Recognition of Daily Activities by embedding hand-crafted features within a semantic analysis

Francesco Verrini
University of Genoa
Genoa, Italy
francesco.verrini@hotmail.com

Carlos Fernando Crispim-Junior
Université Lumière Lyon 2
Lyon, France
Carlos.CrispimJunior@univ-lyon2.fr

Manuela Chessa
University of Genoa - DIBRIS
Genoa, Italy
manuela.chessa@unige.it

Fabio Solari
University of Genoa - DIBRIS
Genoa, Italy
fabio.solari@unige.it

Francois Bremond
INRIA
Sophia Antipolis, France
Francois.Bremond@inria.fr

***Abstract*—The recognition of complex actions is still a challenging task in Computer Vision especially in daily living scenarios, where problems like occlusion and limited field of view are very common. Recognition of Activity Daily Living (ADL) could improve the quality of life and supporting independent and healthy living of older or/and impaired people by using information and communication technologies at home, at the workplace and in public spaces.**

This paper proposes to embed spatio-temporal information into ontology models to improve action recognition using visual words. Actions detected by visual words are implemented as Primitive States in the scenario and then used as Components of Composite States to merge them with spatio-temporal patterns that the people display while performing ADLs. In a challenging dataset, such as SmartHome, where a high variance intra-class and low variance inter-class is present, recognition results for some actions improve in precision and recall thanks to spatial information.

***Index Terms*—Semantic logic, a prior-knowledge formalism, detection and tracking, event recognition, action recognition, ontology model**

I. Introduction

With the advance and prevalence of low-cost sensors, computing devices and wireless communication networks, pervasive computing has become an achievable and deployable computing paradigm. As a result, research is now being conducted in all areas related to pervasive computing, ranging from low-level data collection, to intermediate level information processing, and then to high-level applications [1].

Smarthomes (SH) are residences (or homes) that are augmented or simulated with equipments, such as sensors, actuators and information processing systems. Within a SH the Activities of Daily Living (ADL) of its inhabitants, usually the elderly and the disabled people, can be monitored and analyzed so that personalized context-aware assistive living can be provided.

In automated monitoring systems for assisted living, the accuracy of event recognition is vital and event time intervals must be precisely assessed. Apart from alerting family members and caregivers about emergencies, video event understanding can aid medical evaluation and patient rehabilitation.

Action recognition is still a very challenging task, methods based on machine or deep leaning are very good in video classification, but still far from desirable results in untrimmed video, where multiple actions are performed, e.g. real life scenario and ADLs.

In [2], a method based on visual feature and a spatial grid are used inside the person bounding box, by computing Fisher Vector and HOG descriptor [3] for each spatial cell, but this method does not take into account spatial information of the scene. Action recognition methods could be based on hand-crafted features, such as in [4], where feature points are sampled for each frame and then tracked on the basis of displacement information from a dense optical flow field. These trajectories cover local motion information in videos, which can be used to identify events. Though, this method has problem in recognizing actions with low amount of motion.

In the last years, methods based on deep learning are becoming more and more popular, such as [5], [6], [7]. The latter is composed of two different deep neural networks: a spatial and a temporal stream. The spatial one captures the discriminating appearance features for action understanding, the temporal aims to learn the effective motion features. The drawback of this method is the request of a large number of labeled videos for training. Moreover, the deep learning datasets are concerned with the performance of a single activity, performed by a single person, whereas in ADL scenarios multiple activities could happen, performed by more than one person.

In conclusion, the methods listed above are very good in short video classification, though there is a lack of precision in real life scenarios.

Simultaneously, an approach based on the use of knowledge driven method, usually associated to an ontological formalism to define concepts and their interrelations, has been developed in [8], [9], [10] and [11].

A Kinect device and the skeleton joints that are obtained

with this low-cost sensor have been used in [12] and [13] to perform action recognition, but this skeleton detection is not robust enough to be used in a daily activity scenario.

In [14] the sequences of the joints, during an action, are used to draw RGB images in three dimensions, and then a CNN is used to classify them. However, in this method the scene environment is not taken into account as well.

Event recognition in *SmartHome* environment has been computed in [15] by using Hidden Markov Models to detect actions, but it is only based on representation of data and a training phase is needed. In [16], where a first person camera is used for ADL recognition, authors proposed to use detected objects with which person interacts, but excluding person's information such as posture.

