Online Detection of Long-Term Daily Living Activities by Weakly Supervised Recognition of Sub-Activities

Anonymous AVSS submission for Double Blind Review

Paper ID 120

Abstract

In this paper, we address detection of activities in long-term untrimmed videos. Detecting temporal delineation of activities is important to analyze large-scale videos. However, there are still challenges yet to be overcome in order to have an accurate temporal segmentation of activities. Detection of daily-living activities is even more challenging due to their high intra-class and low inter-class variations, complex temporal relationships of sub-activities performed in realistic settings. To tackle these problems, we propose an online activity detection framework based on the discovery of sub-activities. We consider a long-term activity as a sequence of short-term sub-activities. Then we utilize a weakly supervised classifier trained on discovered sub-activities which allows us to predict an ongoing activity before being completely observed. To achieve a more precise segmentation a greedy post-processing technique based on Markov models is employed. We evaluate our framework on DAHLIA and GAADRD daily living activity datasets where we achieve state-of-the-art results on detection of activities.

1. Introduction

With the proliferation of video recording devices capturing countless hours of videos on a daily basis, automatic content analysis is in a high demand. Since most of the recordings are untrimmed, it is the objective of activity detection to detect various occurrences of activities that happen throughout these long-term videos. Given an activity, the detection algorithm should localize it both in time and space providing an answer to "what is the activity?" and "where it happened in the video?" questions. Although numerous methods have been proposed [26, 5, 24] trying to improve activity recognition in videos, activity detection has become a more elusive target to achieve and the most crucial step in video activity analysis. Activity detection is more challenging since long-term untrimmed videos create larger and more versatile spatiotemporal volumes resulting in a higher search space. A favorable activity detection algorithm detects activities of interest while maximizes the temporal overlap of the ground-truth and its intersection with the detected boundaries.

Offline activity detection methods first potentially localize the activities in temporal domain by processing the whole video. Then to recognize the activities in the temporally detected intervals, a trained classifier based on extracted features across video frames is applied to form the final detection result. On the other hand, online approaches that are intrinsically unable to have access to the whole video in the first place, are compelled to perform both localization and classification steps simultaneously. In the case of daily living activities (ADL), the intended activity can go on for a long time. In addition to the original challenges, an early detection has to take place before the activity is fully observed. To be capable of this, online solutions should also cope with the issues regarding processing time complexity in order to produce real-time predictions. Therefore, reliable yet costly features cannot be directly applied due to these real-time processing requirements.

Previously, many methods have been proposed [6, 10, 14] to generate precise localization and well-anchored temporal segmentation of activities. In spite of these efforts, the small size of available datasets with a limited number of samples was an important issue hindering these challenges from being effectively resolved. In recent years, this problem is adequately remedied with the introduction of new challenges and large-scale datasets. For example, THUMOS'14 dataset [11] recollected a large number of untrimmed Youtube videos from 20 different activity categories providing a long-term and diverse set of activities. Similarly, ActivityNet [7] comprises 203 activity classes where each class includes an average of 137 untrimmed videos. Equipped with such datasets, the research community has become more motivated to work on the activity detection problem. Unlike general activity detection datasets that use videos from the web, there is another category of datasets that in particular focuses on activities of daily-living [25, 12, 13] (e.g. cooking, reading, answering the phone, and etc.). Such datasets introduce new challenges since the complexity of ADLs goes beyond activities from the web which have a high inter-class variability. Usually, diverse ADLs are performed with very similar motion patterns (even with no motion such as in reading) which makes them hard to discern. This leads to low inter-class variability and a vague boundary between person and the background due to the subtle variation of consecutive frames.

For online detection of activities, conventional sliding window approaches group sub-parts of activities with various granularities to generate proposals that fit activities with varying lengths. Inspired by these approaches, we propose a novel framework to precisely detect temporal boundaries of ADLs in long-term untrimmed videos with a two-phase algorithm. In the first phase, the candidate sub-activities of each activity class in the dataset are generated by clustering which employs aggregated frame-level features of a fixed window size. The goal is to train a classifier for each activity to recognize its sub-activities. The second phase refines noisy detections at the activity boundaries to improve the precision of temporal segmentation.

