Towards Recognising Collaborative Activities Using Multiple On-Body Sensors

Jamie A. Ward
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany
jamie@jamieward.net

Peter Hevesi
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany
Peter.Hevesi@dfki.de

Gerald Pirkl
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany
gerald.pirkl@dfki.de

Paul Lukowicz
German Research Center for
Artificial Intelligence (DFKI)
Kaiserslautern, Germany
paul.lukowicz@dfki.de

Abstract

This paper describes the initial stages of a new work on recognising collaborative activities involving two or more people. In the experiment described a physically demanding construction task is completed by a team of 4 volunteers. The task, to build a large video wall, requires communication, coordination, and physical collaboration between group members. Minimal outside assistance is provided to better reflect the ad-hoc and loosely structured nature of real-world construction tasks. On-body inertial measurement units (IMU) record each subject's head and arm movements; a wearable eye-tracker records gaze and egocentric video; and audio is recorded from each person's head and dominant arm. A first look at the data reveals promising correlations between, for example, the movement patterns of two people carrying a heavy object. Also revealed are clues on how complementary information from different sensor types, such as sound and vision, might further aid collaboration recognition.

Author Keywords

Wearable sensing; Datasets; Activity recognition

ACM Classification Keywords

I.2.m [Artificial Intelligence]: Miscellaneous

UbiComp/ISWC '16 Adjunct, September 12-16, 2016, Heidelberg, Germany
ACM 978-1-4503-4462-3/16/09. http://dx.doi.org/10.1145/2968219.2971429

Building a Video Wall

- 6 screens tiled in 2x3 formation, 2.5m high
- each 105x60 cm screen weighed 8kg
- 50 screws, 12 screen spacers, and 2 base panels
- parts stored 25m from assembly area
- 4 people took 45 min to build and take apart

Introduction

To date, most work on activity and context recognition has concentrated on recognizing what an individual does and how he or she interacts with the environment (e.g. [4] for a recent survey). Effects from the presence of multiple users were often seen as a “disturbance” (e.g. the “multiple occupancy” problem within the smart home activity recognition domain [1]). The idea of recognising group activities using wearable sensors was recently explored in the doctoral thesis of Dawud Gordon [2]. The iGroups project develops this idea by exploring group activities, structures, and dynamics, with a focus on sensing and recognising collaborative activities within groups[1]. This paper describes the initial steps of iGroups and introduces a dataset based on a group of people collaborating on the task of building a video wall.

The choice of task is motivated by the need to detect worker activities for safety and documentation purposes, for example, on construction sites, during field repairs, or in emergency response scenarios. In these situations collaboration and interaction between people is essential but does not necessarily follow strictly defined workflows. The task is therefore designed to be loosely-structured with participants given the freedom to work as they need. On-body sensing is used to reflect the fact that this sort of work is often done in places where other sensing resources cannot be assumed.

The Task

The overall task was for 4 people (3 male, 1 female) to collaborate in assembling, and then dismantling, a large video wall. The sidebar (left) shows the completed wall and lists what is required to build it.

A short description and guide to the task was given to the subjects, beyond this the group had to organize and execute the work themselves. At least two people were needed to carry each (8kg) screen from a storage room, which is 25m away from the assembly area. Other components, such as spacers and tools, could be carried by one person. Once enough components were at the destination area, the group could start to build the wall by lifting and mounting the screens onto the base panels. Lifting required cooperation of at least two people, whereas tasks such as tightening the screws could be done by one person. The activities also varied in length, from nearly a minute for two people to carry a screen along a corridor, to a few seconds for one person to place a spacer in its track on the video wall. After 20 minutes of set-up and synchronisation, it took 45 minutes for the subjects to complete the task.

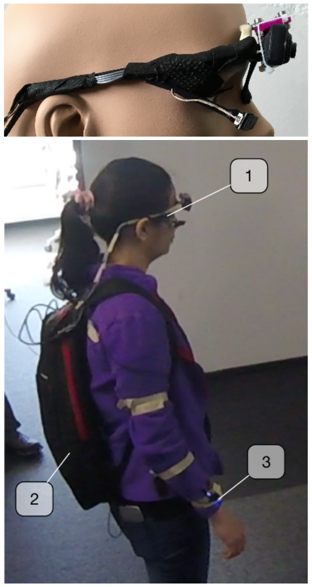
Data Collection

Each subject was equipped with sensors on the arms and on the head to monitor their movements and environment (see sidebar on recording setup). The open-source, head-mounted eye tracking platform, Pupil, was used to record egocentric video, audio data, and eye gaze information of each subject [3]. Data collection was handled by Pupil software running on a laptop carried in each subject’s backpack. Head movements were tracked using additional inertial measurement units (IMU) attached to each eyetracker headband[2]. Each subject also wore IMU devices on both arms[3]. An additional microphone was worn on the dominant arm of each subject (in this instance all subjects were right-handed). To cope with bandwidth limitations of the recording setup, this audio was recorded using the Voice Memos app on an iPhone5 in each person’s pocket. To aid

[1] German Federal Ministry of Education and Research (BMBF) funded

[2] Bosch BNO055: https://www.bosch-sensortec.com/bst/products/all_products/bno055

[3] Sparkfun 9 DOF: https://www.sparkfun.com/products/13033



Recording setup

1. video camera, microphone, video eye tracker; extended with IMU to track head movement
2. laptop in backpack
3. left and right arm IMUs, with microphone on right arm

with synchronisation, a series of clapping and jumping gestures were performed by participants both at the beginning and half-way through the task. Total dataset size is 30GB (including 25GB of video).