The aim of this paper is to merge two approaches to take advantage of ontological models [8] to improve the results of the action recognition by using visual features [2]. With this approach any method, which uses features to recognize the action performed, can be improved by using prior knowledge of the scene, such as zones or the relationships between actions to detect more complex and longer events

II. Recognition of Daily Activities

This Section is organized as follows: first, we introduce the two approaches on which our method is based, and then we describe the proposed approach.

A. Knowledge-Driven Event Recognition

Event recognition using RGB-D sensors has been performed using the Scene Understanding Platform (SUP) [9]: the pipeline in the platform has different modules for detecting moving objects, tracking them over time and recognizing the events. The framework is divided into the following modules: people detection, people tracking and ontology driven event recognition [8] [9].

The people detection step is performed by the Single Shot Detection (SSD) algorithm proposed in [17]. The SSD approach is based on a feed-forward convolutional network that produces a fixed-size collection of bounding boxes and scores for the presence of object class instances in those boxes. The SSD normally start with a VGG pre-trained model, developed by the VGG team in Oxford [7], which is converted to a fully convolution neural network.

Then some extra convolutional layers are attached, which actually help to handle bigger objects. To make the detection more robust, the skeleton of the person has been added by using OpenPose [18], which allows the system to detect a person only if the skeleton is present.

For people tracking, the detected objects in a past temporal window are taken as input of the tracking algorithm, such as in [19]. To build the tracklet, six object descriptors are taken into account: 2D and 3D positions, 2D object area, 2D object shape ratio, color histogram and dominant color. These descriptors are considered to calculate a link score among any detected object in the temporal window, hypothesizing different trajectories based on the values of the links greater than a predefined threshold. All the link scores are then summed to calculate the dependability of the trajectory's hypothesis (see Fig.1).

Fig. 1. Example of detection and trajectory computation.

The constraint-based ontology model, which is based on prior knowledge of the involved objects, considered in this work follows the one already proposed in [20]. The ontology reasoning includes the following parts: Physical Objects (e.g. Person and objects in scene), Constraints and Components. In Fig. 2 how physical objects integrate 3D visual information into the ontological events is shown [8].

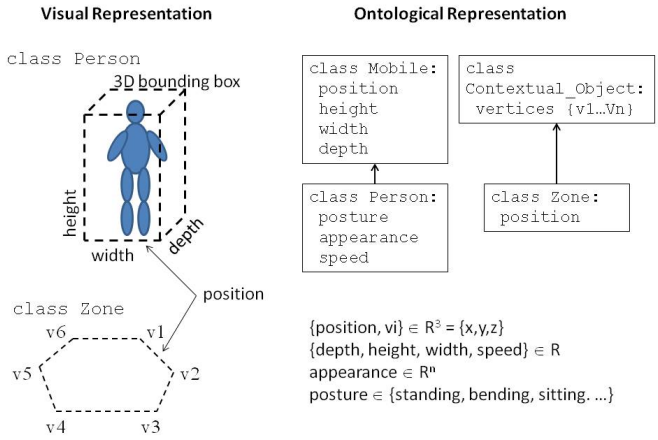

Fig. 2. 3D visual information of the scene is integrated into the ontological events as physical objects.

B. RGB-D Action Recognition With Sliding Window

This method detects the actions that are happening in an untrimmed video, by allowing a real-time action detection using a multi-scale sliding windows approach [2] [21].

Starting from trimmed videos of the action to be recognized, we extract local features with dense trajectory to compute a

Fisher Vector that is used to train a support vector machine (SVM) that will be used later in the pipeline.

To encode the trajectory information, three descriptors can be used as follows. Histogram of gradient (HOG) method [3] tiles the detector window with a dense grid of cells, with each cell containing a local histogram over orientation bins. In Histogram of flow (HOF) [22], the image is divided into small connected regions called cells, and for each cell it is computed a histogram of gradient directions or edge orientations for the pixels within the cell, then each cell is discretized into angular bins according to the gradient orientation. Finally, Motion Boundary Histogram (MBH) [4] is a descriptor, where derivatives are computed separately for the horizontal and vertical components of the optical flow.

Starting from the incoming frame, we use a temporal window (centered on frame t, spanning from $t-w$ to $t+w$, where w denotes a number of frames) to compute a Fisher Vector for every frame included in the window. Each window is classified by generating a label of the action detected. A confidence is obtained for each frame. Windows with confidence below a threshold are removed. The classification is made with the trained SVM.