Our contributions can be summarized as follows:

- We introduce a new online frame-level activity detection pipeline which uses single-sized window approach. A weakly supervised classifier is trained directly on sub-activities discovered by clustering and operates on test videos to capture sub-activities of long videos within a fixed temporal window.
- To alleviate the noisy detections especially in activity boundaries, we propose a novel greedy post-processing method based on Markov models.
- We have extensively evaluated our proposed method on untrimmed videos from DAHLIA [12] and GAADRD [25] datasets and achieved state-of-the-art performances.

2. related Work

For a long time, there were many approaches proposed to solve the problem of temporal activity detection [14, 6, 10, 23]. However, some approaches required certain constraints and used limited data, for example, the authors in [14] focused only on the detection of "drinking" activity in movies, and used one movie for training and another one for testing. In [6] depending on movie scripts, the authors used a weakly-supervised clustering method to segment actions in videos. In [10] 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 [21] the authors used improved dense trajectories and multi-scale sliding window approach with many different window sizes for detection. The method proposed in [16] depends on 1D temporal convolutional layers to directly detect action instances in untrimmed videos. In [2] 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 [27], 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. Action tubes [8] was one of the successful approaches for activity detection, the authors used a two-stage approach to first select the regions which contain human motion, and extract spatial and temporal features from these regions along all frames, followed by SVM classification to label each activity.

For daily-living activities, fewer methods and datasets for detection were introduced. In [1] the authors used a simple method for detection depending on the person's motion; they segment chunks for successive frames that contain motion, then pass it to action recognition stage. The authors in [15] proposed an end-to-end Joint Classification Regression architecture based on LSTM network for both classification and temporal localization. In [20, 19] unsupervised method was used to detect the activities depending on the trajectory of people representing their global motion inside scene regions, the proposed unsupervised model defines these zones automatically during training and use it in test time to detect the activities.

Recently, the DAily Home LIfe Activity Dataset (DAHLIA) was published [25], which is by far the biggest public dataset for detection of daily-living activities. Various methods have been applied to this dataset providing baselines: Online Efficient Linear Search (ELS) [18] utilized the sliding window approach along with features from 3D skeletons in each frame to form a codebook then train SVM classifier. Max-Subgraph Search [4] 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) [3] utilized a voting based method. Each frame codeword has a certain weight to vote for the label of neighboring frames, and the weighting 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 [12] to show robustness of the framework when different types of descriptors (hand-crafted or deep) are used. Our approach overcomes the issue of using multiple-scale window proposals and utilizes the idea of sub-activity discovery for early detection of long activities which is more useful for real-life applications.

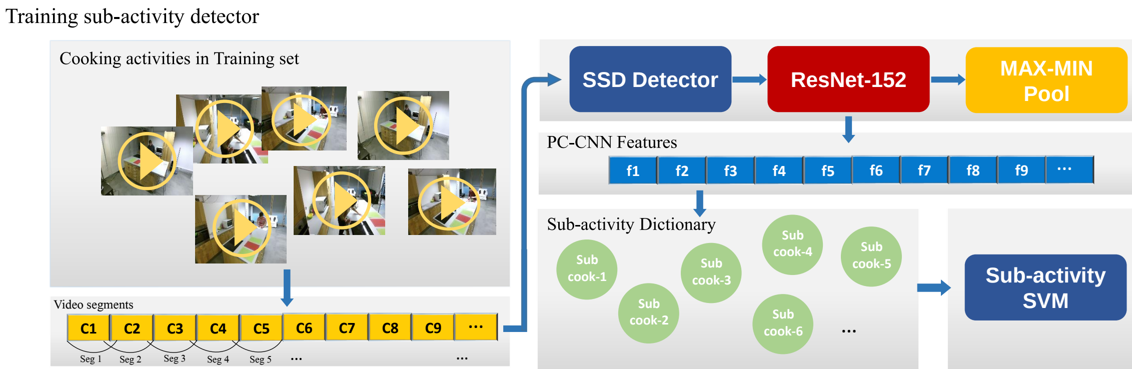

Figure 1. The process of extracting PC-CNN features and training of a weakly supervised sub-activity detector for the "Cooking" activity.

