Appearance and Depth for Rapid Human Activity Recognition in Real Applications

Stavros Tachos
http://www.iti.gr/iti/people/Stavros_Tachos.html
Konstantinos Avgerinakis
http://www.iti.gr/iti/people/KAfgerinakis.html
Alexia Briasouli
http://www.iti.gr/iti/people/Alexia_Briasouli.html
Ioannis Kompatsiaris
http://www.iti.gr/iti/people/Ioannis_Kompatsiaris.html

Centre for the Research & Technology Hellas (CERTH)
Information Technologies Institute (ITI)
Thessaloniki, Greece

Abstract

Human activity recognition has gained a lot of attention in the computer vision society, due to its usefulness in numerous contexts. This work focuses on the recognition of Activities of Daily Living (ADL), which involves recordings constrained to specific daily activities that are of interest in assisted living or smart home environments. We present a novel technique for spatial activity localisation and recognition from colour-depth sequences, tailored to Activities of Daily Living (ADLs), which usually take place in relatively constrained environments. The proposed method significantly reduces the computational cost of activity recognition, while at the same time achieving a competitive accuracy rate, comparable to the State of the Art (SoA). This is achieved by the introduction of appearance and depth based spatiotemporal volumes, the Spatio-Temporal Activity Cells (STACs), extracted using appearance and depth information from successive video frames. A novel adaptive background modelling method follows, to characterize the STACs as "active" or "inactive" and accumulate them into foreground or background history volumes respectively. After activity detection using the STACs, activity recognition takes place using a novel, depth-based descriptor, the Histogram of Surface Normals Projections (HONSP), in combination with well known appearance descriptors (Histograms of Oriented Gradients, HOGs). Fisher encoding aggregates them into a fixed size vector to train a multiclass SVM model, which is then used for activity recognition. Experiments on different ADL datasets recorded with elderly people verify that the suggested algorithm is very appropriate for real life scenarios.

1 Introduction

Over the last decade, researchers have put a lot of effort in understanding various aspects of human activities for accurate activity recognition. Increased interest in ambient assisted living is making computer vision solutions for Activity localisation and Recognition (ALnR) necessary for monitoring and recognition of Activities of Daily Living (ADLs). This work

Pages 38.1-38.12
DOI: https://dx.doi.org/10.5244/C.29.38

presents an improved activity recognition approach, which achieves SoA accuracy at a reduced computational cost by leveraging information provided by colour-depth cameras. The central contributions of this work are:

1. The introduction of **Spatio-Temporal Activity Cells (STACs)** for adaptive spatial activity localisation through their dynamic characterisation as belonging to foreground or background (Sec. 3.2).

2. The introduction of the **Homogeneity** criterion (Sec. 3.2), which re-classifies foreground STACs as inactive when they stop changing, in a simpler manner than existing adaptive background subtraction methods [5], [23].

3. The introduction of novel, depth-based descriptors, the **Histogram of Surface Normals Projections (HOSNP)**, to replace computationally costly optical flow (OF) estimation and/or human detection that are commonly used in the current activity recognition SoA [8], [21]. The HOSNP supplement appearance information and help overcome issues caused by local noise and occlusions, while indirectly including motion-related information by accumulating changing appearance and depth data over time.

It should also be noted that the local nature of the STACs results in a fine boundary around subjects, unlike the coarse bounding box used in spatial localisation methods that are based on human detection and tracking [8, 16]. Furthermore, it goes beyond background subtraction, as it leverages depth information, enhancing the discrimination of moving from static regions in an adaptive manner (Sec. 3.2). Our proposed method leverages the STAC-based dynamic spatial localisation of activities for computationally efficient, yet accurate activity recognition, by simultaneously extracting trajectory, appearance (Histogram of Oriented Gradients, i.e. HOG) and depth descriptors (Histogram of Oriented Surface Normals Projections, i.e. HOSNP) from spatio-temporal volumes comprising of successive foreground STACs. The experiments show that, indeed, we achieve a significant reduction in computational cost, without loss in activity recognition accuracy. Experiments take place on three real-world datasets for monitoring elderly individuals from a dataset, which is available for research purposes upon demand.

