Online temporal detection of daily-living human activities in long untrimmed video streams

Abhishek Goel*, Abdelrahman Abubakr†, Michal Koperski‡, Francois Bremond§, and ‖Gianpiero Francesca
*†‡§INRIA, Sophia Antipolis, ‖Toyota Motor Europe
*†‡§2004 Rte des Lucioles, 06902, Valbonne, France. ‖Hoge Wei 33, B - 1930 Zaventem
{*abhishek.goel †abdelrahman-gaber.abubakr ‡michal.koperski §francois.bremond}@inria.fr ‖gianpiero.francesca@toyota-europe.com

***Abstract*—Many approaches were proposed to solve the problem of activity recognition in short clipped videos, which achieved impressive results with hand-crafted and deep features. However, it is not practical to have clipped videos in real life, where cameras provide continuous video streams in applications such as robotics, video surveillance, and smart-homes. Here comes the importance of activity detection to help recognizing and localizing each activity happening in long videos. Activity detection can be defined as the ability to localize starting and ending of each human activity happening in the video, in addition to recognizing each activity label. A more challenging category of human activities is the daily-living activities, such as eating, reading, cooking, etc, which have low inter-class variation and environment where actions are performed are similar. In this work we focus on solving the problem of detection of daily-living activities in untrimmed video streams. We introduce new online activity detection pipeline that utilizes single sliding window approach in a novel way; the classifier is trained with sub-parts of training activities, and an online frame-level early detection is done for sub-parts of long activities during detection. Finally, a greedy Markov model based post processing algorithm is applied to remove false detection and achieve better results. We test our approaches on two daily-living datasets, DAHLIA and GAADRD, outperforming state of the art results by more than 10%.**

***Index Terms*—daily-living activity recognition, human activity detection, video surveillance, smart-homes**

I. Introduction

Many approaches were proposed to solve the problem of activity recognition in short clipped videos, which achieved impressive results with hand-crafted and deep features [1] [2] [3]. However, it is not practical to have clipped videos in real life, where cameras provide continuous video streams in applications such as robotics, video surveillance, and smart-homes. Here comes the importance of activity detection to help recognizing and localizing each activity happening in long video streams. Activity detection is more challenging than recognition. The detection system should recognize the activities, and observe the changes that happen between different activities and quickly recognize the new activity. In addition, the system should differentiate between real activities and background "neutral" activities. Finally, in some applications it is necessary to have an online detection system that takes live video streams as input, and localize start and end of each activity happening to take some decisions.

Although there are many previous approaches to address the problem of activity detection [4] [5] [6], the datasets used were small and the number of activity samples was limited. Recently, some challenges and datasets were introduced for the task of activity detection, for example, THUMOS'14 [7] contains a large number of youtube untrimmed videos for 20 classes, and ActivityNet [8] introduced 203 activity classes with an average of 137 untrimmed videos per class. The introduction of such datasets motivated more researchers to work on the problem of activity detection, as will be discussed in section II. Unlike general activity detection and datasets that use videos from web, there is another category of datasets with main focus on activities of daily-life (ADL) [9] [10] [11], where all the videos contain usual daily-living activities such as cooking, eating, reading, answering the phone, etc. Daily-living activities are different from general activities from web as some activities have similar motion, identical background, and the person can be occluded with different objects. In addition, the duration of the activities is usually long and different activities vary in duration.

In this work, we focus on the problem of daily-living activity detection. we specifically provide an online solution for the detection of such activities from long video streams. Our proposed algorithm depends on the sliding window approach, however, unlike other algorithms that try to use different window scales as proposals to fit all sizes of different activities, we train the classifier with sub-parts of the activities in training split with certain window size. After that while online detection the algorithm do frame-level recognition depending on previous W frames, where W is the same as window size used in training. Finally, a greedy markov model based post processing step is applied to filter out any noisy detection and boost the performance of the detection. The proposed pipeline will be discussed in section III.