C. Proposed approach

The novelty of the method proposed in this paper lies in merging a knowledge-driven pipeline with the machine learning pipeline as in Fig. 3. The upper pipeline has been explained in Section II-A and the lower in Section II-B.

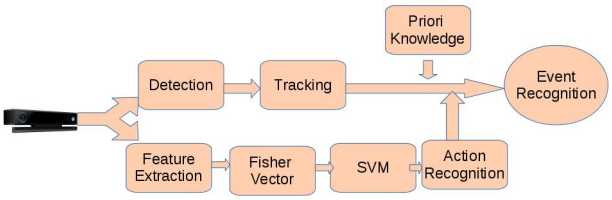

Fig. 3. Fusion of the two pipelines explained in the previous Sections.

The advantage of fusing the two pipelines is found in improving the results coming from the machine learning algorithm by taking the missing spatio-temporal information from the prior-knowledge. The improvement comes mainly with diminution of action recognized if the person or the actions taken from the lower pipeline does not respect constraints of the semantic logic.

In this paper we have done an effort in event modelling. The result is the development of 45 Primitive States, 39 Composite States and 19 new Attributes to the class Person. In addition, 2 new operators have been developed.

Using an already developed software, it has been possible to create zones according to a prior knowledge about the activities that the person could possibly perform, such as in Fig.4, where the zones of the kitchen are drawn. In this way, it is possible to define peculiar actions for every zone, e.g. Clean Dishes could only happen in the zone where the sink is located.

Fig. 4. View of the kitchen with the relative zones, where people activities can happen.

The ontological model of the Composite Event Cook.Clean_dishes_sink is visible in the following:

```
CompositeState(Cook.Clean_dishes_sink,
 PhysicalObjects((p1 : Person), (z1 : zone))
 Components(
  (c1:PrimitiveState Cook.Clean_dishes_ac(p1))
  (c2:PrimitiveState Person_Inside_Zone_Sink(p1,z1))
 )
 Constraints((c1->Interval and c2->Interval))
 Alarm ((Level : URGENT))
)
```

where one sub-event detected by the machine learning (e.g. Cook.Clean_dishes_ac) is fused with a primitive state that is derived from spatio-temporal patterns that people can do while performing activities of daily living (e.g. Person_Inside_Zone_Sink). The operator *and* is a logical constraints and is true only if the two components are detected together. In addition to logical operators ,others are available: such as temporal operators, where the time interval between two actions is used to detect a more complex one, or spatial operators, where the position of the person is considered.

The skeleton provided by OpenPose [18] is used to improve the people detection that will be performed on the physical object only if the skeleton is present, thus avoiding false detection. The skeleton has also been added to the class of interest Person. Thus, it is possible to use the skeleton and its joints as arguments of an operator by allowing the use of the pose of the person to detect complex actions, such as ”drink”, where the distance between the hand and the head of the person is taken into account to avoid false detection from the feature extraction method.

Moreover, the depth map obtained by the Kinect device is very noisy or completely missing, due to the distance from the

sensor to the scene. To solve this problem, an interpolation algorithm has been developed. The trajectory is also made more robust with the use of a median filter, which takes the median of the last twenty values of the depth data z.

A software to evaluate the approach has been developed. In this way, we can compare our output with the ground truth of the dataset that we used. The final output is an XML file, easy to be handled by non-experts of software (e.g. medical staff), containing the results.

III. Results

In this Section, we evaluate the proposed method by using a large dataset recorded in a real environment.

A. Dataset

The *SmartHome* dataset was recorded in 2015. The goal was to record large scale dataset, with daily-living actions performed in most realistic way as possible. At that time it was biggest daily-living action datasets contained a couple of hundreds of video clips, in addition they were recorded in quite constrained and controlled environment. In publicly available datasets, actions are performed by students, while one of the key application of daily-living action recognition is patient monitoring. Thus, in this dataset age of people performing actions varies from 60-80 years. Throughout recording process it was managed to gather more than 1000 hours of video footage. The manual annotation of the videos took more than 6 months, and *SmartHome* dataset was ready at the end of 2016. The house is composed of four rooms with two or three cameras installed in each of them to have different point of views. One of the cameras is pointing at the kitchen worktop with only the view of the hands, so it has been discarded in this work. The high intra-action variance and low inter-action variance make the dataset very challenging for action recognition's task [2]. The dataset is a private, thus we aim in future to use our approach also with a public one in order to compare our method with the state of the art in action recognition.