This paper is organized as follows: Sec. 2 elaborates on ALnR related work for both RGB and depth video data. Sec. 3 analyses the methodology that we propose for activity localisation and recognition. Sec. 4 presents the experimental evaluation and comparisons with related work while we finalize this paper with conclusions and future work in Sec. 5.

2 Related Work

Early spatial activity localisation, i.e. background subtraction, were mainly variations of the GMM-based modeling of pixel intensities over time [5, 19]. This work goes beyond those approaches, by leveraging both colour and depth information for more accurate background removal, thanks to the increasing use of colour-depth cameras. Appearance information is enriched with depth data, to discriminate between changes in a pixel's intensity caused by actual motion, or by its physical location in space, at a different depth. Indeed, the inclusion of depth information leads to more accurate extraction of the foreground than standard, GMM-based background removal methods, as shown in Fig. 1.

Temporal activity localisation, i.e. activity detection, has been examined in [8], and more recently in [16], where a human detector and tracker was used. The authors trained an

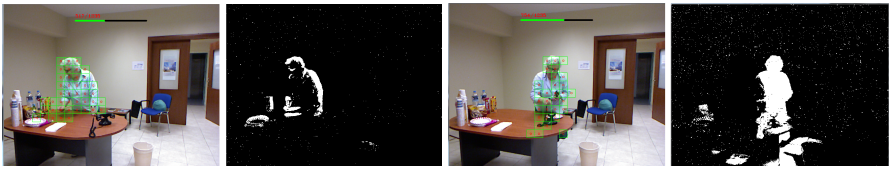

Figure 1: STAC activity localisation on the coloured figures against background subtraction proposed by [19] on the binary masks.

SVM as a human classifier, providing a great deal of upper human body training examples as input. The main disadvantage of this technique was the great computational cost of the classifier, and the large number of training samples needed, since the human detector could not be generalised to new video samples, as it led to a high false alarm rate without sufficient training. An alternative indirect activity localisation algorithm, very common in early techniques [4], [11], uses Spatio-Temporal Interest Points (STIP). However, this led to sparse features which led to reduced activity recognition accuracy.

While spatial activity localisation is common in activity detection, it has not been used often in activity recognition, until recently in [10], where the person's location is treated as a latent variable to be inferred simultaneously with activity recognition. This replaces expensive human detection and tracking with a simple and lightweight machine learning problem. However, [10] relies heavily on the modelling of the classifier, the amount of training data and their resemblance to the test videos, leading to solutions that cannot be easily generalized. Another Activity localisation and Recognition (ALnR) approach [17] introduced hierarchical space-time segments: two levels of hierarchies, with the human body and body parts respectively, were suggested to replace human detection and tracking. An unsupervised step models the localisation, and is followed by the recognition step, leading to improved accuracy. Spatial localisation algorithms in [20] improved activity recognition, inspired from spatial pyramids [12] that were used for image encoding. Spatial activity localisation also takes place with the computation of Motion Boundaries Activity Areas (MBAA), proposed in [1] for the recognition of Activities of Daily Living (ADLs).

Currently, activity recognition uses the depth sequences provided by colour-depth cameras [7] via features that rely on depth value differences between an interest point and its neighbourhood in [3], or features like HON4D that represent the surface normal distribution around the interest point [14] [18]. The main advantage of these methods is that they can be used in a wide range of applications as they have a natural ability to deal with occlusion, and lead to good performance in the recognition of complex actions with person-to-person or human-to-object interactions. Inspired by these works, we introduce a depth histogram, the Histogram of Surface Normal Projections (HOSNP). The HOSNP differs from HON4D as it uses the projections of surface normals to compute the final spatio-temporal histogram. The surface normal projections can be calculated quickly and with higher accuracy than the 3D surface normals used to estimate HON4D [3], leading to a faster and more precise activity representation. In particular, we compute the projections of the surface normals from finite differences for each pixel, whereas 3D surface normals require computationally costly SVD analysis to be calculated with accuracy.

