kinect™v1, and provided RGB images, Depth map and skeleton. The 16 performed actions were daily usage oriented such as "drink", "eat", "read book", "play game", "sit down", etc.

Sung *et al.* published in [32] a dataset called *CAD-60* in which 12 actions were captured within 5 different environments namely bathroom, bedroom, kitchen, living room and office, to diversify background of captured videos.

These datasets only consider one single point of view in each example. In order to increase intra-class variability, but also to benefit from different points of view during testing step, multi-view action recognition datasets were introduced [7], [1], [21], [18], [14]. *UWA3D Multiview* [23] and *NJUST* [31] datasets were captured with only one camera and subjects were asked to perform each action several times, under different side views. Yet, most of these sets were captured with two or three kinect™sensors simultaneously [18], [14].

ATC4^2[7], one of the first multi-view dataset, was collected in 2012. It contains 14 classes corresponding to daily actions such as "Drink", "MakePhoneCall", "ReadBook", "Throw-Away", etc. but also two actions in relation to healthcare applications, namely "Collapse" and "Stumble". Color, depth and skeleton data extracted from 4 Kinect™sensors are registered.

Berkeley MHAD Dataset [21] was captured with different sensor types: mocap system, 4 stereo vision cameras, 2 Kinect™, 6 accelerometers and 4 microphones, to explore the complementarity of several modalities.

More recently, and to overcome the lack of examples in action recognition dataset, Shahroudy *et al.* [30] created *NTU RGB+D* dataset containing 56880 examples, captured with Kinect™v2. This new dataset aims to make easier the use of neural network in action analysis domain.

2) Detection and Online Recognition oriented datasets: In assisted living oriented applications, action detection and online recognition are much more realistic than action recognition. Indeed, streams extracted from uncontrolled settings are not segmented and, still, we want algorithms to locate and recognize occurring actions. Following this idea, datasets composed of continuous sequences made of several actions were captured [46], [27].

First, simple actions ("stand up", "sit down", "carry", "wave", "drink", etc.) were performed in an unsegmented stream [12]. In [40], Wei *et al.* provided a dataset in which subjects perform several actions simultaneously (amongst 12 actions). Furthermore, these actions may interact with each other. In [27], multiple actions can also occur simultaneously and algorithms have to detect actions in both temporal and spatial domains.

Wu *et al.* published in [42] a relatively large dataset acquired with the Kinect™v2 sensor. Actors were asked to perform several actions in either a kitchen or an office in continuous sequences. Thus, videos are untrimmed and suits well for segmentation algorithms. This dataset was captured at 13 different locations, with only one point of view per example.

Rohrbach *et al.* presented in [25] and its extension [26] a dataset containing continuous sequences of cooking activities: the MPII Cooking Activities Dataset. It is composed of 65 different activities such as "cut slice", "cut dice", "peel" performed by 12 actors. They increased variability by giving actors verbal instructions to prepare one out of 14 dishes such as *sandwich, salad....* Video duration vary from $3\,\mathrm{min}$ to $41\,\mathrm{min}$.

In the DMLSmartActions dataset [1], actors were also asked to perform series of actions without any interruption and without precise instructions. It was recorded using 2 HD cameras and one kinect™v1 sensor, providing 3 different points of view and depth map of the scene.

These 2 latest datasets are the closest to ours since actors behave in a natural way within a realistic environment.

If many gestures [19], [32], [21] or actions [37], [46], [42] databases exist in the literature, no long-term activity databases has been published to our knowledge. Thus, we introduce here the DAHLIA Dataset.

III. Proposed Dataset

In this section, we present the DAHLIA Dataset which will be made available to the computer science community. We first present the experimental protocol, included modalities, evaluation protocol and finally compare it to currently avail-

able datasets. Then, we introduce some evaluation metrics to evaluate algorithm performances.

A. Experimental protocol

We recorded 51 long sequences performed by 44 different persons during lunch time in a realistic kitchen. Before recording, we gave participants very simple instructions such that they performed the activities in a very natural way. Particularly, after a quick presentation of the kitchen, we asked them to perform 7 daily life activities in various orders (some of them are naturally ordered as, for example, setting the table before eating). The so-obtained dataset has high variety in the way of performing activities.

The 7 proposed high level activities, inspired from smart home applications, are:

1- Cooking: The subject prepare his lunch. This includes taking food, containers, cutting tomatoes and cheese, preparing oil and vinegar dressing,

2- Laying Table: The subject prepare the table for his lunch: the plates, the cutlery, the food, the water, etc.

