Automatic action recognition for assistive robots to support MCI patients at home

G. Stavropoulos
Information Technologies Institute
Thessaloniki 57001, Greece
stavrop@ece.upatras.gr

D. Giakoumis
Information Technologies Institute
Thessaloniki 57001, Greece
dgiakoum@iti.gr

K. Moustakas
Department of Electrical and Computer Engineering,
University of Patras
Patras 26504, Greece
moustakas@ece.upatras.gr

D. Tzovaras
Information Technologies Institute
Thessaloniki 57001, Greece
dimitrios.tzovaras@iti.gr

ABSTRACT

This paper presents a novel approach for automatic human action recognition, focusing on user behaviour monitoring needs of assistive robots that aim to support Mild Cognitive Impairment (MCI) patients at home. Our action recognition method utilizes the human's skeleton joints information, extracted from a low-cost depth sensor mounted on a service robot. Herein, we extend the state of art EigenJoints descriptor to improve recognition robustness for a series of actions involved in common daily activities. Specifically, we introduce novel features, so as to take into account action specificities such as the joints' travelled distance and their evolution trend in subsequent frames to the reference one. In addition, we use information related to the user's manipulated objects, taking into account that several actions may be similar, yet performed with different objects, as well as the fact that real, practical applications involve continuous input video streams rather than pre-segmented action sequences. Through experimental evaluation on the MSR Action3D dataset, our approach has been found to outperform the state of art in action recognition performance. Evaluation has also been performed on a custom dataset, providing further promising results for future practical applications of our overall action recognition framework.

CCS CONCEPTS

• **Computing methodologies → Activity recognition and understanding**;

KEYWORDS

Assistive Robots; Activity recognition; Mild Cognitive Impairment

ACM Reference format:
G. Stavropoulos, D. Giakoumis, K. Moustakas, and D. Tzovaras. 2017. Automatic action recognition for assistive robots to support MCI patients at home. In *Proceedings of PETRA '17, Island of Rhodes, Greece, June 21-23, 2017,* 6 pages.
DOI: http://dx.doi.org/10.1145/3056540.3076185

DOI: http://dx.doi.org/10.1145/3056540.3076185

1 INTRODUCTION

The automatic monitoring of domestic daily activities is a significant challenge, towards future systems for assistive living, that will be capable to support persons in their daily domestic life, especially when considering older adults with cognitive impairments [10]. In this context, future service robots are anticipated to be capable to monitor their user's activities, identify abnormalities and try to provide assistance so as to improve the seniors' capacity for independent living and maintenance of quality of life [8]. This is of particularly importance when considering older persons with Mild Cognitive Impairment (MCI), who may be considered in general capable to autonomously perform daily activities at home, however at the same time, their behaviour in the scope of e.g. cooking or medication intake may involve abnormalities, from a forgotten fridge door left open through to the oven left turned on or pills intake forgotten. In order to detect such cases and provide proactive assistance, either with reminders through human-robot communication or robotic manipulations, a domestic service robot should first be capable to effectively monitor and assess user activities. Such assessment of daily activities can be performed on the basis of automatic recognition of lower-level actions, which in turn compose the higher level activities of the user's daily life.

More specifically, we take into consideration that complex daily activities, such as "cooking", comprise a series of lower-level actions, which are typically performed by the monitored person through interaction with objects of the home environment such as "open the fridge". By monitoring such lower-level actions, a robotic system can be enabled to understand the way that the user performs more complex activities and detect abnormalities, so as to then trigger its assistive behaviours. For instance, an assistive service robot may identify that after an "open the fridge" action, its counterpart, i.e. "close the fridge" has not been performed. In other cases, a service robot may detect the absence of user's hydration activities (e.g. water drinking) during a day, or a gradual increase in the time spent looking for objects stored in cupboards, or searching within multiple cupboards until s/he finds what is looked for. The automatic detection of such cases is necessary for robots that aim to assist MCI patients at home [8], yet it requires robust action recognition methods.