We can summarize our contributions in this paper as follows:

1) Proposing new algorithm to do early detection of long activities by recognizing its sub-parts. The algorithm does frame-level detection, which is suitable for applications that need online activity detection.
2) Introducing new discriminative deep features that can capture the person and objects used in the scene, which helped to improve the detection accuracy of fine-grained activity classes.
3) Proposing a post-processing technique to further im-

prove the results of labeled frames generated by online detection step.

4) Analyzing and testing our proposed algorithm with hand-crafted and the new proposed deep features, proving that our pipeline outperforms state-of-the-art results for two public datasets, namely, DAHLIA [9] and GAADRD [10].

II. Related work

Many approaches and techniques were proposed to solve the problem of temporal activity detection. Laptev and Perez [6] used boosted space-time window classifiers and introduced the idea of "keyframe priming", but they focused only on the detection of "drinking" activity, and used one movie for training and one for testing. In [4] depending on movie scripts, they used a weakly-supervised clustering method to segment actions in videos. In [5] the authors proposed a framework for joint video segmentation and action recognition, the recognition model is trained using multi-class SVM, and segmentation is done using dynamic programming. In [12] features extracted from 3D skeleton data were used along with multi-scale sliding window approach. After introducing new big datasets such as THUMOS'14 [7] and ActivityNet [8], many new approaches were proposed, in [13] the authors used improved dense trajectories as features, and multi-scale sliding window approach with 11 different window sizes for detection followed by non-maximum suppression. The method proposed in [14] is a Single Shot Action Detector network based on 1D temporal convolutional layers to directly detect action instances in untrimmed videos. In [15] the authors proposed an end-to-end deep recurrent architecture that outputs action detections directly from a single-pass over the input video stream. In [16], an end-to-end Region Convolutional 3D Network was introduced, it encodes the video streams using a 3D convolutional network, then generates candidate temporal proposals followed by classification.

For daily-living activities, less methods and datasets for detection were introduced. In [17] the authors used a simple method for detection depending on the person's motion. They segment chunks for successive frames that contain motion and keep track of all interest points, then pass it to action recognition stage where HOG-HOF features around the sampled interest points are computed followed by Bag of Words representation and SVM classifier. The authors in [18] proposed an end-to-end Joint Classification Regression architecture based on LSTM network for both classification and temporal localization.

Recently, the DAily Home LIfe Activity Dataset (DAHLIA) was published [9], it is the biggest public dataset for detection of daily-living activities. Many algorithms were applied to this dataset as baseline; Online Efficient Linear Search (ELS) [19] utilized the sliding window approach along with features from 3D skeletons in each frame forming a new feature called "gesturelets", which is used to form a codebook then train SVM classifier. Max-Subgraph Search [20] represents action sequences as a space time graph, then try to identify the max-subgraphs that represent video subsequences having an activity of interest. Deeply Optimized Hough Transform (DOHT) [21] utilized a voting based method, each frame codeword has certain weight to vote for the label of neighboring frames, the weight function is learned using a new optimization method (mapped to a linear programming problem). In our work we used DAHLIA as the main dataset to test our proposed approach, along with smaller dataset such as GAADRD [10]. Our approach overcomes the issue of using multiple-scale window proposals by introducing the idea of sub-video recognition for early detection of long activities, which is more useful for real-life applications. Problem definition and the proposed approach will be introduced in next section III.

TABLE I
Average duration of activities in set 1 of DAHLIA dataset

Activity label	Avg length (frames)	Avg duration (minutes)
cooking	5663	6.2
laying table	1397	1.55
eating	8563	9.51
clearing table	1462	1.62
washing dishes	5882	6.53
housework	2694	2.99
working	7826	8.69
neutral	130	0.14

III. Proposed Approach

To understand the problems of temporal detection for long-term activities, we will analyze split 1 of DAHLIA dataset. Table I shows the average duration of the 8 activities in DAHLIA, from which we can notice that some activities are very long (9.5 minutes for "eating"), while other activities takes small duration (1.5 minutes for "laying table" and 9 seconds for "neutral"). The big difference in activities duration makes it difficult to apply multi-scale sliding window approach, and the long duration of some activities doesn't help for single window approach that try to fit the whole activity duration. In addition, for systems that require online detection, it is not practical to wait for long time until reading all frames of long activities to recognize its label. Furthermore, the borders between activities are not usually clear to decide if the activity is finished.

