
Peer reviewed version

Link to published version (if available):
10.1109/HealthCom.2015.7454583

Link to publication record in Explore Bristol Research
PDF-document

This is the author accepted manuscript (AAM). The final published version (version of record) is available online via IEEE at http://ieeexplore.ieee.org/document/7454583/. Please refer to any applicable terms of use of the publisher.

University of Bristol - Explore Bristol Research
General rights

This document is made available in accordance with publisher policies. Please cite only the published version using the reference above. Full terms of use are available: http://www.bristol.ac.uk/pure/about/ebr-terms
A Comparative Home Activity Monitoring Study using Visual and Inertial Sensors

L. Tao, T. Burghardt, S. Hannuna, M. Camplani, A. Paiement, D. Damen, M. Mirmehdi, I. Craddock
Visual Information Laboratory, Faculty of Engineering, University of Bristol, United Kingdom
Email: {lili.tao, tb2935, sh1670, massimo.camplani, csatmp, cxsda, m.mirmehdi, ian.craddock}@bristol.ac.uk

Abstract—Monitoring actions at home can provide essential information for rehabilitation management. This paper presents a comparative study and a dataset for the fully automated, sample-accurate recognition of common home actions in the living room environment using commercial-grade, inexpensive inertial and visual sensors. We investigate the practical home-use of body-worn mobile phone inertial sensors together with an Asus Xmotion RGB-Depth camera to achieve monitoring of daily living scenarios. To test this setup against realistic data, we introduce the challenging SPHERE-H130 action dataset containing 130 sequences of 13 household actions recorded in a home environment. We report automatic recognition results at maximal temporal resolution, which indicate that a vision-based approach outperforms accelerometer provided by two phone-based inertial sensors by an average of 14.85% accuracy for home actions. Further, we report improved accuracy of a vision-based approach over accelerometry on particularly challenging actions as well as when generalising across subjects.

I. INTRODUCTION

In this paper we focus on monitoring daily household activities in the home environment. It is here where patients, for whom activity monitoring is most challenging and necessary, spend most of their time after hospital discharge. Monitoring the level and type of patients’ physical activity is of general interest to clinicians across a wide variety of subjects, including obesity, diabetes, and depression-related research, as well as regarding aftercare for orthopaedic, cardiac and other surgery [1]. Traditionally, physical activity levels have been monitored using questionnaires, occasional clinical check-ups, and more recently, wearable devices – with a focus on a coarse categorisation of activity levels by wrist-worn inertial sensors [2]. To date, the use of wearable accelerometers remains a popular choice as source for inferring human activity levels due to its low cost, low energy consumption and data simplicity. Among these, triaxial accelerometers are the most broadly used motion sensors to recognise ambulation activities [3].

Visual sensor based techniques have emerged over recent years for which there exists a significant body of literature describing the inference of activities from 2D colour intensity imagery [4]. However, the increasing availability of depth-measuring sensors, especially the introduction of the Microsoft Kinect, has generated an opportunity for utilizing depth in conjunction with traditional RGB camera data allowing for richer and more fine-grained analysis of human activity [5].

Recent work by Chen et al. [3] presents a comparative study of such RGB-D (colour and depth imaging) sensors versus accelerometer sensors. The work also introduces a fusion approach for both modalities of data. They show that RGB-D and accelerometer data can be used to generate comparable results when tested on the Berkeley MHAD dataset [6]. This dataset is, however, recorded in a laboratory environment where most actions in the dataset are related to body exercises (jump, punch etc.).

Developing a reliable home monitoring system has drawn much attention in recent years due to the growing demands for integrated health care. Existing approaches to current home monitoring systems often include custom-fit environmental, physiological and vision sensors, such as in the SPHERE project [1]. Such systems can enable several types of application, to increase personal safety for elderly patients and to facilitate clinicians to diagnose and monitor patients. This new patient-clinician interactive mode improves the reliability and effectiveness of diagnosis to some extent, significantly shortens the travel time and hospital stay for patients, and reduces the work load for clinicians [7]. Visual sensors in particular have the potential to address several limitations of current systems [8]: they are data-rich and able to capture multiple events simultaneously, and they are easy to integrate into already existing living environments.

