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

Introduction: As the demand for better assisted living and smart home environments continuously rises, the development of highly accurate and efficient human activity recognition algorithms becomes a necessity. We propose a novel technique for activity localisation and recognition from colour-depth sequences recorded with the Kinect sensor, specifically tailored for the recognition of Activities of Daily Living (ADLs). Comparative analysis with SoA [1, 2] on three challenging ADL datasets indicates that our algorithm is very appropriate for real life scenarios as it achieves SoA accuracy while performing 10-20 times faster. Our RGB-D video processing framework consists of four stages (Fig.1): i) the depth image is refined in order to fill the missing values produced by the sensor. ii) Activity is detected on a grid of Spatio-Temporal Activity Cells (STACs) over the video sequence. iii) Activity Representation takes place by extracting features from the 3D volume comprised of all STACs that contain activity, iv) the features are encoded with Fisher Vectors of fixed size and activity recognition is implemented with a SVM classifier.

Depth Refinement: Depth data provided by low cost colour-depth sensors contains missing values (depth holes) often caused by light reflection or high frequency light sources that add noise to the IR signal. Our refinement strategy operates pixelwise by maintaining a small set of the most recent depth values that are used to compute a median depth value for each pixel. Afterwards, holes are eliminated by filling missing values with the computed medians. Since this technique is being applied at every frame, the median values change through time and adapt to future variations on the depth image.

Activity Localisation: A grid is applied throughout the video sequence, comprised of *STACs* of size 24x24xW_{STAC}, with W_{STAC}=3 frames. Each grid cell over time contains a spatiotemporal $STAC_i$, which is characterised by a HOG and a Histogram of Depth (HoD), extracted around its center. Localisation is performed by our adaptive background model that exploits two statistical criteria to determine whether a $STAC_i$ is "active" (i.e foreground) or "inactive" (i.e. background): i) the minimum Chi-Square distance and ii) the homogeneity criterion. More specifically, two History Volumes of temporal size $N = 2 * fps$ are maintained for every $STAC_i$: the Foreground History Volume ($FgHV_i$) containing the N most recent "active" *STACs* and the Background History Volume ($BgHV_i$) containing the N most recent "inactive" STACs. The minimum Chi-square distance criterion indicates whether a $STAC_i$ can be considered as background based on its corresponding $BgHV_i$, while the homogeneity criterion serves to detect foreground *STACs* with no significant changes over time and need to become part of the background.

Activity Recognition: The STACs characterised as "active" through a video sequence form the 3D activity volume. This volume is continuously sampled, and sampled points are tracked over time using the KLT tracker, as long as they remain inside the volume. HOG (Histogram of Oriented Gradients) and HOSNP (Histogram of Surface Normal Projections) descriptors that represent the activity performed are extracted around every tracked point, concatenated with the points' 3D-Trajectory. The HOSNP, a 9-bin histogram of the orientation of the projections of surface normals offers significant surface information at a very low computational cost, mainly because it is extracted from finite depth differences of pixels within a small neighbourhood. Fisher encoding is applied on the HOG+HOSNP+3Dtrajectory descriptors computed in the former stage in order to get a compact representation of the activity performed. Final activity recognition is carried out by the use of a multiclass SVM.

Results: Experiments took place on three datasets (DemCare1, DemCare2, DemCare3) of elderly people performing ADLs, available for benchmark purposes upon request (http://www.demcare.eu/results/datasets). The

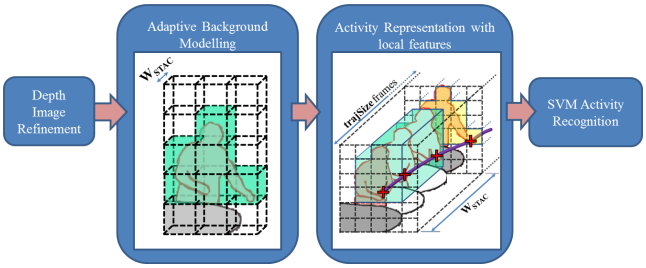

Figure 1: Overview of our Activity localisation and Recognition solution: (from left to right) a) depth frame refinement corrects noisy depth values. b) Adaptive background modelling uses HOG and HoD to separate *active* from *inactive* STACs. c) HOG, HOSNP and 3D trajectories are accumulated over time to represent human activities. d) Fisher encoding over the whole video trains a multiclass SVM model.

Datasets	HOG	HOSNP	HOG+ HOSNP	HOG+ HOSNP+ 3D Traj	[1]	[2]
DemCare1	70.2	81.4	85.3	89.6	85.1	**90.2**
DemCare2	66.9	69.9	75.0	79.9	**83.2**	79.9
DemCare3	80.7	79.1	88.6	**94.5**	93.3	91.7

Table 1: Average Accuracy of the proposed method for different combinations of descriptors and average accuracy of SoA methods[1,2]

640x480 videos of these datasets contain a variety of activities (e.g. Drink Beverage, Eat Snack, Talk to Visitor, Start Phonecall, End Phonecall, Prepare Hot Tea, Read Article, etc) and many anthropometric differences between subjects. Moreover, each dataset is recorded in a different environment at a unique sampling rate.

Results demonstrate that our method achieves accuracy that is highly competitive to SoA algorithms [1, 2] that make use of Optical Flow, while maintaining a very low computational cost. More specifically, the proposed method resulted in a -0.6% accuracy compared to SoA for D1 while performing 14.8 **times faster**. Similarly, a -3.3% accuracy deficit from SoA was reported for DemCare2 with a 11.8 **faster** computation, and lastly, our algorithm outperformed SoA on DemCare3 ($+1.2\%$ accuracy) while performing 21.4 **times faster**.

Further analysis exposes the value of the descriptors chosen for activity representation. Activity Recognition was carried out with different combinations of descriptors and as shown in table 1. The combination of HOG and HOSNP significantly increases the mean accuracy, demonstrating that these descriptors incorporate different aspects of the video data. Lastly, the concatenation of 3D trajectories boosts the accuracy even further, as more motion information is introduced to the final descriptor.

Conclusions: Our work proposes a new approach for feature-based activity localisation and recognition from RGB-D image sequences. Our proposed algorithm achieves SoA accuracy while maintaining a low computational cost. The development of a full spatio-temporal activity localisation system is scheduled as a future work.

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