3- Having Lunch: This activity contains entire lunch time, without specific instruction, the actors is simply asked to enjoy his lunch.

4- Clearing Table: The participant remove items from the table and put them near the sink or back to his original location.

5- Washing Dishes: The subject washes dirty dishes and, potentially, dry them.

6- Working: Participants were asked to answer a test by searching information in documents. The test was regularly changed to add variations.

7- House working: Participants were asked to sweep around the table, to clean the table and to change trash bag if necessary.

In order to be as realistic as possible, participants were invited to perform this experiment at lunch time and to actually have lunch during the data acquisition.

Average time of overall sequences is $39\,\mathrm{min}$ ranging from $24\,\mathrm{min}$ to $64\,\mathrm{min}$ for a total of 33.4 hours. Besides, since some activities as "Having lunch" takes far more time than "Laying table", high duration variability exists between classes. Fig. 1 represents the mean duration of the 8 classes (the seven previously defined classes plus a *neutral* one, introduced when none of the defined classes is present). Activities "Working" and "Having Lunch" represent 25% of the dataset while the activity "Cleaning Table" represents only 5% of the DAHLIA dataset.

Participants were 29 males and 15 females aged from 23 years old to 61 years old, increasing intra-class variability.

B. Environment settings and recording

The dataset was acquired within a fully monitored kitchen, surrounded by 3 synchronized kinect$^{\mathrm{TM}}$v2 as shown on Fig. 2. Several occlusions can appear depending on the user location and on the objects present on the table.

We provide four streams with a 15 fps frame rate.

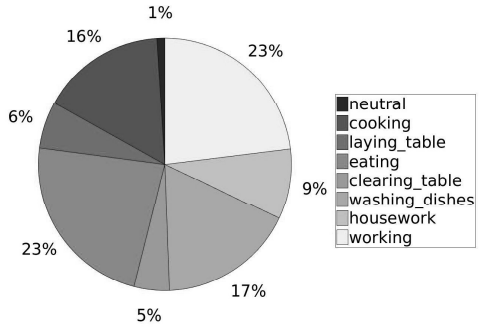


Fig. 1: Mean duration of the 8 classes (the seven classes previously defined and "neutral" one).

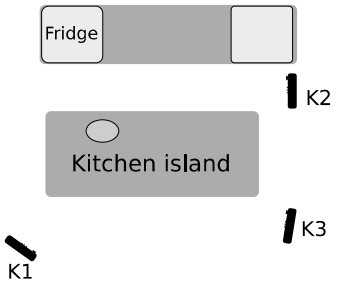


Fig. 2: Approximated camera location for DAHLIA Dataset recording.

RGB Videos recorded at high resolution (1920×1080 pixels) and compressed with H.264 at $2\,\mathrm{Mbits/s}$ bitrate.

Depth map with a 512×424 resolution. For each pixel, the 16 bits encoded value represent the distance between the sensor and corresponding point. We applied a low pass filter consisting in three median filters (on the x, y and temporal dimension). The filtered value is kept if it differs with the original one for more than $d = 10\,\mathrm{cm}$.

Skeleton data extracted using the SDK associated to the Kinect$^{\mathrm{TM}}$v2 sensor. It consists in the 3D locations of 25 joints of the human body. The sensor returns coordinates of joints in two different spaces (the depth map space and the 3D point space) and *tracking state* information. This *tracking state* is 0 if the joint is *not tracked*, 1 if it is *inferred* and 2 if *tracked*. Finally, all false detections that may have occurred have been manually removed.

Body Index. This pixel map indicates which pixels in the depth map are associated with the detected body.

Since the skeleton is extracted from the Kinect$^{\mathrm{TM}}$sensor, the body joint location estimation quality is irregular. Moreover, as a result of the kitchen configuration, the lower part of the body is mostly not tracked - because occluded by the central kitchen island.

Since all examples were captured in the same kitchen, we added variance by asking actors to perform the actions at various places in the scene. An extracted example from the dataset is shown on Fig. 3, representing representing the RGB views and depth maps coming from the three cameras.

Human skeleton is superimposed on depth maps. Skeleton joints have poor quality in case of large occlusion as in Fig. 3d or Fig. 3f.

C. Annotations and evaluation protocol

The 51 long-range sequences of the DAHLIA database are composed of multiple activities of daily life. Each frame of sequences has been manually labelled with one of the 7 activities or neutral.