3. Proposed method

3.1. Overview

Our framework produces frame-level activity labels in an online manner by two major steps followed by a novel greedy post-processing technique. In order to handle long activities, the activities are decomposed into a sequence of fixed-length overlapping temporal clips. We then extract deep features from the clips. In order to characterize each activity with constituent sub-activities, we use K-means to cluster that activity's clips and construct a specific sub-activity dictionary. Therefore, we will have one sub-activity dictionary for each one of the activities. We represent an activity sequence with sub-activity assignments using the trained dictionary. Then, for each activity class, we train a binary SVM classifier (one versus all) based on its sub-activities. The trained classifiers are then simultaneously used to produce frame-level activity labels with the help of a sliding window architecture. It should be noticed that unlike multi-scale sliding window methods [23, 21], we only use a single fixed-size temporal window thanks to recognition of fixed length sub-activities. Finally, a greedy, Markov model based, post-processing technique is used for refinement of the obtained activity boundaries.

3.2. Feature Extraction

To align with the requirements of an efficient online detector, instead of applying feature extractors in a holistic manner, we use a local feature extractor. To avoid redundancy of holistic methods that misleads the classifier, we use a person-centric approach that rather than extracting not so useful static background features at every frame, focuses on the spatial context of a person in the scene. This approach not only helps the framework to obtain the best discriminative representation of the activities but also reduces the processing cost of expensive yet powerful CNN features by focusing on smaller patches. Inspired by CNN features introduced in [5], we name our feature Person-centric CNN or PC-CNN features (Fig. 1). Meanwhile, our framework is designed to be generic toward different feature types where the performance of the framework can be improved by replacement or modification of the features. To extract the features, first, Single Shot MultiBox Detector (SSD) [17] is used to get a bounding box around the person. SSD detector is used because of its accuracy and real-time performance without requiring region proposal network. The bounding box is extended by 20 pixels in the right, left and bottom of the box and resize to 244x244 in order to capture contextual information of the scene around the person. The resized images are fed to ResNet-152 [9] and deep features from last flatten layer are extracted resulting in a feature vector of size 2048. The temporal context of the videos is handled by the aggregation operator using max and min pooling. The frame descriptors are combined over time where the pooling mechanism helps to choose more salient values of the feature maps.

3.3. Sub-activity Recognition

Activity detection in long-term videos such as in ADLs is challenging due to temporal evolution of the activities. In particular, it is critical for an online framework to be able to detect an activity segment just by observing a fraction of a long activity. For instance, it is not efficient to wait until the end of "Cooking" activity to detect it. While Recurrent Neural Networks (RNN) are popular for predicting activities at each time by considering the observation at that time and previous hidden states of the model and model temporal progression of activities, these models fail due to not properly penalizing the incorrect predictions. In order to incorporate such properties in an activity detection framework, different from these methods, we use a weakly supervised classifier to discover sub-activities and predict the intended activity.

Consider a collection of V videos collected from "Cooking" instances in a dataset where each video consists of F frames (Figure 1). First, we decompose all the videos to a

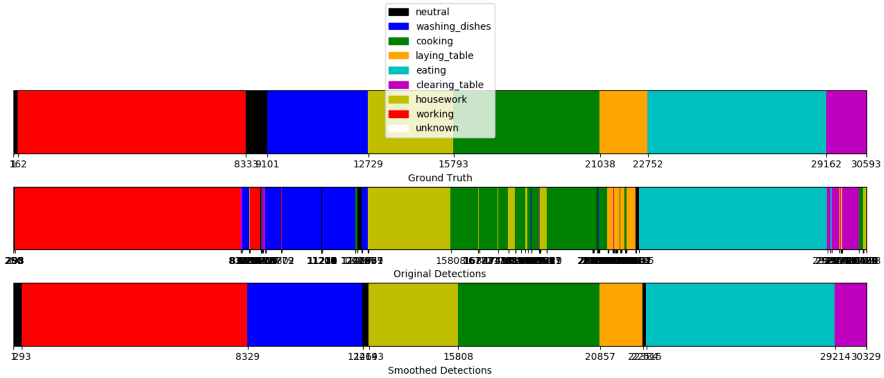

Figure 2. Visualization of temporal detection before and after post-processing for Subject 36 from camera view 1 in DAHLIA dataset (S36_A1_K1). First row is ground truth, second is online recognition, and finally the post-processed result.