We use a real world, publicly available (upon request) ADL dataset that contains a wider range of activities, not limited to the kitchen. Most benchmark ADL datasets currently avail-

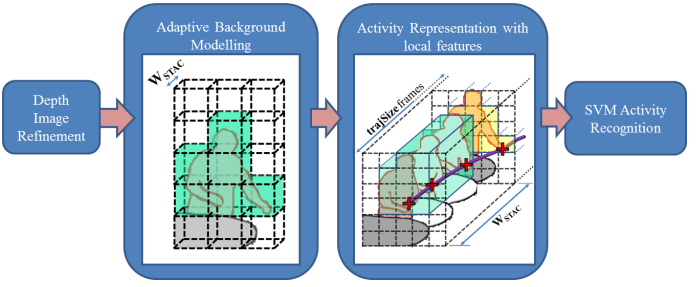

Figure 2: Overview of our ALnR solution: depth frame refinement corrects noisy depth values. Adaptive background modelling uses HOG and HoD to separate "active" from "inactive" STACs. HOG, HOSNP and 3D trajectories accumulated over time to represent human activities. Fisher encoding over the whole video trains a multi-class SVM model.

able only include colour information, such as the KIT [6] robo-kitchen dataset that monitors a kitchen environment and the ADLs that occur in it and the University of Rochester ADL dataset [13]. The MSR dataset, used in [3], will be included in future work, as it includes colour-depth information and has been used in senior monitoring applications, similarly to our data.

3 Methodology

This section describes the aim of this work which focus on exploiting both depth and intensity image characteristics for spatial activity localization and recognition. We introduce adaptive spatial activity localisation using so-called Spatio-Temporal Activity Cells (STACs), which are denoted as "active" or "inactive" based on the Histograms of Oriented Gradients (HOG) and new appearance metrics, the Histograms of Depth (HoD). Similar history-based adaptive background modelling methods rely only on appearance information [5], [23]. Activity representation follows by aggregating HOGs, 3D trajectories and a novel representation feature introduced in this work, the Histograms of Oriented Surface Normals Projections (HOSNP) that is based on the surface normals of [22]. The overall procedure of our system is depicted in Fig. 2.

3.1 Depth Frame Refinement

The depth data provided by colour-depth cameras eliminates the need for costly depth estimation, while providing a rich description of the scene. However, it is often noisy or has missing values (holes), usually caused by: (a) light reflection, (b) shadows around the human boundary or around objects that are attributed to different positions of the infrared camera and infrared projector position (which constitute the depth-estimating components of the colour-depth camera) and (c) high frequency light sources that add noise to the IR signal.

We follow a refinement strategy similar to [9] to eliminate holes in the depth image: the depth history of every pixel in a hole is examined over W_0 frames and its nonzero values are accumulated into a vector. This vector's median over time replaces the missing depth value, under the assumption that it contains the most likely value for that pixel. The resulting cor-

rected depth image does not contain holes, but the hole-filling strategy may have introduced outliers, e.g. if moving entities occluded that pixel during the W_0 frames. For this reason, it is spatially filtered with a median filter to eliminate any potential noise artefacts, resulting in a depth image that does not contain holes, and is spatially coherent.

3.2 Activity Localisation from Chi-Square Distance and Homogeneity

We perform spatially local adaptive background modelling by introducing a grid of Spatio-Temporal Activity Cells (STACs), which are dynamically characterised as "active" or "inactive". Each $STAC_i$ has dimensions 24×24 and is characterised by HOGs and Histograms of Depth (HoD), extracted around the grid points over W_{STAC} frames, with $W_{STAC} = 3$ chosen empirically to retain sufficient activity-related information.

Our adaptive background model uses two statistical criteria to determine if $STAC_i$ is active or inactive and to update the model appropriately: (i) the minimum Chi-square (χ^2) distance and (ii) the homogeneity criterion. Particularly, $STAC_i$ are accumulated over $N \simeq 2 \cdot fps$, forming two "History Volumes" for each one of them, which are characterised either as a Foreground History Volume ($FgHV_i$) or a Background History Volume ($BgHV_i$). All $BgHV_i$ are initialised with data from the first 15 frames, assuming that only background is present in them, while (a) the N most recent active STACs form the set of Foreground History Volumes (FgHV), (b) the N most recent inactive STACs form the set of Background History Volumes (BgHV), as shown in Fig. 3. Once a new HOG/HoD descriptor is extracted at each $STAC_i$, the algorithm computes its minimum χ^2 distance from the current BgHV: if it is below a predefined threshold th_{χ^2}, the cell is similar to the background and marked as "inactive", with the corresponding STAC incorporated in the $BgHV_i$. If the minimum distance is greater than th_{χ^2}, the algorithm pushes the current descriptor into the corresponding $FgHV_i$ and characterises it as "active".