In order to allow fair performance comparisons, we define here two accurate evaluation protocols that have to be followed to report results on this database.

a) Cross-Subject Evaluation: We defined two groups of participants: *group A* and *group B*, provided with the dataset. Both represent about half of the dataset and should be used as training set and testing set alternatively. Result for Cross-Subject Evaluation is then the average performance of the two configurations.

b) Cross-Subject and Cross-view Evaluation: To counterbalance the lack of environment variations, we also defined a cross-view evaluation protocol. In this protocol, a training set is defined for each combination of view and group of subject (*A* and *B*) while testing set should be done on each complementary set, for instance, one of the training/testing set would be trained on view 1, group A and testing on view 2, group B and view 3, group B independently. Then, average of the 12 defined sets is retained.

D. Comparison with existent databases

Zhang *et al.* [49] propose a survey on RGB-D-based Action Recognition datasets and introduced some criteria to compare them and particularly:

Characteristics of dataset acquisition

Mode 1: each action sample is stored in a sequence.

Mode 2: each sequence contains a continuous set of labelled sub-actions.

Mode 3: each sequence contains a continuous set of actions with the same order.

Mode 4: each sequence contains a continuous set of actions with random order.

Background clutter and occlusion

Low: the background is fixed and clean. There is no occlusion.

Medium: the background is fixed but is cluttered. Some occlusion can appear.

High: the background is not fixed and/or is cluttered. Occlusions are present and may affect action.

Kinematic complexity

Low: the movements are simple and with short duration.

Medium: the movements are of medium complexity and the duration is longer than in the previous case.

High: the movements are complex, with long duration.

Very High: The movements are very complex and composed of several sub-actions.

Variability amongst actions

Low: the variation of complexity levels amongst actions within a dataset is low.

Medium: the variation of complexity levels amongst actions within a dataset is medium.

High: the variation of complexity levels amongst actions within a dataset is high.

Table I summarises these different characteristics evaluated on several previously published dataset. It has been inspired from tables introduced in [49] with adapted criteria presented above. We evaluated the DAHLIA Dataset relatively to these criteria: the background is stable among action samples and cluttered. Cameras location and the kitchen configuration lead to partial occlusions of subjects. The DAHLIA dataset contains very long-range sequences compound of several activities that can be proceed in a variable order depending on the subject and sample. These activities involve many interactions with objects from the scene and the movements are very complex. Therefore, we evaluate the Kinematic complexity to "very high" and the variability amongst actions to "high".

Our DAHLIA Dataset has been performed by **44 subjects**, which is one of the highest subject number, with 7 **high semantic level activities**, contrary to previous dataset which dealt with low-level actions. Note that if 7 classes is one of the lowest class number, these activities could be decomposed in many sub-actions, leading to a **very high kinematic complexity**. This high semantic level yield to class lasting around $6\,\mathrm{min}$ **-the longest mean duration of analysed datasets-** in **untrimmed** (Mode 4) videos. This dataset has been recorded in a realistic kitchen, with clutter background.

Thus, the DAHLIA Dataset in composed of longer activities with a higher semantic level and both high intra-variability and high inter-variability.

E. Evaluation metric

In order to evaluate and rank algorithms, precise metrics are used and highly depend on the end-application. In previous work, several metrics were defined [13], [17], [8], [44], [25], [39], [41] to evaluate and/or understand algorithm performances.

Since we aim to provide a comprehensive baseline, we present here the metrics on which the DAHLIA Dataset has been evaluated.

For each class c to be detected in the dataset (binary classification), we define TP^c, FP^c, TN^c and FN^c as the number of True Positive, False Positive, True Negative and False Negative frames.

1) Frame-wise Accuracy: One common metric is the frame-wise accuracy which represents the ratio of correctly classified frames to all frames in the dataset (1). Note that this metric is sensitive to the class distribution but provides an intuitive measure of the algorithm ability to recognize actions.

$$\mathcal{FA}_1 = \frac{\sum_{c\in\mathcal{C}} TP^c}{\sum_{c\in\mathcal{C}} N_c} \tag{1}$$

where N_c is the number of frames labelled c in the ground truth.