Given the importance of automatic action recognition for assisted living applications in general, significant research efforts have been made over the last decades, towards the development of corresponding systems that could be practically applicable in the future. Typically, user activity monitoring can be performed through wearable sensors [15] [4], sensors installed in the house environment [2] [6], or computer vision techniques applied on video streams taken from mobile cameras [20]. The latter line of research is especially relevant for mobile service robots, as these are typically equipped with RGB-D sensors that can be used to monitor the user's activity. Nevertheless, relying solely on the RGB-D streams of a service robot's camera for both user activity and home environment monitoring, introduces a series of challenges [13], especially when considering practical applications of domestic assistive robots. In this context, the present study builds upon the current state of art in computer vision-based automatic action recognition and tries to advance it, towards further increased robustness in realistic assistive robot applications.

1.1 Related work

Human action recognition has been a highly active area of research in the past decades. While initially, the focus was on recognizing human actions from monocular RGB video sequences [1], limitations such as sensitivity to illumination, various viewpoints etc., as well as the lack of 3D space information, significantly hindered such methods' effectiveness and real application potential. The recent advent of cost-effective depth sensors has alleviated some of these problems, as RGB-D sensors can provide 3D depth data of the scene, robust to illumination changes, and can offer sufficient information to recover 3D human skeletons. These advances have resulted in to an ever-increasing interest in human action recognition based on markerless skeleton extraction and body pose tracking [19].

Such approaches try to model the motion of either individual joints or combinations of joints using various features. Indicatively in [7], human skeletons were represented by their 3D joint locations, and the joint trajectories were modelled using a temporal hierarchy of covariance descriptors. A similar representation was also used with HMMs in [5]. Sheikh et al. [11] used a set of 13 joint trajectories in a 4-D XYZ-T space to represent a human action, and their affine projections were compared using a subspace angles-based view-invariant similarity measure. In [14], the human skeleton was represented by the pairwise relative positions of the joints, and the temporal evolutions of this representation were modelled using a Fourier coefficients hierarchy. An actionlet-based approach was utilized, in which discriminative joint combinations were selected using a multiple kernel learning approach. Yang & Tian [18][17] represented the human skeleton using relative joint positions, temporal joint displacement and offset of the joints with respect to the initial frame, formulating an eigenJoints-based action recognition framework. Action classification was performed using the Naive-Bayes nearest neighbor rule in a lower dimensional space constructed using principal component analysis (PCA). Zhu et al. [21] utilized a similar skeletal representation in conjunction with random forests. A view invariant human skeleton representation was proposed in [16], by quantizing the 3D joint locations into histograms based on their orientations with respect to a coordinate system fixed at the hip centre. The temporal evolutions of this view-invariant representation were modelled using HMMs.

Approaches like the above mentioned ones have demonstrated promising results in action recognition when evaluated on publicly available benchmarks, such as the MSR Action3D dataset [9]. Notably, the eigenJoints method is still considered a highly effective state of art approach. Nevertheless, the performance of state of art methods, even when evaluated on benchmark, acted datasets, shows that there is still space for improvement [17], especially when considering their future application in real domestic assistive service robots. The large variations in viewpoints, as well as the occlusions encountered in practical robotics applications call for further improvements in the employed descriptors that would make them more robust to the diverse and noisy inputs of realistic settings.

1.2 Contribution

In the present study, we focus on computer vision-based action recognition for assistive, domestic service robots, in need of user behaviour assessment during daily activities, considering this as a pre-requisite for the future improvement of corresponding systems' cognitive capacities, and proactive assistance provision capabilities. We utilize information of the user's tracked pose, as provided through markerless skeleton extraction from depth input video streams, which are taken from a low-cost RGB-D sensor [12]. The tracked user's skeleton is processed through our action recognition method, which builds upon the state-of-art eigenJoints [18] descriptor and extends it, so as to improve robustness in realistic scenarios. More specifically, we propose novel features that take into account further descriptive information of user actions, such as the joints' travelled distance and how the user pose evolves in subsequent frames to the reference one. We also examine an eigenJoints descriptor variation where only correspondent joints are taken into account for feature extraction. Alongside, we take into account interactions with objects, used within the user's daily activities. Object-related information is used to both perform action segmentation in continuous monitoring streams, as well as to better discriminate actions with similar movement characteristics, yet involving different manipulated objects.