sequence of fixed-size segments (250 frames). There is 50 frames overlap between adjacent segments. Although there might be some redundant segments without containing any semantic interpretation, most of the segments will include meaningful sub-activities of the main activity. Moreover, clustering process handles the redundancy of the segments by assigning them to the main sub-activity clusters. We then extract PC-CNN features from the obtained segments that result in a pool of features ($f = \{P(t)\}^{t=1:250}$ where P is the PC-CNN feature extractor). The final feature vector F_{all} is a concatenation of the features of the obtained segments. In order to build the sub-activity representation of the main activity, we run K-means to group the feature segments and produce sub-activity dictionary where the cluster centers represent discovered sub-activities. Using the sub-activity dictionary we can assign clusters to the video segments and represent a long video as a sequence of sub-activities. This is done by mapping the segments to the nearest sub-activity in the dictionary. We have selected K-means clustering algorithm to discover sub-activities:

$$\arg\max_{C} \sum_{j=1}^{K} \sum_{P_{all}(i) \in C_j} (\parallel F_{all}(i) - \mu_j \parallel^2) \tag{1}$$

where $C = \{C_1, \ldots, C_K\}$ is a set of clusters representing sub-activities and μ_j is the mean of the feature component values in cluster C_j. Therefore, given a certain value K, we use K-means algorithm over spatiotemporal features to generate the set of discovered sub-activities ($\psi = \{\psi_0, \ldots, \psi_{K-1}\}$. The exact number of the sub-activities is not known since the sub-activities are not labeled in the evaluated datasets. Therefore, to infer the ideal number of sub-activities (k) automatically Bayesian Inference Criterion (BIC) model selection is utilized [22]. To calculate the BIC score, assume the features F_{all} and a set of alternative models are given. To chose the best model BIC score representing the posterior probabilities of the models are calculated:

$$BIC(M_j) = \hat{l}_j(F_{all}) - \frac{p_j}{2}.logR \tag{2}$$

$\hat{l}_j(F_{all})$ is the log-likelihood of the jth model. p_j is the number of parameters in M_j and R is the total number of data points belonging to the centroids under consideration. The model with the highest BIC score is selected as the best model and its k value is taken as the ideal number of sub-activities for a given activity. The sub-activity dictionary generated with the ideal K is used for assigning sub-activities to the video segments. In order to recognize sub-activities, we train an SVM classifier using PC-CNN features of the training segments and the assigned sub-activity cluster codes as their labels. Give a test video segment, the classifier can infer what sub-activity it contains. In the training process of the classifier, the segments from a target activity are taken as positive samples and conversely, all the other segments are considered as negative samples. The same sub-activity discovery process is repeated for all other activities in the dataset to learn their sub-activities. The obtained set of classifiers are used in an online sliding window configuration with fixed length and stride to recognize sub-activities of a given test video. In the sliding windows, the previous n frames (n=250) are employed to label the current frame (frame-level labeling).

3.4. Post Processing

Refinement of the composed activities by the sub-activity proposals is crucial to develop an efficient activity detection framework. While sub-activity detector uses local window information to generate frame-level recognition, a refinement process can consider the context of the whole activity. After prediction of the frame-level sub-activity proposals, we link them to form the spatiotemporal sequence of sub-activities that helps to detect the entire video. Usually, false detection of sub-activities either occurs in the activity

(a) Cooking (b) Working (c) Prepare Drink (d) Watering Plant

Figure 3. Instances of daily activities provided in DAHLIA (a and b) and GAADRD (c and d) datasets.

boundaries where the borders are vague or in the middle of longer activities where common sub-activities are confused. This is mainly because of similar sub-activities co-exist in different activities. For example, the sub-activity detector can get confused between “Using Sink” sub-activity which is possessed in common between “Cooking” and “Washing Dishes” activities.

A greedy post-processing approach benefiting from duration (average duration of activities obtained from training instances) and temporal progression information of activities is adopted to resolve this issue. We can assume that usually there is a temporal order among the sub-activity sequences of a realistic ADL. For example, there is a high probability that the “Eating” activity is followed by a “Washing Dishes” activity. Markov models are suitable to model temporal sequences. We train a model that learns the sub-activity links from the training data. First, the model generates a stochastic Matrix M where each entry $\mathrm{M_{i,j}}$ is a probability showing that activity i is followed by activity j. Then, during post-processing, the Markov matrix is used to check all consecutive activities and if the probability of $\mathrm{M_{i,j}}$ is less than a certain threshold, activity j is considered as false detection and takes the same label of activity i (Figure 2).