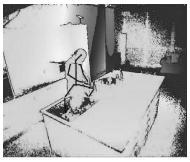
(a) Camera 1

(b) Camera 2

(c) Camera 3

(d) Depth from cam 1

(e) Depth from cam 2



(f) Depth from cam 3

Fig. 3: Example of "Cooking" activity from the DAHLIA database

Dataset	Year	#Subject	#Action	Mean Duration	Modality	#View	Dataset acquisition	Background	Kinematic	Variability	Semantic level
TUM [33]	2009	4	9	2 s	C,S	4	Mode 4	Low	Medium	Low	Gestures
MSR-Action 3D [S8]	2010	10	20	2.8 s	D,S	1	Mode 1	Low	Low	Low	Actions
CAD-60 [32]	2011	4	12	45 s	C,D,S	-	Mode 1	Medium	Medium	Medium	Actions
RGBD-HuDaAct [19]	2011	30	12	-	C,D	1	Mode 1	Medium	Medium	Medium	Actions
MSRDaily-Activity3D [37]	2012	10	16	-	C,D,S	1	Mode 1	Medium	High	Low	Actions
UTKinect [43]	2012	10	10	1 s	C,D,S	1	Mode 3	Medium	Medium	Low	Actions
ACT4 Dataset [7]	2012	24	14	4 s	C,D,S	4	Mode 1	Low	Low	Medium	Actions
MPII Cooking Activities [25]	2012	12	65	6 s	C,S	1	Mode 4	Medium	Medium	Medium	Actions
CAD-120 [14]	2013	4	10	17 s	C,D,S	1	Mode 2	High	High	High	Actions
UCFKinect [9]	2013	16	16	2 s	S	1	Mode 1	Any	Low	Low	Actions
3D Online [45]	2014	36	7	3 s	C,D,S	1	Mode 1/4	Medium	Medium	Low	Actions
Northwestern UCLA [38]	2014	10	10	-	C,D,S	3	Mode 1	Low	Medium	Medium	Actions
RGB-D activity [42]	2015	7	21	5 s	C,D,S	1	Mode 4	High	High	High	Actions
UTD-MHAD [5]	2015	8	27	2 s	C,D,S	1	Mode 1	Low	Low	Low	Actions
UWA 3D Multiview Dataset [22]	2016	10	30	-	C,D,S	4	Mode 1	-	-	-	Gestures
NTU RGB+D [30]	2016	40	60	5 s	C,D,S	80	Mode 1	Low	Medium	Medium	Actions
DAHLIA (proposed)	2016	**44**	7	**6 min**	C,D,S	**3**	**Mode 4**	High	**Very High**	High	**Activities**

TABLE I: Existing datasets with name and reference of the database | Year of publication | Number of subjects performing action | Number of actions | Mean duration of actions | Modality of the database: Color, Depth and/or Skeleton | Number of views | Characteristics of dataset acquisition | Background clutter and occlusion | Kinematic complexity | Variability amongst actions | Class semantic levels. These notions have been defined in Section III-D

2) F-Score: This metric combines *Precision* $\mathcal{P}^c$ and *Recall* $\mathcal{R}^c$ for each class c and is defined as the harmonic mean of these two values:

$$\mathcal{P}^c = \frac{TP^c}{TP^c + FP^c} \qquad \mathcal{R}^c = \frac{TP^c}{TP^c + FN^c} \tag{2}$$

$$\textit{F-Score} = \frac{2}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \times \frac{\mathcal{P}^c \times \mathcal{R}^c}{\mathcal{P}^c + \mathcal{R}^c} \tag{3}$$

3) Intersection over Union (IoU): this common metric was used to assess segmentation in the PVOC challenge [10]:

$$\text{IoU} = \frac{1}{|\mathcal{C}|} \sum_{c \in \mathcal{C}} \frac{TP^c}{TP^c + FP^c + FN^c} \tag{4}$$

IV. Baselines

In order to supply algorithm baselines on the DAHLIA dataset, we evaluated three algorithms, namely Deeply Optimized Hough Transform (DOHT) [3], Efficient Linear Search (ELS) [16] and Max-Subgraph search [6]. These methods have been chosen since they were designed for online action detection and/or recognition.