Based on these points, and to solve the problem in online manner, we think of the problem as early detection of long activities, which does not need to have the full activity duration to recognize its label. Thus, we propose a new algorithm that can recognize sub-parts of the long activities instead of trying to fit the duration of all activities. In next sub-sections, we will introduce the proposed pipeline. First, we discuss the training process using activities sub-videos instead of full clips of activities, then, we explain the proposed online frame-level recognition depending on single window approach. After that, a novel post-processing greedy algorithm is explained, that depends on average activities duration, followed by applying Markov model to capture the relationships between successive activities. Finally, the features used will be discussed including the proposed Person-Centered CNN features.

A. Classifier Training

In the basic single sliding window approach, the classifier is trained using the features of the full activities duration, which can be considered as an intuitive baseline for using single window, but the results of using this approach were always not good with DAHLIA dataset. More about this will be shown in experimental results section IV. Therefore, to solve the problem of long activities duration discussed before, we propose a different way of training our classifier. First, all training videos are divided into relatively small windows of size W frames, which represent activity sub-videos (sub-parts). Then the features are generated for all these windows and the training is done with linear SVM classifier using all activities sub-videos. Figure 1 shows the process for training the single sliding window classifier for our proposed approach. The size of W is set to 250 frames by cross validation.

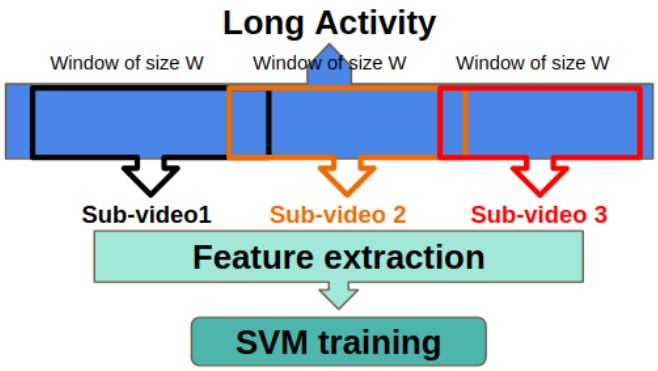

Fig. 1. The training process for our classifier. First, training videos are divided into windows of size W frames, representing activity sub-parts. Then the features are generated for these windows and the training is done with SVM

B. Online frame-level detection

After training the classifier, Online frame-level detection is done. Figure 2 shows the proposed detection framework; for each frame in the video, a sliding window extracts the feature of previous W frames, then these features are fed to the SVM classifier to recognize the label of current frame depending on the features of previous W frames. This process is done for all the frames in the video, and each frame gets new label from the classifier. Using this approach, an early detection for the activities can be done for each coming frame, even without knowing the frames coming in the future.

C. Features Used

1) **Dense Trajectories features :** First features we tried are dense trajectories [1], which we consider as the baseline to test our approach. Dense Trajectories method depends on sampling key-points from each frame in the video, Then these points are tracked using optical flow. Action is represented by trajectory shape descriptor, HOG [22], HOF [1], and MBH [1], [22] computed in a cropped patches around each trajectory points.

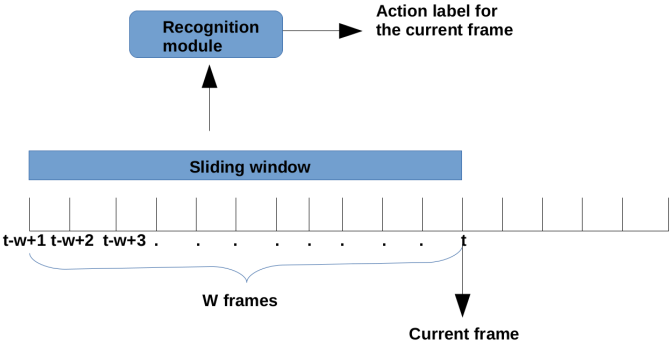

Fig. 2. Illustration of the proposed online detection framework. For each coming frame in the video, a sliding window extracts the feature of previous W frames, then these features are fed to the trained SVM classifier to recognize the label of current frame depending on the features of previous W frames.