B. Evaluation

The results of the proposed approach are here compared with the ground truth file: a file containing the list of the actions of the video, which has been previously annotated, and the frame, in which actions are performed.

Table I shows on the left the results of action recognition using a sliding window of 200 frames, where features of the images are extracted and then classified with an SVM, and on the right there are the results of our fused approach that takes the prior knowledge of the room into account. From this table, comparing the left with the right side, we can see an improvement of some events, in particular in the most influenced by the semantic logic.

Specifically, we describe the following events:

Cook.Cleandishes The recognition of this event takes advantage from the definition of Zone Sink, avoiding the recognition of this action outside that zone. For this reason, the average of True Positive is mostly the same, but a substantial decrease of the False Positive is found, allowing a general increase of Precision (see Fig. 5).

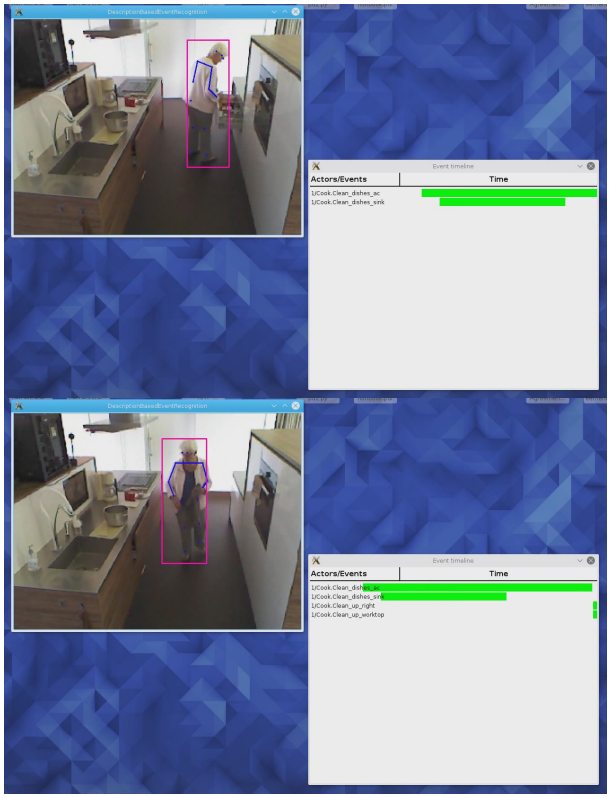

Fig. 5. The action predicted by the feature extraction pipeline is discarded thanks to the knowledge of the zones of the scene.

Cook.Cleanup A small improvement can be found in this action mainly due to the fact that our system merges two actions if they happen in a temporal distance of about hundred seconds. Furthermore, this action is not involved in the semantic logic, since it is an action that could be done everywhere in the scene.

Enter and Leave For these two actions the output of the machine learning pipeline is not taken into account by using only the semantic logic for their definition. Results improve by taking advantage of the possibility of recognizing an empty or an active scene.

Drink The presence of joints' positions, as a constraint, avoids false detection, if the distance between hands and nose is more that a certain value (see Fig. 6). On the other side, this action is sometimes performed outside the view of the camera, worsening the results.

Make Coffee And Make Tea The results are almost the same for all actions involving the making of tea or coffee, since there is no advantage from the semantic logic.

Walk The detection of this action improves (i.e. a decrease of the false positive) through the fusion of the action detected by the machine learning pipeline and of the displacement of the trajectory.

TABLE I

ON THE LEFT SIDE THE AVERAGE OF TRUE POSITIVE, FALSE POSITIVE, FALSE NEGATIVE, PRECISION AND RECALL COMPUTED BY THE ACTION RECOGNITION USING A SLIDING WINDOW OF 200 FRAMES IN THE PAST AND AN SVM FOR CLASSIFICATION; ON THE RIGHT THE RESULTS OF THE PROPOSED APPROACH. THE BETTER PERFORMANCES ARE MARKED IN BOLD.