A. Deeply Optimized Hough Transform [3]

For this baseline, we present results computed with the DOHT algortihm on both skeleton and dense trajectories descriptors as [35]. Concerning vote parameters, we kept the same value for both features. Since activities defined in the DAHLIA dataset are much longer than those extracted in the original DOHT paper [3], we set the maximum considered time displacement M *-i.e. size of the temporal window for*

	Skeleton			Trajectories			HOG			Traj+HOG		
	$\mathcal{FA}_1$	F-Score	IoU	$\mathcal{FA}_1$	F-Score	IoU	$\mathcal{FA}_1$	F-Score	IoU	$\mathcal{FA}_1$	F-Score	IoU
View 1	0.60	0.58	0.42	0.74	0.73	0.58	0.80	0.77	0.64	0.73	0.73	0.59
View 2	0.63	0.60	0.44	0.78	0.76	0.62	0.81	0.79	0.66	0.79	0.78	0.64
View 3	0.73	0.71	0.56	0.76	0.74	0.59	0.80	0.77	0.65	0.77	0.76	0.62
Multiviews	0.65	0.64	0.48	0.81	0.80	0.67	0.85	0.82	0.71	0.82	0.80	0.68
Cross-Cam	0.34	0.31	0.19	view-dependent descriptors								

TABLE II: Result for DOHT on DAHLIA dataset

the Hough votes- to 1000 and parameter C (data attachment) to 4, which provided the best results.

a) Body joints: To exploit extracted Body Joints from the Kinect™, we first normalized raw data applying a normalization similar to the one presented by Raptis *et al.* in [24]. Note that this normalization is robust to view point variation and voluntarily ignores the person's location in the room since we do not want the algorithm to benefit from this information for recognizing activities.

Due to the room configuration and camera locations (indicated in Fig. 2), occlusions occur differently on each sensor. To overcome these occlusion issues, we combined information from all views following [35] workflow. We consider the following joints : *Head*, the two *Hands, Elbows, HandTip, Wrist, Shoulders, Hips, Knees* and *Spinebase*. In each frame, we only consider joints associated to the *tracked* state (provided by the sensor). Frames where shoulders have a low confidence status (because of the skeleton normalization strategy) do not provide votes.

b) Dense trajectories descriptor [36]: Following [35], we applied the DOHT algorithm on RGB data from DAHLIA dataset. Both Trajectory shape and HOG (Histogram of Gradient) descriptors [36] were used, as well as a concatenation Traj+HOG as presented in [35].

c) Results : Table II summarises obtained results with the DOHT algorithm. In each configuration, the DOHT algorithm outputs best predictions when used with HOG features. It emphasizes the importance of spatial context in the process of online activity recognition since the HOG descriptor captures the local shape of an image. The use of multiple views in the training and testing process leads to higher results since data extracted from different views complement one another.

Results are lower in the cross-view protocol, which is consistent with the fact that features extracted on one view can be occluded in another. For the DOHT algorithm, this protocol is the most challenging one. Note that since Dense Trajectories descriptor is a view-dependant descriptor, we did not run the cross-View protocol for these features.

Finally, Table III presents per-class results with the HOG descriptor which is the one associated to the highest results.

	Multicam Hog	
	F-Score	IoU
Cooking	0.75	0.60
Laying Table	0.69	0.53
Eating	0.91	0.84
Clearing Table	0.75	0.59
Washing Dishes	0.87	0.77
Housework	0.86	0.75
Working	0.92	0.86

TABLE III: Per-class results obtained with DOHT algorithm computing cross-subject protocol using HOG in a multiviews approach.

	Skeleton		
	$\mathcal{FA}_1$	F-Score	IoU
View 1	0.18	0.18	0.11
View 2	0.27	0.26	0.16
View 3	0.52	0.55	0.39
Cross-Views	0.31	0.32	0.21

TABLE IV: Result for ELS [16] on DAHLIA dataset

B. Online Efficient Linear Search (ELS)

Meshry *et al.* [16] proposed an online action detection method based on 3D skeleton sequences. A codebook is generated from skeleton features and weights are learnt for each entry of this codebook through a linear SVM. Then, online recognition consists in an identification of the maximum sub-interval score based on these weights. We refer an interested reader to [16] for more details. We used the source code provided by the authors to run the following baseline.

Online ELS algorithm was computed with the local descriptor described in the original paper: a weighted concatenation of the angles desciptor Θ [20], its velocity $\delta\Theta$ and an adaptation of Moving Pose descriptor [47] P, its first and second derivatives δP and $\delta^2 P$, respectively. The final form of their descriptor is $[P, \alpha\delta P, \beta\delta^2 P\psi\Theta, \delta\Theta]$ with α, β and ψ three weigths parameters. We evaluated several sets of parameters and present the best results we obtained on the DAHLIA dataset: $\alpha = 0.1, \beta = 0.1, \psi = 0.1$. We set the latency parameters to 2 frames and kept other parameters as in the original paper. Results obtained with this algorithm are presented in Table IV in the cross-subject and single view or cross-view protocol. We can observe that these results highly depend on the considered view. More precisely, higher results were obtained when the Camera 3 is used. This camera is the one with the least self-occlusion because of its location and angle relatively to the scene.