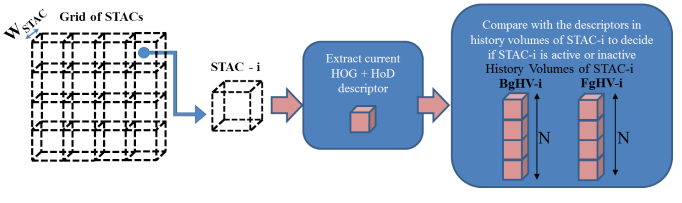

Figure 3: Adaptive background modelling: HOG/HoDs are extracted for each STAC and assign it to BgHV or FgHV. The BgHV or FgHV are dynamically reclassified according to the minimum chi-square distance and homogeneity.

When a $STAC_i$ is assigned to the $FgHV_i$, our method examines if it has been in the foreground for a long time: when foreground STACs do not change over time, we consider that they have become part of the background. This is determined by measuring their **homogeneity**, a new, intuitively meaningful metric that we introduce to define data similarity by taking the inverse mean pairwise Chi-Square distance among all elements of each history volume. As soon as $BgHV_i$ and $FgHV_i$ homogeneities are computed, they are compared to each other: if the $BgHV_i$ has larger homogeneity than $FgHV_i$, the two volumes remain unaltered. If the $FgHV_i$ is more homogeneous, the pixels in it have not changed during the last N frames, so it is reclassified as $BgHV_i$, representing the current background. Thus, when $FgHV_i$ cells stop

changing over time, they are integrated into a new background model.

3.3 Activity Recognition

As we progress through a video sequence, the set of all STACs that have been characterised as "active" form a 3D volume. Our activity representation algorithm runs in parallel to the adaptive background modelling and continuously samples this volume: the sampled points are tracked over time with the KLT tracker [2], as shown in Fig. 4. HOGs and HOSNP are then extracted in a $2 \times 2 \times 3$ grid around each tracked point. The HOSNP consists of a 9-bin histogram of the orientation of the projections of surface normals. Weikersdorfer et al. [22] estimate the projection of surface normals at each pixel as follows: first, a disk of radius R centered at $depth(x,y)$ is considered, and its projection $r_0(x,y)$ on pixel (x,y) is defined as:

$$r_0(x,y) = f \cdot R / depth(x,y), \tag{1}$$

where f is the sensor's focal length and $r_0(x,y)$ is the radius projection on (x,y). The normal vector p_n for each STAC pixel is computed from finite differences using the depth gradient:

$$p_n = \nabla depth(x,y) = \frac{1}{2d_W} \begin{pmatrix} depth(x+d_p,y) - depth(x-d_p,y) \\ depth(x,y+d_p) - depth(x,y-d_p) \end{pmatrix} \tag{2}$$

where d_W is a radius equal to $R/2$ and d_p is its corresponding projection on pixel (x,y), calculated by (1). The histogram of all p_n's in the STAC then forms the HOSNP.

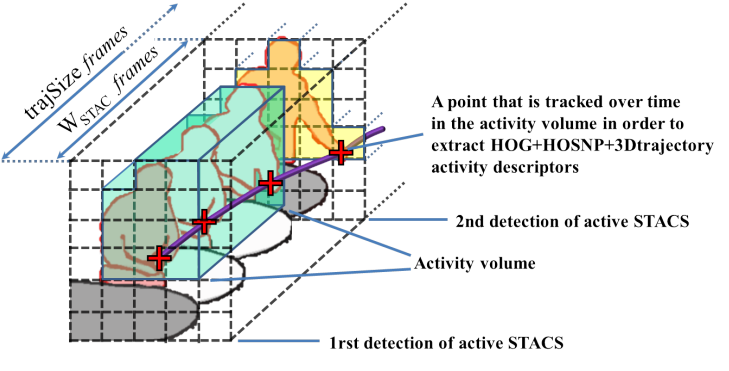

Figure 4: Activity representation in active STACs when the person stretches an arm: active STACs define the activity volume (in green) which is sampled to extract activity descriptors. Sampled points (crosses) that remain inside the activity volume continue to be tracked.