		baseline					our method				
	event	TP	FP	FN	Precision	Recall	TP	FP	FN	Precision	Recall
0	Cook.Cleandishes	4.8095	11.9524	1.0000	0.2884	0.7736	4.2350	6.4416	0.6471	**0.4337**	**0.7807**
1	Cook.Cleanup	2.2273	9.0455	0.7273	0.2137	**0.6087**	2.6667	6.2222	0.6111	**0.2582**	0.5717
2	Cook.Cut	1.1000	6.5000	0.5000	**0.1265**	**0.5167**	0.1667	1.1667	0.0000	0.0556	0.1667
3	Cook.Stir	2.3750	10.1250	1.8750	0.1040	0.2951	3.4000	4.0000	2.2000	**0.2628**	**0.3316**
4	Drink	1.7500	1.2500	0.8750	0.3682	0.4010	1.6364	1.5454	1.0909	**0.4189**	**0.4515**
5	Eat.Snack	0.5000	1.0000	1.0000	**0.3333**	**0.2083**	0.3333	0.3333	1.0000	**0.3333**	0.1667
6	Enter	1.0000	3.2308	1.3077	0.1875	**0.3782**	0.6666	1.0000	0.3333	**0.2704**	0.3333
7	Leave	0.0000	1.5625	0.3125	0.0000	0.0000	0.5215	1.2500	0.0000	**0.2140**	**0.1500**
8	Makecoffee	0.0000	2.2222	1.8889	0.0000	0.0000	0.1250	2.1250	1.5000	**0.0250**	**0.0625**
9	Makecoffee.Pourgrains	0.4444	2.3333	0.8889	**0.1093**	**0.2778**	0.3750	2.3750	0.8750	0.1042	0.2500
10	Makecoffee.Pourwater	0.1429	2.1429	1.0000	**0.0476**	**0.1429**	0.1428	2.4286	1.1428	**0.0476**	**0.1429**
11	Maketea.Boilwater	0.5000	2.2500	0.3333	**0.1833**	**0.4167**	0.4000	2.2000	0.4000	0.1667	0.3000
12	Maketea.Insertteabag	0.2222	0.2222	0.5556	**0.2222**	**0.1667**	0.1667	0.5000	1.0000	0.1667	**0.1667**
13	Pour	0.4167	0.6667	0.3333	**0.1750**	**0.2292**	0.3333	0.6667	0.2222	0.0667	0.0833
14	Usetelephone	1.0000	1.8571	1.2857	**0.4405**	**0.5238**	0.8571	1.7143	0.8571	0.3690	0.5000
15	Walk	4.8000	11.2000	2.0000	0.2735	0.5924	4.9474	9.0531	1.7895	**0.3195**	**0.6510**

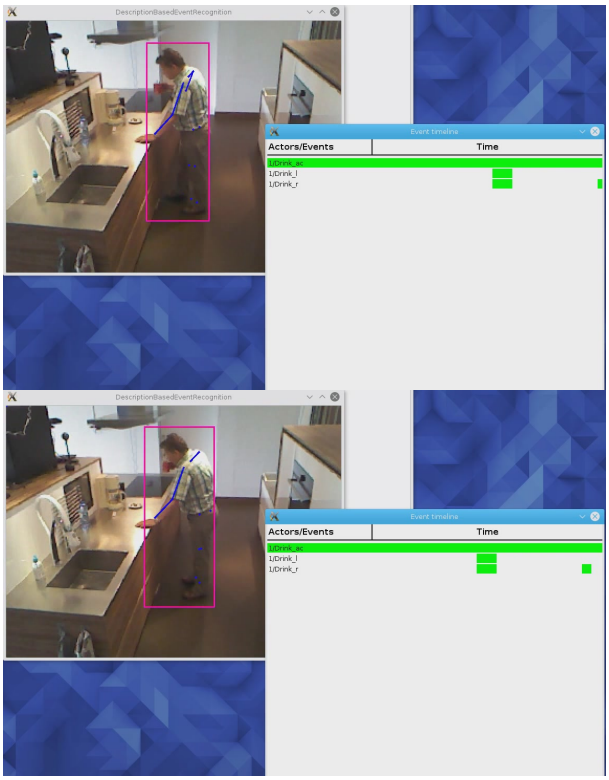

Fig. 6. Frames with Drink action: the proposed approach allows us to detect the action only when the wrist is close to the face.

C. Confusion Matrices

The confusion matrix helps to find which actions are misclassified by our approach. In this way, we can understand if the misclassification is due to a low inter-action variance (e.g. Cook.Cleandishes and Cook.CleanUp have a very low inter-variance so the system is often wrong), or if the approach does not classify correctly two actions with high inter-action variance. In the confusion matrices the class NoAction has been added for the frames that in the ground truth do not have a label assigned.