As well as in the previous section, lowest results are obtained with the Cross-view/Cross-subject protocol since a change in the point of view highly affects extracted descriptors.

C. Max-Subgraph Search

Chen and Grauman [6] proposed an action detection method based on a max-subgraph search named *T-Jump-*

	F-Score	IoU
View 1	0.25	0.15
View 2	0.18	0.10
View 3	0.44	0.31

TABLE V: Result with Max-subgraph Search method [6] on DAHLIA dataset

Subgraph: a graph is constructed for each action to be detected. Each node of this graph is associated to a score computed from descriptors extracted on a time window. The descriptors' weights are estimated through an SVM in a similar way than [16]. Since this algorithm was designed to make detection instead of online recognition, several activities can be triggered in each frame and the framewise accuracy $\mathcal{FA}_1$ cannot be computed. Results are presented in Table V, using the same local descriptor as ELS and a node size of 100.

V. Conclusion

In this paper, we introduced a new challenging dataset for activity recognition. This dataset suits well for applications such as smart-home or video-surveillance since the high-level semantic activities performed by the 44 subjects are continuously stored in 51 long untrimmed videos. We compared our dataset to previously published ones, emphasizing its challenging aspects, mostly regarding video duration and complexity of real human daily-life activities, composed of many (only partly-structured) actions, themselves being composed of several gestures. The mean duration of activities is around 6 minutes, significantly longer than those in existing untrimmed datasets.

Video scenarios were captured in a realistic way, with very simple instructions to increase intra-class and kinematic variability. This annotated dataset will be made available to the scientist community to evaluate algorithms on longer and more realistic class labels. To this end, we provide four streams recorded with the Kinect^{TM}v2. We defined two complementary evaluation protocols which should be followed by further works: cross-subject evaluation and cross-view evaluation. As a baseline, we evaluated three algorithms from the litterature and presented results with several well-known metrics.

References

[1] S. M. Amiri, M. T. Pourazad, P. Nasiopoulos, and V. C. Leung. Non-intrusive human activity monitoring in a smart home environment. In *e-Health Networking, Applications & Services (Healthcom), 2013 IEEE 15th International Conference on*, pages 606–610. IEEE, 2013.

[2] G. Bieber, M. Haescher, and M. Vahl. Sensor requirements for activity recognition on smart watches. In *Proceedings of the 6th International Conference on PErvasive Technologies Related to Assistive Environments*, page 67. ACM, 2013.

[3] A. Chan-Hon-Tong, C. Achard, and L. Lucat. Deeply optimized hough transform: Application to action segmentation. In *Image Analysis and Processing–ICIAP 2013*, pages 51–60. Springer, 2013.

[4] R. Chavarriaga, H. Sagha, A. Calatroni, S. T. Digumarti, G. Tröster, J. d. R. Millán, and D. Roggen. The opportunity challenge: A benchmark database for on-body sensor-based activity recognition. *Pattern Recognition Letters*, 34(15):2033–2042, 2013.

[5] C. Chen, R. Jafari, and N. Kehtarnavaz. Utd-mhad: A multimodal dataset for human action recognition utilizing a depth camera and a wearable inertial sensor. In *Image Processing (ICIP), 2015 IEEE International Conference on*, pages 168–172. IEEE, 2015.

[6] C.-Y. Chen and K. Grauman. Efficient activity detection with max-subgraph search. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1274–1281. IEEE, 2012.

[7] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *European Conference on Computer Vision*, pages 52–61. Springer, 2012.

[8] R. Collins, X. Zhou, and S. K. Teh. An open source tracking testbed and evaluation web site. In *IEEE International Workshop on Performance Evaluation of Tracking and Surveillance*, volume 35, 2005.

[9] C. Ellis, S. Z. Masood, M. F. Tappen, J. J. Laviola Jr, and R. Sukthankar. Exploring the trade-off between accuracy and observational latency in action recognition. *International Journal of Computer Vision*, 101(3):420–436, 2013.

[10] M. Everingham, S. A. Eslami, L. Van Gool, C. K. Williams, J. Winn, and A. Zisserman. The pascal visual object classes challenge: A retrospective. *International Journal of Computer Vision*, 111(1):98–136, 2015.

[11] L. Gorelick, M. Blank, E. Shechtman, M. Irani, and R. Basri. Actions as space-time shapes. *IEEE transactions on pattern analysis and machine intelligence*, 29(12):2247–2253, 2007.