Our proposed action recognition method has been comparatively evaluated over a publicly available benchmark dataset (MSR Action3D [9]), as well as over a custom dataset focusing on actions relevant to behaviour analysis in the scope of robotic assisted living applications. The obtained results show that our proposed features improve action recognition performance, compared to the original eigenJoints method. In addition, the incorporation of object-related information was found to further improve action recognition performance.

2 PROPOSED ACTION RECOGNITION METHOD

Our proposed action recognition method is an extension of the eigenJoints algorithm proposed in [16], which improves the original method's discriminative capabilities.

The original eigenJoints algorithm [17] focuses on extracting information about the relative positions of the joints between frames in a video sequence. It employs the calculation of three metrics; namely f_{cc}, f_{ci} and f_{cp}, which are essentially pairwise differences of the joints in the current frame (f_{cc}), between the current and the initial frame (f_{ci}) and with the previous frame (f_{cp}). The next step is to normalize the data in order to avoid domination of the attributes in higher numerical ranges. Next, in order to reduce redundancy and noise, Principal Components Analysis (PCA) is employed to obtain a compact representation of the concatenated $f_c = [f_{cc} f_{ci} f_{cp}]$ vector, by keeping the first 128 eigenvectors. After the eigenJoints features are extracted, a Naive Bayes Nearest Neighbour (NBNN) classifier is employed. The developed implementation exploits FLANN (Fast Library for Approximate Nearest Neighbours) and calculates Video-to-Class distance in order to output the most probable action given an input video sequence.

In our proposed method, the aforementioned approach is extended, by taking into account:

- The next frame of the video sequence: In this case, instead of restricting the analysis on the previous frames only, the next frame of the sequence is added. In essence a feature f_{cn}, analogous to f_{cp}, but extracted using the next frame, is added. This way, the motion trend is also taken into account. This feature will be referred to as "NF" in the rest of the paper.
- The accumulated travelled distance for each frame over the video sequence. This feature describes the total distance that a joint has travelled over the examined video sequence. This way, a sequence where a joint stays still, will have a different feature value than a sequence where the same joint made a motion and returned in the same position. This feature will be referred to as "TD" in the rest of the paper.
- In the f_{ci}, f_{cp} (and f_{cn} when used) features, only pairs of corresponding joints are used, instead of all joint pairs. This approach aims to reduce the size of the feature vector. The f_{cp} feature thus becomes: $f_{cp} = x_i^c - x_i^p | x_i^c \in X_c, x_i^p \in X_p$, with its size reduced from $N * N * 3$ to $N * 3$, where N is the number of joints. f_{ci} & f_{cn} are computed accordingly. This variation in the overall eigenJoints feature extraction process reduces noise from action irrelevant joints and was found to improve the recognition performance (See section 3); it will be referred to as "CO" in the rest of the paper.

Combinations of the above extensions were also applied and tested, as further explained in the experimental results presented in Section 3 below.

2.1 Use of manipulated objects

A series of domestic activities in realistic settings, such as food preparation, eating or drinking, typically involve the manipulation of household objects. By employing a state of art object recognition algorithm [3] so as to detect known household objects in the camera's field of view, and by detecting cases that the user's tracked hand joints come in contact to those, possible user interactions with specific objects of interest (e.g. cup for drinking, pills box for pill intake) can be inferred.

In our approach, this inference is first used so as to detect candidate segments where the above described action recognition method should be applied. Specifically, we take such points in time as the possible start time of an action, and create a segment from the incoming stream, with its length based on the average sequence length for the specific action from the training set. Moreover, in our method, we take into account the user's interacting object as further, semantic information that can further improve the discrimination among similar actions, which however involve different objects. Indicatively, this could lead to better discrimination between eating and drinking activities, which have a rather similar "hand to mouth" movement as their base, yet they involve different manipulated objects.