4. Experimental Results

The performance of the proposed framework is evaluated on two public daily living activity datasets. The **DAHLIA** [25] consists of 153 long-term videos (51 videos recorded from 3 different views) recorded from 44 people performing ADLs. The average duration of the videos is 39 minutes with 7 different actions (and Neutral class). The considered ADLs are: cooking, laying table, eating, clearing table, washing dishes, housework, and working (3 a,b). The **GAADRD** dataset [12] consists of ADLs performed by 25 older adults. It includes 7 ADLs: reading article, watering plant, preparing drug box, preparing drink, turning on radio, talking on phone and balancing account with no neutral class (Figure 3 c,d).

The evaluations carried out following cross-subject protocol. In order to evaluate the proposed approach, metrics based on frame level accuracy have been used for the evaluation purposes. For each class c in the dataset, we assume $\mathrm{TP^c}, \mathrm{FP^c}, \mathrm{TN^c}$ and $\mathrm{FN^c}$ as the number of True Positive, False Positive, True Negative and False Negative frames respectively. Therefore, Frame-wise accuracy is defined as: $\mathrm{FA_1} = \frac{\sum_{\mathrm{c}\in\mathrm{C}} \mathrm{TP^c}}{\sum_{\mathrm{c}\in\mathrm{C}} \mathrm{N_c}}$ where N_c is correctly labeled frames compared to the ground-truth. F-Score is defined as: $F - Score = \frac{2}{|\mathrm{C}|} \sum_{\mathrm{c}\in\mathrm{C}} \frac{\mathrm{P^c} \times \mathrm{R^c}}{\mathrm{P^c}+\mathrm{R^c}}$ where P^c and R^c are precision and recall metrics of class c respectively. We also define Intersection over Union (IoU) metric as:

$$\mathrm{IoU} = \frac{1}{|\mathrm{C}|} \sum_{\mathrm{c}\in\mathrm{C}} \frac{\mathrm{TP^c}}{\mathrm{TP^c} + \mathrm{FP^c} + \mathrm{FN^c}} \tag{3}$$

where C is the total number of action classes.

Tables 1 and 2 show the results of applying the developed frameworks on GAADRD and DAHLIA respectively. It can be noticed that in DAHLIA dataset we significantly outperformed state-of-the-art results in all of the categories except in camera view 3 when the F-Score metric is used (we underperformed by a small margin of 1%). While we surpass ETS [18] and Max Subgraph [4] methods with a big margin, the closest performance to ours is DOHT [3] which utilizes both skeleton and dense trajectory descriptors. Obtaining similar results from different camera views highlights the robustness of our method to viewpoint variations and different types of occlusion. In order to compare the performance of our framework using hand-crafted and deep features, we reported the results of GAADRD dataset with the two types of features. As it can be seen, even with hand-crafted features our framework produces comparable results. GAADRD dataset is more challenging for activity detection since the videos are not long enough and the frame rate is very low (e.g. “Preparing drug box” and “Watering Plant” activities have instances with only 5-10 frames long). This makes sub-activity discovery and refinement process very challenging. Moreover, as it is recorded from real patients, the temporal order of activities are arbitrary and unpredictable (even sometimes some sub-activities are forgotten).

Method	FA_1	F_score	IoU
simple sliding window(HOG)	0.68	0.52	0.40
simple sliding window(PC-CNN)	0.61	0.55	0.44

Table 1. Detection results obtained on the GAADRD dataset.

	ELS [18]			Max Subgraph Search [4]			DOHT (HOG) [3]			Sub Activity		
	FA_1	F_score	IoU	FA_1	F_score	IoU	FA_1	F_score	IoU	FA_1	F_score	IoU
View 1	0.18	0.18	0.11	-	0.25	0.15	0.80	0.77	0.64	**0.85**	**0.81**	**0.73**
View 2	0.27	0.26	0.16	-	0.18	0.10	0.81	0.79	0.66	**0.87**	**0.82**	**0.75**
View 3	0.52	0.55	0.39	-	0.44	0.31	0.80	**0.77**	0.65	**0.82**	0.76	**0.69**

Table 2. The activity detection results obtained on the DAHLIA. Values in bold represent the best performance.

5. Conclusion

In this paper, we proposed a novel framework capable of temporal segmentation and classification of daily activities in long-term untrimmed videos. We suggested a person-centric feature (PC-CNN) based on SSD detector that satisfies required processing efficiency of online systems. We then proposed a weakly-supervised method for discovery of sub-activities of long-term activities which benefited from clustering and model selection methods to find the optimal sub-activities of the given activities. Finally, assuming temporal progression of sub-activities, we developed a greedy algorithm based on Markov models in order to refine noisy sub-activity proposals in middle and boundary regions of long activities. We evaluated the proposed method on two daily-living activity datasets and achieved state-of-the-art performances. In future work, we are going to improve the sub-activity discovery algorithm by making it capable of distinguishing similar sub-activities in two different activities.