The paper has two key contributions. Firstly, we introduce a dataset for fine-grained action recognition within a real home environment in the SPHERE project’s house [9]. The dataset, exemplified in Figure 1, contains 13 household actions performed over 10 sessions - a total of 130 sequences. The setup consists of 1) an Asus Xmotion RGB-Depth camera mounted at the corner of a living room, and 2) two three-axis mobile phone accelerometer sensors worn at the waist and the dominant wrist. Secondly, we present a comparative study towards activity monitoring using these visual and accelerometer sensors in a living environment. We outline areas where a visual approach exceeds the performance of an accelerometer sensor, showing its merits in (a) detecting particularly challenging actions, and (b) in generalising across subjects.

II. VISUAL DATA COLLECTION AND PROCESSING

Visual Data Collection. We simultaneously collect RGB and depth images using the commercial product Asus Xmotion. Motion information can be recovered best from RGB data as it contains rich texture and colour information. Depth information, on the other hand, reveals details of the 3D configuration. We extract and encode both motion and depth features over the area of a bounding box as returned by the human detector and tracker provided by the OpenNI SDK [10]. Figure 2 gives an overview of the feature extraction process and illustrates the motion and depth information extracted. To
normalise the utilised image region due to varying heights of
the subjects and their distance to the camera, the bounding
box is scaled by fixing its longer side to $M = 60$ pixels
while maintaining aspect ratio. The scaled bounding box is
then centred in a $M \times M$ square box and horizontally padded.

Motion Feature Encoding. Motion information can generally
be readily extracted from this box independent of varying hu-
man appearance. Inspired by [11], optical flow measurements
are taken and split into horizontal and vertical components.
These are re-sampled to fit the normalised box and a median
filter with kernel size $5 \times 5$ is applied to smooth the data
in each direction. The normalised bounding box is divided
into an $N \times N$ non-overlapping grid and the orientations of
each grid cell are quantised into $n_b$ bins. The parameters
for our experiments are empirically determined as (optimal)
values of $N = 2$, $n_b = 9$. The second and third rows in
Figure 3 show optical flow patterns and its motion features
for different actions. Here, we only show the magnitude of
horizontal flow $F_H$ and vertical flow $F_V$ in one figure to save
space.

These patterns in hand, a local motion feature descrip-
tor $F = F_H \cup F_V$ is constructed by concatenating the histogram
of optical flow features in each block from both orientations.
To encode motion spatio-temporally, we choose a temporal
window of 15 frames suggested in [6] which is approximately
half a second around the current frame to be concatenated for
establishing the final descriptor.

Depth Feature Encoding. Information computed from struc-
tured light alleviates the effect of appearance and lighting
variations, allowing for independent depth recovery. For home
environments with partial occlusions and unconstrained object
interaction, however, we found that the performance of avail-
able skeleton trackers [12] is unreliable – specially when the
subject is not facing the camera. Instead, we opt to extract
features directly from depth imagery employing the histogram
of gradients (HOG) feature on raw depth images [13] applied
to the normalised box. The last row in Figure 3 shows the
visual representation of the HOG feature for different actions.
Essentially, these descriptors are able to encode a person’s
silhouette, its contours and the edges and depth gradients
within its area.

The complete visual feature descriptor consists of appear-
ance features extracted from the depth image of the current
frame and the 15-frame motion context from colour images.

III. INERTIAL DATA COLLECTION AND PROCESSING
As shown in [3], accelerometers placed on wrist and waist were
found to be the most effective for human action recognition.
Having more sensors [6] may improve performance, but it is
Different subjects may perform actions differently and can vary significantly across snapshots from "Stretching" due to diverse subjects and living habits. Figure 4 shows snapshots from "stretching" as an example that actions may be performed differently and can vary significantly across different subjects.

V. Experimental Results and Evaluation

For evaluation, we perform leave-one-subject-out cross validation (CV1) where final recognition results reported are averaged over all subjects to remove any bias. We utilise a Support Vector Machine (SVM) with a Radial Basis Function (RBF) kernel to classify the data. Table II lists the number of frames for each action and sequence. Associated with this, we show the percentage of the total number of frames for each action. The recognition rates reported in the table include appearance features only (HOG), motion features only (FLOW), appearance and motion feature fusion (FLOW+HOG), accelerometer data (ACC), and the fusion of both visual and accelerometer data (Visual+ACC).

In general, the overall recognition rate of visual sensors is found to be 14.85% higher than that of accelerometers. Notice that a substantial recognition improvement can be attributed to actions, for which part of the containing movements are similar. For example, sitting down and standing up are classified as “picking up” by accelerometers, but can be recognised by cameras due to different body poses. Some actions, e.g. wiping the table and dusting, are confused by both sensors, as these actions are performed in a very similar way. It can also be observed that a combined visual-inertial approach does not lead to a significant recognition improvement than when using only visual data. Figure 5 depicts the recognition confusion matrices corresponding to the use of inertial and visual sensors, respectively.