Exploring in more detail the confusion matrix for a specific video of the dataset (Fig. 7), it is visible a good amount of False Positive, most of which come from the NoAction column. The absence of True Positive for NoAction seems to point out that our system has problem in detecting when none of the action is happening. Looking at the Cook.Cleandishes row five False Positive are present in the NoAction columns, pointing out that most of the time for this action the corresponding label in the output file is missing. The same assumption could be applied to other actions.

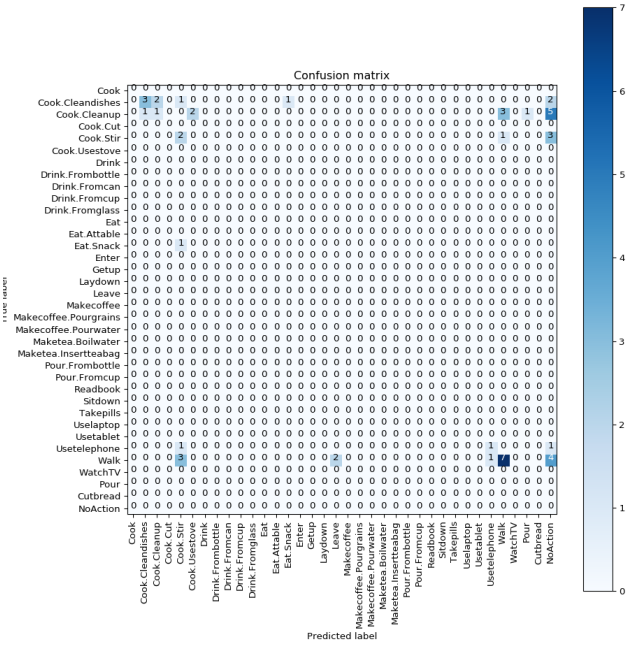

Fig. 7. Confusion matrix of a recorded video.

IV. Conclusions

In this paper, we have proposed a method that merges two approaches, by taking advantage of ontological models to improve the results of action recognition using visual features. The fusion of the two pipelines helps to improve the results of the visual features pipeline, by applying a sort of filter with the semantic logic. For this reason, the final results show a better precision not in terms of increase of True Positive, but in terms of the decrease of the False Positive detected. The main limitation of this method lies in the impossibility of increasing the number of True Positive, which are strictly dependent on the machine learning pipeline. The improvement in precision of some events is obtained by decreasing the number of False Positive.

In a future work, we plan to use skeleton joints for the extraction of features and for the analysis of the posture. Moreover, in the current work zones are drawn manually, but they can be learned, such as in [9], by clustering trajectory points corresponding to people's location on the ground. Another future goal would be to evaluate our method on a public dataset to compare our method with the state of the art algorithms.

Acknowledgment

References

[1] B. Yuan and J. Herbert, "Context-aware hybrid reasoning framework for pervasive healthcare," *Personal Ubiquitous Comput.*, vol. 18, no. 4, pp. 865–881, Apr. 2014. [Online]. Available: http://dx.doi.org/10.1007/s00779-013-0696-5

[2] M. Koperski and F. Brémond, "Modeling spatial layout of features for real world scenario RGB-D action recognition," in *13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016, Colorado Springs, CO, USA, August 23-26, 2016*, 2016, pp. 44–50.

[3] N. Dalal and B. Triggs, "Histograms of oriented gradients for human detection," in *2005 IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2005), 20-26 June 2005, San Diego, CA, USA*, 2005, pp. 886–893.

[4] H. Wang, A. Kläser, C. Schmid, and C. Liu, "Action recognition by dense trajectories," in *The 24th IEEE Conference on Computer Vision and Pattern Recognition, CVPR 2011, Colorado Springs, CO, USA, 20-25 June 2011*, 2011, pp. 3169–3176.

[5] H. Xu, Q. Tian, and J. Wu, "Exploring the influence of motion boundary sampling to improved dense trajectories for action recognition," in *Proceedings of the 7th International Conference on Internet Multimedia Computing and Service, ICIMCS 2015, Zhangjiajie, Hunan, China, August 19-21, 2015*, 2015, pp. 62:1–62:5.

[6] C. R. de Souza, A. Gaidon, E. Vig, and A. M. L. Peña, "Sympathy for the details: Dense trajectories and hybrid classification architectures for action recognition," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part VII*, 2016, pp. 697–716.