3 EXPERIMENTAL RESULTS

3.1 Evaluation on the MSR Action Dataset

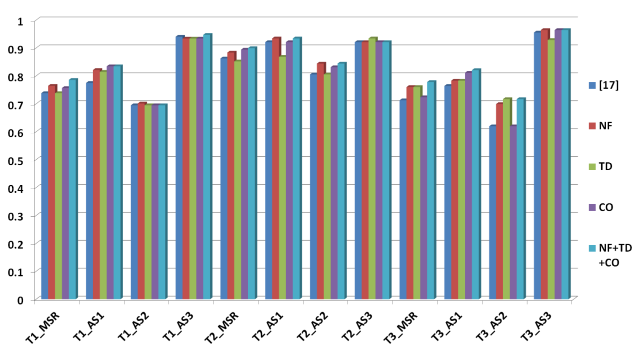

Figure 1: Overall Action Recognition Results on the MSR action dataset

In order to evaluate the proposed extensions to the eigen Joints algorithm, the tests presented in [17] were conducted using the original method, as well as the proposed variations. In short, the actions were separated in the three proposed action sets of [17], while also including an experiment using all the dataset actions. On these sets, three tests were conducted: T1 & T2 where 1/3 and 2/3 of the subset were used for training respectively, and T3 where subjects 1,3,5,7 & 9 were used for training and the rest for testing (cross subject test in [17]).

Overall action recognition results are displayed in Figure 1, where it can be seen that the proposed extensions (NF, TD and CO) improved the recognition performance in all the experiments, more so in the ones that used all the dataset's actions (*_MSR), as well as in the cross subject experiments (T3_*). Also, confusion matrices for the most challenging case of the cross subject action set 2 experiment, are presented in Figures 2 and 3, for the original method and the combination of the proposed extensions respectively. From

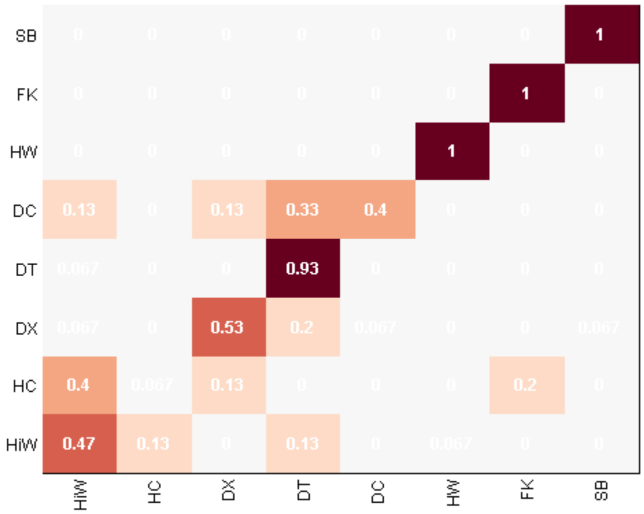

Figure 2: Confusion matrix on the MSR dataset using [17]. Activities Legend: HiW = High Wave, HC = Hand Catch, DX = Draw X, DT = Draw Tick, DC = Draw Circle, HW = Hands Wave, FK = Forward Kick, SB = Side Boxing.

these figures, it can be seen that, our proposed extensions to the original algorithm improve the action recognition results, especially between actions with similar movements, i.e. draw actions (DX, DT and DC) where our proposed method performs significantly better. For instance for action "Draw X" (DX) the original method has a correct classification score of 0.53, while ours has 0.73.