[12] D. Huang, S. Yao, Y. Wang, and F. De La Torre. Sequential max-margin event detectors. In *European conference on computer vision*, pages 410–424. Springer, 2014.

[13] R. Kasturi, D. Goldgof, P. Soundararajan, V. Manohar, M. Boonstra, and V. Korzhova. Performance evaluation protocol for text, face, hands, person and vehicle detection & tracking in video analysis and content extraction (vace-ii). *Protocol Document*, 2005.

[14] H. S. Koppula, R. Gupta, and A. Saxena. Learning human activities and object affordances from rgb-d videos. *The International Journal of Robotics Research*, 32(8):951–970, 2013.

[15] H. Kuehne, H. Jhuang, E. Garrote, T. Poggio, and T. Serre. Hmdb: a large video database for human motion recognition. In *2011 International Conference on Computer Vision*, pages 2556–2563. IEEE, 2011.

[16] M. Meshry, M. E. Hussein, and M. Torki. Linear-time online action detection from 3d skeletal data using bags of gesturelets. In *Applications of Computer Vision (WACV), 2016 IEEE Winter Conference on*, pages 1–9. IEEE, 2016.

[17] J. Munkres. Algorithms for the assignment and transportation problems. *Journal of the society for industrial and applied mathematics*, 5(1):32–38, 1957.

[18] B. Ni, Y. Pei, P. Moulin, and S. Yan. Multilevel depth and image fusion for human activity detection. *Cybernetics, IEEE Transactions on*, 43(5):1383–1394, 2013.

[19] B. Ni, G. Wang, and P. Moulin. Rgbd-hudaact: A color-depth video database for human daily activity recognition. In *Consumer Depth Cameras for Computer Vision*, pages 193–208. Springer, 2013.

[20] S. Nowozin and J. Shotton. Action points: A representation for low-latency online human action recognition. *Microsoft Research Cambridge, Tech. Rep. MSR-TR-2012-68*, 2012.

[21] F. Ofli, R. Chaudhry, G. Kurillo, R. Vidal, and R. Bajcsy. Berkeley mhad: A comprehensive multimodal human action database. In *Applications of Computer Vision (WACV), 2013 IEEE Workshop on*, pages 53–60. IEEE, 2013.

[22] H. Rahmani, A. Mahmood, D. Huynh, and A. Mian. Histogram of oriented principal components for cross-view action recognition. 2016.

[23] H. Rahmani, A. Mahmood, D. Q. Huynh, and A. Mian. Hopc: Histogram of oriented principal components of 3d pointclouds for action recognition. In *European Conference on Computer Vision*, pages 742–757. Springer, 2014.

[24] M. Raptis, D. Kirovski, and H. Hoppe. Real-time classification of dance gestures from skeleton animation. In *Proceedings of the 2011 ACM SIGGRAPH/Eurographics symposium on computer animation*, pages 147–156. ACM, 2011.

[25] M. Rohrbach, S. Amin, M. Andriluka, and B. Schiele. A database for fine grained activity detection of cooking activities. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1194–1201. IEEE, 2012.

[26] M. Rohrbach, M. Regneri, M. Andriluka, S. Amin, M. Pinkal, and B. Schiele. Script data for attribute-based recognition of composite activities. In *European Conference on Computer Vision*, pages 144–157. Springer, 2012.

[27] M. Ryoo, J. Aggarwal, and U.-I. Dataset. Icpr contest on semantic

description of human activities (sdha), 2010.

[28] C. Schuldt, I. Laptev, and B. Caputo. Recognizing human actions: a local svm approach. In *Pattern Recognition, 2004. ICPR 2004. Proceedings of the 17th International Conference on*, volume 3, pages 32–36. IEEE, 2004.

[29] L. Sevrin, N. Noury, N. Abouchi, F. Jumel, B. Massot, and J. Saraydaryan. Characterization of a multi-user indoor positioning system based on low cost depth vision (kinect) for monitoring human activity in a smart home. In *2015 37th Annual International Conference of the IEEE Engineering in Medicine and Biology Society (EMBC)*, pages 5003–5007. IEEE, 2015.

[30] A. Shahroudy, J. Liu, T.-T. Ng, and G. Wang. Ntu rgb+ d: A large scale dataset for 3d human activity analysis. *arXiv preprint arXiv:1604.02808*, 2016.