[7] K. Simonyan and A. Zisserman, "Very deep convolutional networks for large-scale image recognition," *CoRR*, vol. abs/1409.1556, 2014.

[8] C. F. C. Junior, A. G. Uría, C. Strumia, M. Koperski, A. König, F. Negin, S. Cosar, A. Nghiem, D. P. Chau, G. Charpiat, and F. Bremond, "Online recognition of daily activities by color-depth sensing and knowledge models," *Sensors*, vol. 17, no. 7, p. 1528, 2017.

[9] F. Negin, M. Koperski, C. F. Crispim, F. Brémond, S. Cosar, and K. Avgerinakis, "A hybrid framework for online recognition of activities of daily living in real-world settings," in *13th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2016, Colorado Springs, CO, USA, August 23-26, 2016*, 2016, pp. 37–43.

[10] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu, "Sensor-based activity recognition," *IEEE Trans. Systems, Man, and Cybernetics, Part C*, vol. 42, no. 6, pp. 790–808, 2012.

[11] C. F. C. Junior, Q. Ma, B. Fosty, R. Romdhane, F. Bremond, and M. Thonnat, "Combining multiple sensors for event detection of older people," in *Health Monitoring and Personalized Feedback using Multimedia Data*, 2015, pp. 179–194.

[12] L. Xia, C. Chen, and J. K. Aggarwal, "View invariant human action recognition using histograms of 3D joints," in *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, USA, June 16-21, 2012*, 2012, pp. 20–27.

[13] G. T. Papadopoulos, A. Axenopoulos, and P. Daras, "Real-time skeleton-tracking-based human action recognition using Kinect data," in *Proceedings of the 20th Anniversary International Conference on MultiMedia Modeling - Volume 8325*, ser. MMM 2014, 2014, pp. 473–483.

[14] M. Liu, H. Liu, and C. Chen, "Enhanced skeleton visualization for view invariant human action recognition," *Pattern Recognition*, vol. 68, pp. 346–362, 2017.

[15] G. Singla and D. J. Cook, "Interleaved activity recognition for smart home residents," in *Intelligent Environments 2009 - Proceedings of the 5th International Conference on Intelligent Environments - Barcelona, Spain 2009*, 2009, pp. 145–152.

[16] H. Pirsiavash and D. Ramanan, "Detecting activities of daily living in first-person camera views," in *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, USA, June 16-21, 2012*, 2012, pp. 2847–2854.

[17] W. Liu, D. Anguelov, D. Erhan, C. Szegedy, S. E. Reed, C. Fu, and A. C. Berg, "SSD: single shot multibox detector," in *Computer Vision - ECCV 2016 - 14th European Conference, Amsterdam, The Netherlands, October 11-14, 2016, Proceedings, Part I*, 2016, pp. 21–37.

[18] Z. Cao, T. Simon, S.-E. Wei, and Y. Sheikh, "Realtime multi-person 2D pose estimation using part affinity fields," *arXiv preprint arXiv:1611.08050*, 2016.

[19] C. Gárate, S. Zaidenberg, J. Badie, and F. Brémond, "Group tracking and behavior recognition in long video surveillance sequences," in *VISAPP 2014 - Proceedings of the 9th International Conference on Computer Vision Theory and Applications, Volume 2, Lisbon, Portugal, 5-8 January, 2014*, 2014, pp. 396–402.

[20] V. Vu, F. Brémond, and M. Thonnat, "Automatic video interpretation: A novel algorithm for temporal scenario recognition," in *IJCAI-03, Proceedings of the Eighteenth International Joint Conference on Artificial Intelligence, Acapulco, Mexico, August 9-15, 2003*, 2003, pp. 1295–1302.

[21] D. Cavaliere, L. Greco, P. Ritrovato, and S. Senatore, "A knowledge-based approach for video event detection using spatio-temporal sliding windows," in *14th IEEE International Conference on Advanced Video and Signal Based Surveillance, AVSS 2017, Lecce, Italy, August 29 - September 1, 2017*, 2017, pp. 1–6.

[22] N. Dalal, B. Triggs, and C. Schmid, "Human detection using oriented histograms of flow and appearance," in *Computer Vision - ECCV 2006, 9th European Conference on Computer Vision, Graz, Austria, May 7-13, 2006, Proceedings, Part II*, 2006, pp. 428–441.