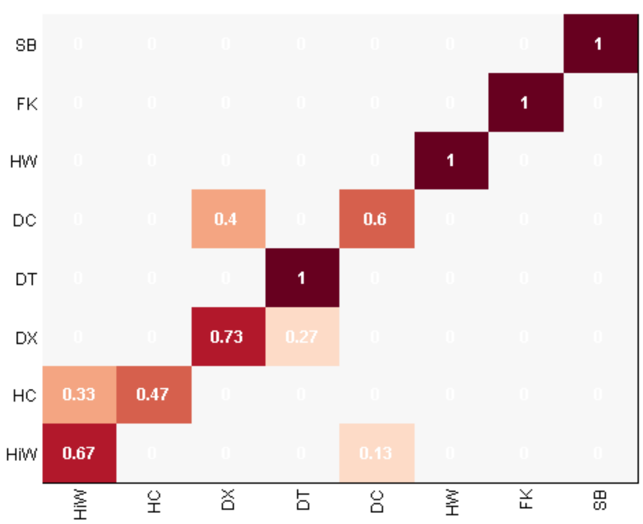

Figure 3: Confusion matrix on the MSR dataset using the proposed extensions. Activities Legend: HiW = High Wave, HC = Hand Catch, DX = Draw X, DT = Draw Tick, DC = Draw Circle, HW = Hands Wave, FK = Forward Kick, SB = Side Boxing.

3.2 Evaluation on proprietary domestic activities dataset

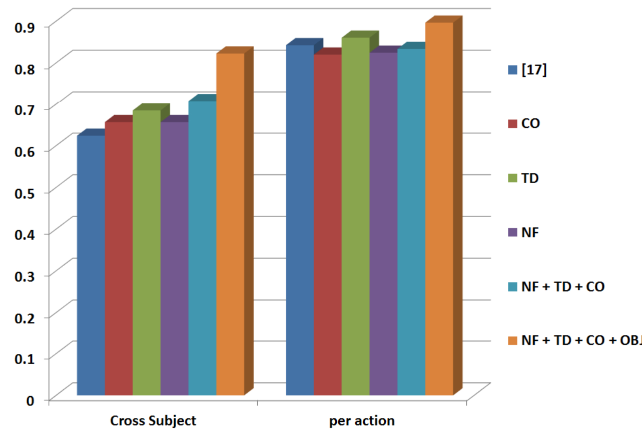

Figure 4: Recognition Results on the continuous actions dataset

The proposed method was also evaluated using a proprietary dataset, containing 20 actions involved in typical activities of daily living (ADL) which was collected using 15 subjects, and each subject performed all of the actions, 3 times each. The action sets included: *Take pill seated (A1), Reach mid seated (A2), Retract low (A3), Retract high (A4), Retract mid (A5), Reach low (A6), Reach high (A7), Reach mid (A8), Open Cupboard low (A9), Open Cupboard high (A10), Open Cupboard mid (A11), Eat seated (A12), Alter seated (A13) (i.e. manipulating a pill box at hand), Drink seated (A14), Drinking Standing (A15), Seat Idle (A16), Stand Idle (A17), Walk (A18), Sit2Stand (A19), Stand2Sit (A20),* . The indexes 'mid', 'low' and 'high' indicate that the person performing the action is bending, the action is taking place at chest or head height, or the person performing the action is stretching to reach the necessary height, respectively.

For testing in this dataset, two experiments were performed: *Cross subject*, where half the subjects were used for training and the rest for testing, and *per action*, where 2/3 of the subset was used for training (similar to T2 in [17]). The overall activity recognition results are presented in Figure 4. Also, confusion matrices are provided in Figures 5, 6 and 7, for the cross subject experiment using the approach from [17] and the proposed extensions, with and without the manipulated object, respectively.