The HOG and HOSNP that lie in the same cuboid are accumulated and normalized and then concatenated with 3D trajectory coordinates $(x,y,depth(x,y))$ to form local activity descriptors. The activity recognition algorithm applies Fisher encoding on the descriptors [15], for a more compact dense representation than the commonly used Bag-of-Words approaches, and a multiclass SVM follows for activity recognition.

Figure 5: DemCare1 (top row), DemCare2 (middle row) and DemCare3 (bottom row) video frame examples. From left to right DemCare1: Eat Snack, Enter Room, HandShake, Read Paper, DemCare2: Serve Beverage, Start Phonecall, Drink Beverage and HandShake, DemCare3: Prepare Drug Box, Prepare Drink, Turn On Radio, Water Plant.

4 Experiments

Experiments take place on three datasets of elderly people carrying out semi-supervised ADLs, which are available for research purposes upon request[1]. The videos were recorded with a static camera, at a 640×480 pixel resolution and contain various activities, recorded in different environments, while also featuring anthropometric variations between the subjects, as they carry out the ADLs in different ways. Characteristic frames are shown in Fig. 5. The videos are split into training/testing following a One-Subject-Out-against-All logic in our experiments, so as to reduce the anthropometric variance on each SVM model.

We examined various combinations of the descriptors to assess their significance and role, so Tables 2-4 compare recognition accuracy with: (1) HOG, (2) HOSNP, (3) HOG-HOSNP, (4) HOG-HOSNP and 3D trajectory. Comparisons on accuracy and speedups are provided with SoA activity recognition methods [21], and methods that focus on ADL recognition [1]. The processing time of [21] is used as baseline for measuring time complexity, and is compared to the speedup achieved by our method and also that of [1]. The parameters for each video are shown empirically to be related to the fps of the video recordings: the temporal length of the BgHV and FgHV is approximately equal to $2\times$ fps, while the best trajectory length is approximately equal to the fps and the optimal threshold values are found to be near 10 (from 9 to 13), as shown in Table 1.

Activity Recognition on DemCare1: The first dataset, DemCare1, consists of 1 hour and 52 min recordings at 8 fps of 32 elderly people performing ADLs in a home-like environment. The camera is in front of the subject while they perform the following activities:

[1] Dem@Care action datasets

Dataset	fps	BgHV, FgHV temporal size N	threshold th_{χ^2}	trajectory size
D1	7.75 ± 0.61	20	9.5	9
D2	18.53 ± 1.98	29	9	15
D3	2.69 ± 0.70	9	13	6

Table 1: Experimental parameters in relation to the fps.

Activity	HOG	HOSNP	HOG + HOSNP	HOG+HOSNP +3D Traj	[1]	[21]
CU	69.4	88.9	**91.7**	88.9	85.3	81.1
DB	74.5	70.2	83.0	87.2	**87.8**	87.7
EP	57.6	75.8	84.8	**93.9**	84.4	90.9
ER	95.3	98.4	100.0	100.0	100.0	100.0
ES	39.1	54.3	58.7	**78.3**	71.1	89.6
HS	93.8	96.9	96.9	**100.0**	93.8	87.5
PS	41.2	50.0	61.8	64.7	68.6	**80.0**
RP	90.6	100.0	96.9	**100.0**	87.5	93.7
SB	50.0	82.4	73.5	85.3	82.4	**94.1**
SP	69.7	84.8	90.9	**93.9**	78.8	87.8
TV	90.6	93.8	**100.0**	93.8	96.9	**100.0**
Av. Accuracy	70.2	81.4	85.3	89.6	85.1	**90.2**
Speed	$\times 14.8$				$\times 2.2$	$\times 1$

Table 2: DemCare1 dataset accuracy over all classes for different combinations of descriptors (appearance only, depth only, combined appearance-depth, with and without the 3D trajectory). Average accuracy and time complexity in fps are reported, showing that the proposed method recognises activities with high accuracy and at a lower computational cost.