2) **Person-Centered CNN (PC-CNN) features :** One of the problems of the proposed algorithm is the ambiguity between sub-parts of different activities. For example, in cooking activity the person usually wash some vegetables, which can be easily confused by washing dishes activity. To solve this kind of problems, we need good discriminative features that can capture the person and the objects used in each activity, in addition to capturing their movement in the scene. Therefore, we propose novel features called Person-Centered CNN (PC-CNN), that helps to get better visual and temporal cues from activities sub-parts.

To extract PC-CNN features, first, the person is detected using the Single Shot Multibox Object Detector (SSD) [23]. SSD is chosen due to its high speed and efficiency. After that, the detection bounding box is resized to 224x224 and the cropped spatial window is sent to ResNet-152 deep network [24] to extract the deep features from the last flatten layer. After obtaining the features from each frame, they are aggregated along W frames, in which a max and min pooling of the features is done to represent the temporal evolution of the features. In addition, the dynamic features are obtained by taking the difference between the features of each fourth frame, then getting max and min pooling of the features difference. Figure 3 shows the pipeline of extracting the PC-CNN features. As will be discussed later, these features combined with our proposed pipeline helped to outperform state-of-the-art results for DAHLIA dataset.

D. Post-processing

Human activities are complicated and usually people do not follow a certain way of doing the same activity. Therefore, there is always a chance of misclassification and confusion between different activities sub-parts while applying the proposed frame-level recognition. Another reason for these false detections is the similarity between some activities sub-parts as discussed before. This problem usually appears after online frame-level recognition, where there are some noisy false

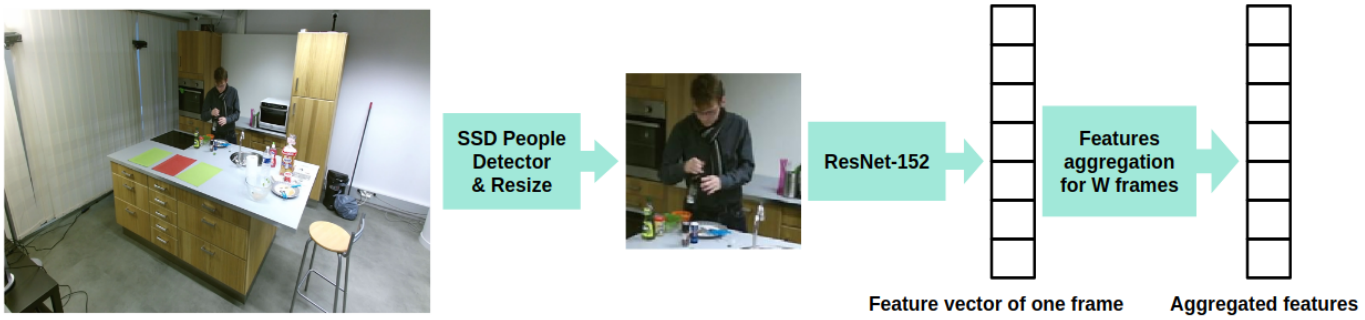

Fig. 3. Full pipeline to extract the Person-Centered CNN (PC-CNN) features for each frame. First. SSD object detection is applied. After that, the detection bounding box is resized to 224x224 and the cropped spatial window is sent to ResNet-152 deep network to get the deep features. After obtaining the features from each frame, they are aggregated along W frames using max and min pooling.

detected frames that occur in the middle of long activities and in the border between two different activities.