As in the case of the MSR dataset, the proposed extensions again improved action recognition performance, while the use of the manipulated object improved results even further. More in detail, the use of the proposed extensions as well as the manipulated object, improved the action recognition performance in cases of actions with similar movement. For example, as shown in Figure 5 the original method had many mis-classifications between actions *Eat seated (A12), Alter seated (A13), Drink seated (A14)*, which involve similar hand movements. More in detail, in the original method the correct classification score for activity A12 was 0.42 (Figure 5) where in the proposed method improved to 0.87 (Figure 7). Similarly for activity A13, the improvement was from 0.47 in the original algorithm to 0.67 using the proposed one. Additionally, for actions *Open Cupboard low (A9), Open Cupboard high (A10), Open Cupboard mid (A11)* that can as well be considered very similar, the proposed extensions were found to help distinguish between them more effectively. In these cases, the correct recognition rate, was found to improve from 0.60 to above 0.80 (Figures 5 & 7 respectively).

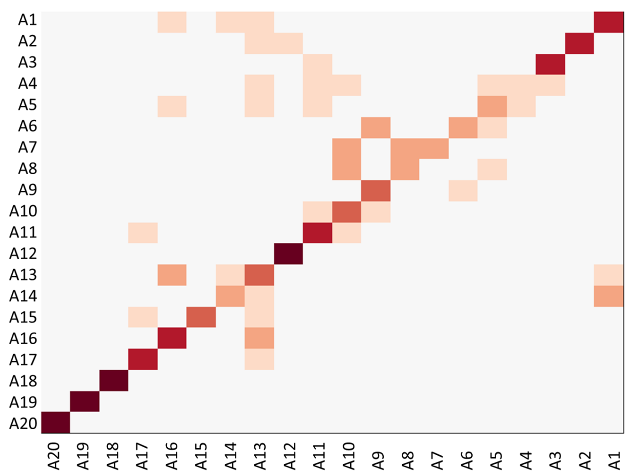

Figure 5: Confusion Matrix using [17] on a proprietary domestic activities dataset. Activities descriptions can be seen in Section 3.2. Darker colors represent values closer to 1, similarly to Figure 3.

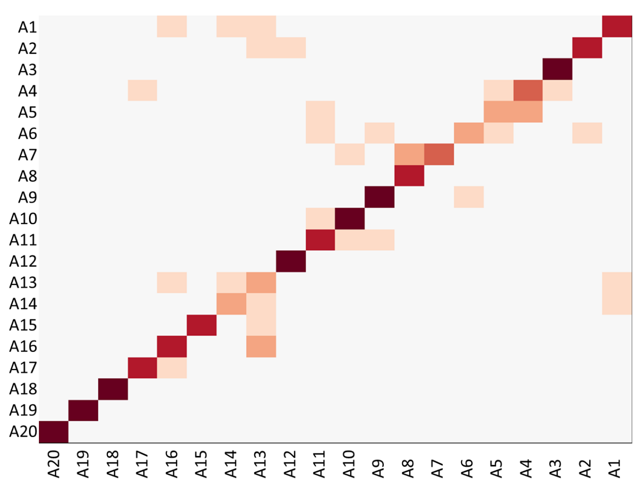

Figure 6: Confusion Matrix using proposed extensions on a proprietary domestic activities dataset. Activities descriptions can be seen in Section 3.2. Darker colors represent values closer to 1, similarly to Figure 3.

At this point, it should be underlined that the specific dataset, contrary to the MSR Action dataset, comprised action sequences

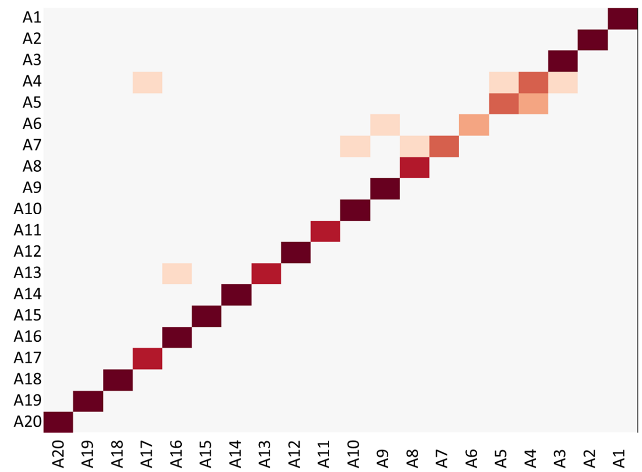

Figure 7: Confusion Matrix using proposed extensions and manipulated object on proprietary domestic activities dataset. Activities descriptions can be seen in Section 3.2. Darker colors represent values closer to 1, similarly to Figure 3.