[31] Y. Song, J. Tang, F. Liu, and S. Yan. Body surface context: A new robust feature for action recognition from depth videos. *IEEE Transactions on Circuits and Systems for Video Technology*, 24(6):952–964, 2014.

[32] J. Sung, C. Ponce, B. Selman, and A. Saxena. Human activity detection from rgbd images. *plan, activity, and intent recognition*, 64, 2011.

[33] M. Tenorth, J. Bandouch, and M. Beetz. The tum kitchen data set of everyday manipulation activities for motion tracking and action recognition. In *International Conference on Computer Vision Workshops*, 2009.

[34] Y.-J. Tu, W. Zhou, and S. Piramuthu. Identifying rfid-embedded objects in pervasive healthcare applications. *Decision Support Systems*, 46(2):586–593, 2009.

[35] G. Vaquette, C. Achard, and L. Lucat. Information fusion for action recognition with deeply optimised hough transform paradigm. In *11th International Conference on Computer Vision and Applications (VISAPP)*, 2016.

[36] H. Wang, A. Kläser, C. Schmid, and C.-L. Liu. Action Recognition by Dense Trajectories. In *IEEE Conference on Computer Vision & Pattern Recognition*, pages 3169–3176, Colorado Springs, United States, June 2011.

[37] J. Wang, Z. Liu, Y. Wu, and J. Yuan. Mining actionlet ensemble for action recognition with depth cameras. In *Computer Vision and Pattern Recognition (CVPR), 2012 IEEE Conference on*, pages 1290–1297. IEEE, 2012.

[38] J. Wang, X. Nie, Y. Xia, Y. Wu, and S.-C. Zhu. Cross-view action modeling, learning, and recognition. In *Computer Vision and Pattern Recognition (CVPR), 2014 IEEE Conference on*, pages 2649–2656. IEEE, 2014.

[39] J. A. Ward, P. Lukowicz, and G. Tröster. Evaluating performance in continuous context recognition using event-driven error characterisation. In *International Symposium on Location-and Context-Awareness*, pages 239–255. Springer, 2006.

[40] P. Wei, N. Zheng, Y. Zhao, and S.-C. Zhu. Concurrent action detection with structural prediction. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 3136–3143, 2013.

[41] C. Wolf, E. Lombardi, J. Mille, O. Celiktutan, M. Jiu, E. Dogan, G. Eren, M. Baccouche, E. Dellandréa, C.-E. Bichot, et al. Evaluation of video activity localizations integrating quality and quantity measurements. *Computer Vision and Image Understanding*, 127:14–30, 2014.

[42] C. Wu, J. Zhang, S. Savarese, and A. Saxena. Watch-n-patch: Unsupervised understanding of actions and relations. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, pages 4362–4370, 2015.

[43] L. Xia, C.-C. Chen, and J. Aggarwal. View invariant human action recognition using histograms of 3d joints. In *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops*, pages 20–27. IEEE, 2012.

[44] D. P. Young and J. M. Ferryman. Pets metrics: On-line performance evaluation service. In *Joint IEEE International Workshop on Visual Surveillance and Performance Evaluation of Tracking and Surveillance (VS-PETS)*, pages 317–324, 2005.

[45] G. Yu, Z. Liu, and J. Yuan. Discriminative orderlet mining for real-time recognition of human-object interaction. In *Asian Conference on Computer Vision*, pages 50–65. Springer, 2014.

[46] J. Yuan, Z. Liu, and Y. Wu. Discriminative subvolume search for efficient action detection. In *Computer Vision and Pattern Recognition, 2009. CVPR 2009. IEEE Conference on*, pages 2442–2449. IEEE, 2009.

[47] M. Zanfir, M. Leordeanu, and C. Sminchisescu. The moving pose: An efficient 3d kinematics descriptor for low-latency action recognition and detection. In *Proceedings of the IEEE International Conference on Computer Vision*, pages 2752–2759, 2013.

[48] J. Zhang and S. Gong. Action categorization with modified hidden conditional random field. *Pattern Recognition*, 43(1):197–203, 2010.

[49] J. Zhang, W. Li, P. O. Ogunbona, P. Wang, and C. Tang. Rgb-d-based action recognition datasets: A survey. *Pattern Recognition*, 60:86–105, 2016.

[50] N. Zouba, F. Brémond, M. Thonnat, A. Anfosso, E. Pascual, P. Mallea, V. Mailland, and O. Guerin. A computer system to monitor older adults at home: Preliminary results. In *Gerontechnology Journal*, volume 8, pages 129–139, 2009.