Algorithm 1 The proposed post-processing algorithm depending on activities duration

Result: Post processed activity intervals
input_dataframe = start - end frames and name of activities
avg_length = Lookup containing average lengths of activities
n_start = start frame of the fine-tuned activities (init = 0)
intervals_to_delete = index of intervals identified as noise
threshold = Hyperparameter (0.1 by validation)
counter = -1
for *action, start_frame, end_frame in input_dataframe* **do**
  counter ← counter+1
  activity_length ← end_frame - start_frame + 1
  avg_length_action ← avg_length[activity]
  greedy_criterion ← (activity_length/avg_length_activity)
  if *greedy_criterion* $<$ *threshold* **then**
    if *n_start = 0* **then** $n\_start \leftarrow start\_frame$;
    else continue;
    intervals_to_delete.add(counter)
  else
    if *n_start* $\neq$ *0* **then**
      input_dataframe['start_frame'][counter] ← n_start;
    else continue;
  end
end
final_dataframe = input_dataframe.remove(intervals_to_delete)

To address this problem, we adopt a post-processing greedy algorithm that can filter out false detections based on the average duration of the activities calculated from training split. The pseudo code 1 shows the details of the proposed greedy algorithm. In addition, as the datasets used represent real life activities, there is a logical relationship between activities that can be helpful to spot false detections, for example in DAHLIA dataset, "clearing table" activity has high probability to happen after "eating lunch", and "eating lunch" always happen after "laying table". To model these relationships, we used a markov model that is obtained using the relationships between consecutive activities in training split forming a right stochastic Matrix M with each entry $M_{i,j}$ represents the probability that activity i is followed by activity j. Then while post-processing, the Markov matrix is used to check all consecutive activities and if the probability of $M_{i,j}$ is less than certain threshold, activity j is considered as false detection and takes the same label of activity i. The pseudo code of the greedy algorithm is shown in algorithm 1, and figure 4 shows a sample of the timeline for detection before and after post-processing.

IV. Experimental Results

To evaluate our proposed pipeline, we tested it with two different datasets; DAHLIA [9] and GAADRD [10]. The details and results of these datasets are explained in next subsections.

A. DAily Home LIfe Activity Dataset (DAHLIA)

First dataset to test our algorithm is the DAily Home LIfe Activity dataset (DAHLIA) [9], which is the biggest public dataset for detection of daily-living activities. This dataset consists of 51 long sequences recorded for 44 different people during lunch time in kitchen. The average duration of the sequences is 39 min with 7 different activity classes plus the neutral class. The activity classes are cooking, laying table, eating, clearing table, washing dishes, housework, working, in addition to the neutral class. Figure 5 shows some sample frames from DAHLIA dataset. The size of this dataset, and its similarity to real life situation make it an ideal choice for evaluating the proposed pipeline

Figure 4 shows a colored visualization of the detected activities in each frame for a sample test video. Each activity has one color, the ground truth is shown in first row, then the online frame-level detection in second row, and finally the best results after post-processing. From the visualization we can notice the false detected frames in the middle of correctly-detected frames, and how the post-processing helped to remove these noisy detection and improve the final results.