that were not precisely segmented. In other words, sequences contained noise/random movement before and after the actual activity. In the activities that did not contain a manipulated object (e.g. *Stand2Sit*), a rolling window was used in order to perform the detection. On the other hand, for activities that contained a manipulated object (e.g. *Eat Seated* or *Drink Seated*), the manipulated object detection was used to correctly detect the onset of an activity and thus segment the sequence more precisely.

As was expected, the manipulated object detection was found to significantly improve the recognition results, not only due to the aforementioned use of it for more precise segmentation of a sequence, but also as it provided important contextual information, which the recognition algorithm used to distinguish between actions with very similar movement, e.g. drinking and eating, both hand to mouth -related movement, yet with different manipulated object. These improvements are illustrated by comparing for instance missclassifications rate between activities A13 & A14, where from 0.2 in Figure 6 drops to 0.0 in Figure 7, and A12 & A14 dropping from 0.1 to 0.0 respectively.

4 CONCLUSIONS

In this work, a novel human action recognition method for assistive service robots has been presented. Our method is based on the EigenJoints descriptor and advances the state-of-art by introducing novel features that express further kinematic characteristics of monitored actions, towards improved action recognition robustness. In addition, by incorporating an object recognition module, our overall action recognition framework manages to gain further improved performance and also to have the capacity to operate in realistic scenarios, which need automatic segmentation of input video streams. Experimental evaluation on the public MSR Action3D dataset showed that our novel features lead to improved action recognition performance. In addition, by incorporating also semantic knowledge related to the user's manipulated objects, action recognition performance was further improved over a proprietary dataset which focused specifically on actions involved in typical activities of daily living. The robust recognition of such actions is a key prerequisite for future assistive robots that aim to support MCI patients at home, by assessing user behaviour during her/his daily life, and inferring on proactive assistance provision.

ACKNOWLEDGMENTS

This work has been supported by the EU Horizon 2020 funded project "Robotic Assistant for MCI Patients at home (RAMCIP)", under grant agreement no: 643433.

REFERENCES

[1] J. K. Aggarwal and M. S. Ryoo. 2011. Human activity analysis: A review. *ACM Computing Surveys (CSUR), 43(3), 16* (2011).

[2] L. Chen, J. Hoey, C. D. Nugent, D. J. Cook, and Z. Yu. 2012. Sensor-Based Activity Recognition. *IEEE Transactions on Systems, Man, and Cybernetics, Part C (Applications and Reviews), vol. 42, no. 6, pp. 790-808* (2012).

[3] A. Doumanoglou, R. Kouskouridas, S. Malassiotis, and T.K. Kim. 2016. Recovering 6D Object Pose and Predicting Next-Best-View in the Crowd. *International Conference on Computer Vision and Pattern Recognition 2016 (CVPR 2016), Las Vegas, USA* (2016).

[4] D. Biswas et al. 2014. On the sensor choice and data analysis for classification of elementary upper limb movements. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Valencia, pp. 744-747* (2014).

[5] F. FLv and R. Nevatia. 2006. Recognition and segmentation of 3-d human action using hmm and multi-class adaboost. *European conference on computer vision (pp. 359-372). Springer Berlin Heidelberg* (2006).

[6] D. Giakoumis, G. Stavropoulos, D. Kikidis, M. Vasileiadis, K. Votis, and D. Tzovaras. 2014. Recognizing Daily Activities in Realistic Environments through Depth-Based User Tracking and Hidden Conditional Random Fields for MCI/AD Support. *European Conference on Computer Vision (ECCV '14), pp. 822-838* (2014).