Cleaning Up (CU), Drink Beverage (DB), End Phonecall (EP), Enter Room (ER), Eat Snack (ES), Hand-Shake (HS), Prepare Snack (PS), Read Paper (RP), Serve Beverage (SB), Start Phone-call (SP) and Talk to Visitor (TV). Table 2 shows that the combination of appearance (HOG), depth (HOSNP) and 3D trajectory information led to high accuracy for most activities. The use of only appearance and depth information for "Cleaning Up" led to slightly better results than when the trajectory information was not included, which can be attributed to small inaccuracies in trajectory information. The method of [1] led to better results for Drink Beverage (+0.6%), Prepare Snack (+3.9%) and Talk to Visitor (+3.1%), showing that OF was more important than depth in these activities, as they feature small variations in depth. The use of colour-depth and 3D-trajectories led to an average accuracy of 89.6% over all classes, outperforming [1]. There is a significant improvement in activities that are very difficult to recognize, like End Phonecall (+9.5%), Eat Snack (+7.2%), Read Paper (+12.5%), as they feature small motions that are not extracted very accurately, so they are better distinguished through the use of appearance and depth. The proposed algorithm's mean processing speed was about 14.8 times faster than [21] and 6.7 times faster than [1]. **Thus, for this dataset our appearance-depth-3D trajectory descriptor leads to comparable average accuracy with the SoA (-0.6%), and is 14.8 times faster**.

Activity Recognition on DemCare2: DemCare2 consists of 1 hour and 59 min recordings at 18 fps, where 24 subjects perform ADLs in a different room than DemCare1, with the colour-depth sensor placed further from the activities. The ADL classes are: Cleaning Up (CU), Drink Beverage (DB), End Phonecall (EP), Enter Room (ER), Eat Snack (ES), Hand-Shake (HS), Prepare Snack (PS), Read Paper (RP), Serve Beverage (SB), Start Phonecall (SP), Talk to Visitor (TV) and Use Closet (UC). Most activity classes are the same as in the

Activity	HOG	HOSNP	HOG + HOSNP	HOG+HOSNP +3D Traj	[1]	[21]
CU	48.0	60.0	64.0	76.0	**88.6**	76.6
DB	17.6	11.8	26.5	20.6	35.6	**64.9**
EP	65.2	78.3	65.2	**95.7**	92.0	92.0
ER	96.0	96.0	96.0	96.0	96.0	89.6
ES	**94.3**	84.9	90.6	86.8	91.9	82.5
HS	86.4	81.8	90.9	**95.5**	91.7	88.0
PS	25.0	35.7	39.3	75.0	**81.3**	56.2
RP	95.7	87.0	91.3	95.6	**96.0**	88.0
SB	36.7	43.3	53.3	43.3	58.7	**63.6**
SP	91.7	91.7	91.7	91.7	**98.0**	81.4
TV	54.5	77.3	**95.5**	90.9	79.2	80.0
UC	91.3	91.3	**95.7**	91.3	90.0	96.0
Av. Accuracy	66.9	69.9	75.0	79.9	**83.2**	79.9
Speedup	×11.8				×2.3	×1

Table 3: DemCare2 dataset accuracy over all classes for different combinations of descriptors (appearance only, depth only, combined appearance-depth, with and without the 3D trajectory). Average accuracy and time complexity in fps are reported, showing that the proposed method recognizes activities with high accuracy and at a lower computational cost.

previous dataset, they take place in a different room, so they are not performed the same way. Table 3 shows that [1] led to better results for most activities, with the highest average accuracy of 83.2%, while the proposed approach achieved 79.9%. This can be attributed to the fact that the colour-depth camera is placed at a longer distance from the activities, providing less accurate depth information. Also, this recording included more motion, not sufficiently captured in the changing depth and appearance features. The combined use of appearance and depth led to the best results for TV and UC, as they are not characterised by significant motion, and depth plays a more important role in their description. The HOG+HOSNP descriptor led to an improvement of almost 5% over HOSNP, while the inclusion of 3D trajectory information led to a further improvement of almost 5% over HOG+HOSNP, resulting in 79.9% average accuracy. **Thus, for DemCare2 our method led to** -3.3% **average accuracy than [21] and [1], but was about** 11.8 **and** 5.1 **times faster respectively.**