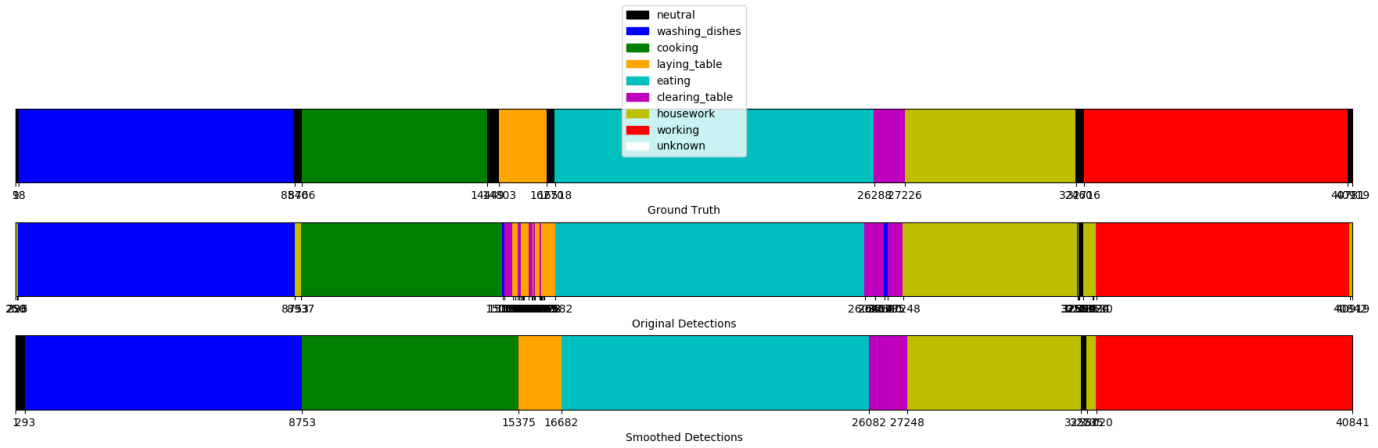

Fig. 4. Visualization of the detection results for S35 from DAHLIA datset. First row is the ground truth, second is the frame-level online detection, and last one is the detection after post-processing. Best viewed in color.

(a) Cooking (b) Working (c) House work (d) Eating

Fig. 5. Sample frames from DAHLIA dataset representing different activities and camera views

TABLE II
THE RESULTS OF OUR PROPOSED APPROACH WITH DENSE TRAJECTORIES AND PC-CNN FEATURES

Method	FA_1	F_score	IoU
ELS	0.18	0.18	0.11
Max Subgraph Search	-	0.25	0.15
DOHT(HOG) [state-of-the-art]	0.80	0.77	0.64
Simple sliding window + DT(HOG)	0.63	0.55	0.45
Simple sliding window + DT(HOF)	0.38	0.42	0.31
Our algorithm + DT(HOG)	0.78	0.70	0.59
Our algorithm + PC-CNN	0.88	0.75	0.66
Our algorithm + DT(HOG) + post-proc	0.87	0.75	0.68
Our algorithm + PC-CNN + post-proc	**0.90**	**0.79**	**0.71**

The evaluation criteria used to evaluate our approach are the same defined in the original work of DAHLIA dataset [9]. Frame-level accuracy represents the ratio of correctly classified frames to all frames in the dataset. The F-score metric combines Precision and Recall for each class and is defined as the harmonic mean of these two values. Finally, Frame-level IoU is the intersection over union between detection and ground truth frames. For all metrics, maximum value is 1.0, and higher values represent better performance. From table II, we can see the quantitative results of this evaluation. We can see that our approach outperforms state-of-the-art results for all evaluation metrics when combined with the PC-CNN features. In addition, our algorithm has comparable results with state-of-the-art even using the hand crafted dense trajectories features. Finally, we can notice that the results improved by around 2% - 6% when applying the post-processing to the detected labels. This shows the effectiveness of the online algorithm and the accuracy improvement by offline post-processing technique.

B. GAADRD dataset

Second dataset to test our approach is GAADRD [10]. This dataset comprises of daily living activities performed by 25 elderly people, with age 65 plus. The pool of these people comprises of patients suffering from dementia along with the healthy ones, which makes order of actions even more random and realistic. It contains 7 activities: reading article, watering plant, preparing drug box, preparing drink, turning on radio, talking on phone and balancing account, with no neutral class. Figure 6 shows some typical frames extracted from GAADRD dataset. Table III shows the results of GAADRD dataset. We note that for this dataset, the videos are not long enough and the frame rate is very low (e.g. “Preparing drug box” and “Watering Plant” activities are only 5-10 frames long). Therefore, we tested this dataset with the simple sliding window approach but showed that our proposed features are better than Dense trajectories features.

V. CONCLUSION

In this paper we proposed a new algorithm for human activity detection in long untrimmed videos. The pipeline is composed of two parts; recognition of activities sub-parts that