Activities performed by different subjects may vary significantly. To investigate the transfer of learned action descriptions across different subjects, we conduct an experiment by using one sequence for training and another sequence from the same subject for testing (CV2). The results listed in Table II show the overall recognition rate over all actions for CV1 and CV2. It is noticeable that in CV2 there is a significant improvement of accelerometer performance compared to the CV1 test, while similar results are produced by visual data and the fusion of visual and accelerometer data. In practice, it is unrealistic to have all the users’ data available for system training. This demonstrates one advantage of using visual sensors for action recognition.
Comparing vision and inertial sensors for actions in the home

tions among different applications, residents, and households.
sensors are low cost, easy to operate, and suitable for deploy-

cover home-typical human activities in 130 sequences of 70

The confusion matrices by (a) accelerometer data and (b) visual data.

VI. CONCLUSION AND FUTURE WORK

This paper presented a comparative study and relevant
dataset for the fully automated recognition of common home
activities via inertial (accelerometer) and visual sensors. We
introduced the challenging SPHERE-H130 action dataset\(^3\) that
covers home-typical human activities in 130 sequences of 70
minutes of multi-modal recordings. Both vision and inertial
sensors are cost, easy to operate, and suitable for deploy-
ment among different applications, residents, and households.
Comparing vision and inertial sensors for actions in the home
environment, results indicated that a vision-based approach
outperforms body-worn accelerometer sensors by an average of

14.85% accuracy for the dataset. A combined descriptor only
marginally outperformed vision descriptors. More importantly,
we found that vision provides better generalisation across
subjects and is able to differentiate some complex actions
that accelerometry fails to decouple. We conclude that visual
approaches should play a role in future monitoring systems
for the home. Future work will include comparisons between
different modalities for particular target variables including
energy expenditure and related monitoring tasks.

ACKNOWLEDGMENT

This work was performed under the SPHERE IRC funded
by the UK Engineering and Physical Sciences Research Coun-
cil (EPSRC), Grant EP/K031910/1.

REFERENCES

D. Kaleshi, M. Mirmehdi, P. Flach, and I. Coddock, “Bridging the internet
and the internet of things: The SPHERE project,” IEEE Intelligent

expenditure using body-worn accelerometers: a comparison of methods,
sensors number and positioning,” Biomedical and Health Informatics,

recognition using fusion of depth camera and inertial sensors,” Human-
2015.


mhad: A comprehensive multimodal human action database,” in IEEE
Workshop on Applications of Computer Vision, 2013, pp. 53–60.

Zwinderman, “Prospective assessment of falls in parkinson’s disease,”

on vision techniques applied to human behaviour analysis for ambient-
assisted living,” Expert Systems with Applications, vol. 79, pp. 1057–1088,
2017.

in a residential environment,” in IEEE International Conference on
Communications, Workshop on ICT-enabled services and technologies
for eHealth and Ambient Assisted Living, 2015.

Available: http://www.openni.org/documentation

learning,” in European Conference on Computer Vision, 2008, pp. 548–
561.

[12] A. Paement, L. Tao, S. Hannuna, M. Camplani, D. Damen, and
M. Mirmehdi, “Online quality assessment of human movement from

detection,” in Computer Vision and Pattern Recognition, 2005, vol. 1,
pp. 886–893.

1546.

“A triaxial accelerometer and portable data processing unit for the
assessment of daily physical activity,” Biomedical Engineering, IEEE

machines,” ACM Transactions on Intelligent Systems and Technology,
vol. 2, no. 3, p. 27, 2011.

\(^3\)The dataset is released on SPHERE website http://www irc-sphere.ac.uk/
work package-2/ar

<table>
<thead>
<tr>
<th></th>
<th>ACC</th>
<th>Visual</th>
<th>Visual+ACC</th>
</tr>
</thead>
<tbody>
<tr>
<td>CV1</td>
<td>56.67%</td>
<td>71.52%</td>
<td>73.99%</td>
</tr>
<tr>
<td>CV2</td>
<td>70.16%</td>
<td>72.12%</td>
<td>75.01%</td>
</tr>
</tbody>
</table>

TABLE II: Overall recognition rate (%) for CV1 vs CV2 tests.