References

[1] K. Avgerinakis, A. Briassouli, and I. Kompatsiaris. Activity detection and recognition of daily living events. In *Health Monitoring and Personalized Feedback using Multimedia Data*. 2015.

[2] S. Buch, V. Escorcia, B. Ghanem, L. Fei-Fei, and J. Niebles. End-to-end, single-stream temporal action detection in untrimmed videos. In *BMVC 2017*.

[3] A. Chan-Hon-Tong, C. Achard, and L. Lucat. Deeply optimized hough transform: Application to action segmentation. In *ICIAP 2013*.

[4] C. Chen and K. Grauman. Efficient activity detection in untrimmed video with max-subgraph search. *IEEE Trans. Pattern Anal. Mach. Intell.*, 2017.

[5] G. Chéron, I. Laptev, and C. Schmid. P-CNN: pose-based CNN features for action recognition. In *ICCV 2015*.

[6] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *ICCV 2009*.

[7] B. G. Fabian Caba Heilbron, Victor Escorcia and J. C. Niebles. Activitynet: A large-scale video benchmark for human activity understanding. In *CVPR*, 2015.

[8] G. Gkioxari and J. Malik. Finding action tubes. In *CVPR 2015*.

[9] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR 2016*.

[10] M. Hoai, Z.-Z. Lan, and F. De la Torre. Joint segmentation and classification of human actions in video. In *CVPR 2011*.

[11] Y.-G. Jiang, J. Liu, A. Roshan Zamir, G. Toderici, I. Laptev, M. Shah, and R. Sukthankar. THUMOS challenge: Action recognition with a large number of classes, 2014.

[12] A. Karakostas, A. Briassouli, K. Avgerinakis, I. Kompatsiaris, and M. Tsolaki. The dem@ care experiments and datasets: a technical report. *arXiv preprint arXiv:1701.01142*, 2016.

[13] M. Koperski. *Human Action Recognition in Videos with Local Representation*. PhD thesis, Universite Cote d'Azur, 2017.

[14] I. Laptev and P. Perez. Retrieving actions in movies. In *ICCV 2007*, pages 1–8.

[15] Y. Li, C. Lan, J. Xing, W. Zeng, C. Yuan, and J. Liu. Online human action detection using joint classification-regression recurrent neural networks. In *ECCV 2016*.

[16] T. Lin, X. Zhao, and Z. Shou. Single shot temporal action detection. In *ACM on Multimedia*, 2017.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. Reed, C.-Y. Fu, and A. C. Berg. SSD: Single shot multibox detector. In *ECCV*, 2016.

[18] M. Meshry, M. E. Hussein, and M. Torki. Linear-time online action detection from 3d skeletal data using bags of gesturelets. In *WACV 2016*.

[19] F. Negin, S. Cogar, F. Bremond, and M. Koperski. Generating unsupervised models for online long-term daily living activity recognition. In *ACPR 2015*. IEEE.

[20] F. Negin, M. Koperski, C. F. Crispim, F. Bremond, S. Coşar, and K. Avgerinakis. A hybrid framework for online recognition of activities of daily living in real-world settings. In *AVSS 2016*. IEEE.

[21] D. Oneata, J. Verbeek, and C. Schmid. Thumos 2014. 2014.

[22] G. Schwarz et al. Estimating the dimension of a model. *The annals of statistics*, 1978.

[23] A. Sharaf, M. Torki, M. E. Hussein, and M. El-Saban. Real-time multi-scale action detection from 3d skeleton data. In *WACV*.

[24] D. Tran, L. D. Bourdev, R. Fergus, L. Torresani, and M. Paluri. Learning spatiotemporal features with 3d convolutional networks. In *ICCV 2015*.

[25] G. Vaquette, A. Orcesi, L. Lucat, and C. Achard. The daily home life activity dataset: A high semantic activity dataset for online recognition. In *FG 2017*, May.

[26] H. Wang, A. Kläser, C. Schmid, and C. Liu. Action recognition by dense trajectories. In *CVPR 2011*.

[27] H. Xu, A. Das, and K. Saenko. R-c3d: Region convolutional 3d network for temporal activity detection. In *ICCV 2017*.