Activity Recognition on DemCare3: DemCare3 includes 4 hours of 3 fps recordings of 25 subjects performing ADLs. The colour-depth camera is mounted on a higher level than the subject's head and the activities are: Answer the Phone (AP), Establish Account Balance (EAB), Prepare Drink (PD), Prepare Drug Box (PDB), Water Plant (WP), Read Article (RA), Turn On Radio (TOR). This dataset is more challenging than D1 and D2 as some activities are quite far away from the camera, giving poor depth information. Indeed, in Table 4, HOSNP achieves comparable, but lower mean average accuracy (by -1.6%) than HOG, as distant activities are not well discriminated. The combined use of HOG and HOSNP leads to improved results, giving an average accuracy of 88.6%, while the inclusion of 3D trajectory information leads to further improvements, resulting in the highest average accuracy, of 94.5%, while the literature achieves 93.3% in [1] and 91.7% in [21]. **Thus, our method achieves the highest average accuracy at a speed** 5.5 **times faster than [1] and** 21.4 **times faster than [21].**

Activity	HOG	HOSNP	HOG + HOSNP	HOG+HOSNP +3D Traj	[1]	[21]
AP	80.0	88.0	88.0	**100.0**	96.0	96.0
EAB	33.3	8.3	58.3	75.0	**83.3**	**83.3**
PDB	69.6	87.0	82.6	91.3	**95.2**	88.9
PHT	100.0	100.0	100.0	**100.0**	96.0	100.0
RA	95.5	81.8	100.0	**100.0**	90.9	85.7
TOR	100.0	92.9	100.0	**100.0**	96.4	100.0
WP	86.4	95.5	90.9	**95.5**	95.0	88.2
AvAcc	80.7	79.1	88.6	**94.5**	93.3	91.7
Speedup	×21.4				×3.9	×1

Table 4: DemCare3 dataset accuracy over all classes for different combinations of descriptors (appearance only, depth only, combined appearance-depth, with and without the 3D trajectory). Average accuracy and time complexity in fps are reported, showing that the proposed method recognizes activities with the highest accuracy and lowest computational cost.

5 Conclusions

This work presents a new method for activity representation and recognition that leverages the information provided by colour-depth cameras for accurate and computationally efficient recognition of human activities. Focus is on the recognition of ADLs, which is central in many applications, such as monitoring in assisted living and smart home environments. The depth data is used to build Histograms of Oriented Surface Normal Projections (HOSNP), which represent shapes and depth in the scene. HOSNP indirectly capture information about the motion in the scene, while avoiding the costly OF estimation, as changes in depth are indirectly related to motion. Appearance information is also included in our representation via HOGs, which complement the motion/depth description of the HOSNP, as they describe details inside objects. Experiments with publicly available datasets and comparisons with combinations of appearance/depth descriptors and existing methods demonstrate that the proposed algorithm achieves accuracy comparable to the SoA, while reducing the computational cost. Future work involves the extension of the proposed approach to more complex benchmark datasets, (such as the MSR action dataset), comparisons in terms of accuracy and speed with the SoA, e.g. HON4D [14] and Histograms of Surface Normals [18], and finally, the development of a full spatiotemporal activity localisation system.

6 Acknowledgements

This work is funded by the European Commission's 7th Framework Program (FP7 2007-2013), under grant agreement 288199 Dem@Care.

References

[1] K. Avgerinakis, A. Briassouli, and I. Kompatsiaris. Recognition of activities of daily living for smart home environments. In *9th International Conference on Intelligent Environments (IE2013)*, 2013.

[2] J. Bouguet. Pyramidal implementation of the affine Lucas Kanade feature tracker description of the algorithm. In *Intel Corporation*, pages 1–10, 2001.

[3] Z. Cheng, L. Qin, Y. Ye, Q. Huang, and Q. Tian. Human daily action analysis with multi-view and color-depth data. In *Computer Vision, ECCV 2012. Workshops and Demonstrations*, pages 52–61, 2012.