[7] M. E. Hussein, M. Torki, M. A. Gowayyed, and M. El-Saban. 2013. Human Action Recognition Using a Temporal Hierarchy of Covariance Descriptors on 3D Joint Locations. *IJCAI (Vol. 13, pp. 2466-2472)* (2013).

[8] I. Kostavelis, D. Giakoumis, S. Malasiotis, and D. Tzovaras. 2015. RAMCIP: Towards a Robotic Assistant to Support Elderly with Mild Cognitive Impairments at Home. *2015 International Symposium on Pervasive Computing Paradigms for Mental Health (MindCare 2015), Springer International Publishing, pp. 186-195* (2015).

[9] W. Li, Z. Zhang, and Z. Liu. 2010. Action recognition based on a bag of 3D points,. *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition - Workshops, San Francisco, CA, pp. 9-14.* (2010).

[10] G. Papamakarios, D. Giakoumis, K. Votis, S. Segkouli, D. Tzovaras, and C. Karagiannidis. 2014. Synthetic ground truth data generation for automatic trajectory-based ADL detection. *IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Valencia, pp. 33-36* (2014).

[11] Y. Sheikh, M. Sheikh, and M. Shah. 2005. Exploring the space of a human action. *Tenth IEEE International Conference on Computer Vision (ICCV'05) Volume 1, pp. 144-149 Vol. 1.* (2005).

[12] J. Shotton, R. Girshick, A. Fitzgibbon, T. Sharp, M. Cook, M. Finocchio, R. Moore, P. Kohli, A. Criminisi, A. Kipman, and A. Blake. 2013. Efficient human pose estimation from single depth images. *Criminisi, A., Shotton, J. (eds.) Decision Forests for Computer Vision and Medical Image Analysis, Advances in Computer Vision and Pattern Recognition, pp. 175-192. Springer, London* (2013).

[13] M. Vasileiadis, D. Giakoumis, K. Votis, and D. Tzovaras. 2016. A Living Lab Infrastructure for Investigating Activity Monitoring Needs in Service Robot Applications. *In Proceedings of the 9th ACM International Conference on PErvasive Technologies Related to Assistive Environments (p. 72). ACM.* (2016).

[14] J. Wang, Z. Liu, Y. Wu, and J. Yuan. 2012. Mining actionlet ensemble for action recognition with depth cameras,. *2012 IEEE Conference on Computer Vision and Pattern Recognition, Providence, RI, pp. 1290-1297.* (2012).

[15] G. M. Weiss, J. L. Timko, C. M. Gallagher, K. Yoneda, and A. J. Schreiber. 2016. Smartwatch-based activity recognition: A machine learning approach. *2016 IEEE-EMBS International Conference on Biomedical and Health Informatics (BHI), Las Vegas, NV, 2016, pp. 426-429* (2016).

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

[17] X. Yang and Y. Tian. 2014. Effective 3d action recognition using eigenjoints. *Journal of Visual Communication and Image Representation, 25(1), 2-11.* (2014).

[18] X. Yang and Y. L. Tian. 2012. EigenJoints-based action recognition using NaÃŕve-Bayes-Nearest-Neighbor. *2012 IEEE Computer Society Conference on Computer Vision and Pattern Recognition Workshops, Providence, RI, pp. 14-19* (2012).

[19] M. Ye, Q. Zhang, L. Wang, J. Zhu, R. Yang, and J. Gall. 2013. A survey on human motion analysis from depth data. *In Time-of-flight and depth imaging, sensors, algorithms, and applications (pp. 149-187). Springer Berlin Heidelberg* (2013).

[20] H. Zhang and L. E. Parker. 2011. 4-dimensional local spatio-temporal features for human activity recognition. *2011 IEEE/RSJ International Conference on Intelligent Robots and Systems, San Francisco, CA, pp. 2044-2049* (2011).

[21] Y. Zhu, W. Chen, and G. Guo. 2013. Fusing Spatiotemporal Features and Joints for 3D Action Recognition,. *2013 IEEE Conference on Computer Vision and Pattern Recognition Workshops, Portland, OR, pp. 486-491.* (2013).