[4] O. Duchenne, I. Laptev, J. Sivic, F. Bach, and J. Ponce. Automatic annotation of human actions in video. In *Intelligent Environments (IE), 2013 9th International Conference on*, pages 1491–1498, 2009.

[5] A. Elgammal, D. Harwood, and L. Davis. Non-parametric model for background subtraction. In *6th European Conference on Computer Vision, ECCV 2000*, pages 751–767, 2000.

[6] L. Rybok et al. The kit robo-kitchen data set for the evaluation of view-based activity recognition systems. In *Humanoid Robots (Humanoids), 2011 11th IEEE-RAS International Conference on*, 2011.

[7] R. Romdhane et al. Activity recognition and uncertain knowledge in video scenes. In *Advanced Video and Signal Based Surveillance (AVSS), 2013 10th IEEE International Conference on*, 2013.

[8] A. Kläser, M. Marcin, C. Schmid, and A. Zisserman. *Trends and Topics in Computer Vision*, chapter Human focused action localization in video, pages 219–233. Springer, Berlin Heidelberg.

[9] J. Konecny and M. Hagara. One-shot-learning gesture recognition using HOG-HOF features. *Journal of Machine Learning Research*, 15(1):2513–2532, 2014.

[10] T. Lan, Y. Wang, and G. Mori. Discriminative figure-centric models for joint action localization and recognition. In *Computer Vision (ICCV), 2011 IEEE International Conference on*, pages 2003–2010, 2011.

[11] I. Laptev, M. Marszałek, C. Schmid, and B. Rozenfeld. Learning realistic human actions from movies. In *Conference on Computer Vision & Pattern Recognition*, Jun 2008.

[12] S. Lazebnik, C. Schmidt, and J. Ponce. Beyond bags of features: Spatial pyramid matching for recognizing natural scene categories. In *Computer Vision and Pattern Recognition, 2006 IEEE Computer Society Conference on.*, 2006.

[13] R. Messing, C. Pal, and H. Kautz. Activity recognition using the velocity histories of tracked keypoints. In *Computer Vision, 2009 IEEE 12th International Conference on. IEEE,*, 2009.

[14] O. Oreifej and Z. Liu. HON4D: Histogram of oriented 4D normals for activity recognition from depth sequences. In *Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on*, pages 716–723, 2013.

[15] F. Perronnin, Y. Liu, J. Sanchez, and H. Poirier. Large-scale image retrieval with compressed fisher vectors. In *Computer Vision and Pattern Recognition (CVPR), 2010 IEEE Conference on*, pages 3384–3391, June 2010.

[16] A. Prest, V. Ferrari, and Cordelia Schmid. Explicit modeling of human-object interactions in realistic videos. *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pages 835–848, 2013.

[17] I. Nazli S. Ma, J. Zhang and Stan Sclaroff. Action recognition and localization by hierarchical space-time segments. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 2003–2010, 2013.

[18] Nathan Silberman, Derek Hoiem, Pushmeet Kohli, and Rob Fergus. Indoor segmentation and support inference from rgbd images. In *Computer Vision–ECCV 2012*, pages 746–760. Springer, 2012.

[19] Chris Stauffer and W Eric L Grimson. Adaptive background mixture models for real-time tracking. In *Computer Vision and Pattern Recognition, 1999. IEEE Computer Society Conference on.*, volume 2. IEEE, 1999.

[20] M. Ullah, S. N. Parizi, and I. Laptev. Improving bag-of-features action recognition with non-local cues. In *British Machine Vision Conference (BMVC), 2010*, pages 95.1–95.11, 2010.

[21] H Wang and C. Schmid. Action recognition with improved trajectories. In *Computer Vision (ICCV), 2013 IEEE International Conference on*, pages 3551–3558, 2013.

[22] D. Weikersdorfer, D. Gossow, and M. Beetz. Depth-adaptive superpixels. In *Pattern Recognition (ICPR), 2012 21st International Conference on*, pages 2087–2090, 2012.

[23] X. Zhang, T. Huang, Y. Tian, and W. Gao. Background-modeling-based adaptive prediction for surveillance video coding. *Image Processing, IEEE Transactions on*, 23(2): 769–784, Feb 2014.