Initial Data Exploration

Although a full analysis of the data (exploring eye tracking, physical orientation, etc.) is beyond the scope of this paper, the following gives a snapshot of the kind of collaboration information that is provided by sound, hand acceleration and egocentric video. It shows how they might be combined to help tackle the challenges of recognising 1) physical collaboration (carrying a screen) and 2) interaction within a group (two people talking).

Physical collaboration clues: sound and acceleration
Figure 1 shows the sound signals and right-wrist acceleration (with x,y & z axes combined) of the 4 subjects over 15 seconds. During this example P2 and P3 are carrying a screen. Each holds onto the screen with his right hand as he walks. This is reflected in their acceleration signals which reveal not only a regular movement pattern, but also a high degree of correlation. During the 5s period highlighted on Figure 1 however, the acceleration signal of P1 seems correlated with those of P2 and P3. Does this imply that she too is helping carry the screen? Perhaps the larger acceleration amplitude indicates that her hands are freer and that she is simply walking in sync with the others? This confusion is settled by looking at the sound signals. Moving the heavy screen is noisy, and those carrying it periodically bump into it as they move. In this example P2 and P3 are also talking. But these sound signals are missing from P1's recording indicating that it is unlikely that she is involved. (She was walking in another room.)

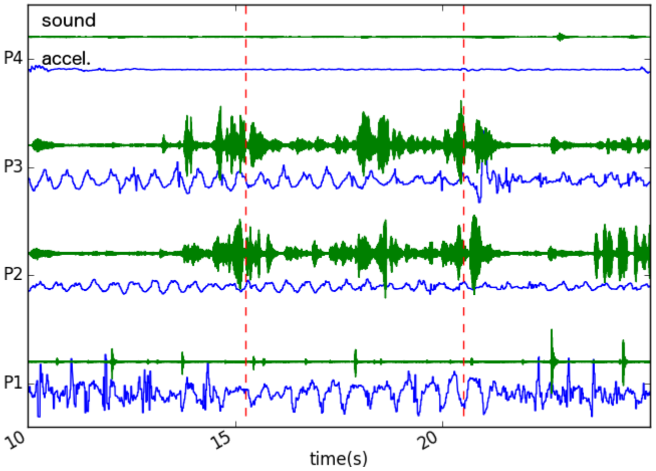


Figure 1: sound and combined acceleration (right wrist) signals from P1 to P4. P2 and P3 are carrying a screen. Acceleration for P2,P3, and P1 appears correlated (between the dotted lines). The sound signals suggest collaboration with P1 is unlikely.

Interaction clues: sound and video
Interaction between group members is a useful indicator of collaboration. The front-facing cameras can be used to reveal who or what each person is looking at. An off-the-shelf face detection algorithm (Haar feature cascade classifier from OpenCV [5]) can be used to highlight the most prominent face in each person's view, as shown in Figure 2.

Audio from the head-worn microphones can then be used to determine who is speaking. A simple heuristic was used that compares the levels of sound picked up by each person's microphone falling within the frequency range of speech. Combining speaker detection with face detection offers the potential to enhance the accuracy of both methods for person identification and for detecting interaction.

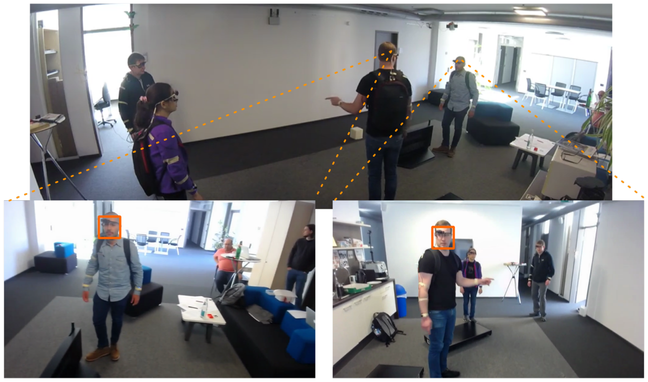


Figure 2: The face detection algorithm is tuned to detect nearby faces, which can be a sign for social interaction especially when combined with speech detection.

Challenges Ahead

The next stage of this work is to investigate different approaches for identifying collaborations and recognising activities, e.g., using time-series similarity measures, and machine-learning (ML). There are 3 main challenges that must be addressed:

Annotation– how do we define and label activities such as 'carry' in a way that is consistent and compatible between participants (e.g., the same activity may begin at different times for both collaborators)?

Missing data– in any multi-sensor recording there are instances of noisy, corrupt, or missing sensor data. ML methods need to be able to cope with this.

Variations– there are many ways to perform a physical activity. The problem is compounded with collaboration (e.g., 'carry' might involve one person walking forward, the other backwards). One solution might be to tackle recognition not only at the raw signal level, but at a higher, abstract level, perhaps using complex features such as body posture.

Many sub-activities in this dataset (e.g. using a screwdriver, lifting, walking) have been studied before and are recognisable using common machine learning methods – at least when carried out by one person. This work builds on these to introduce the added complexity of performing activities in collaboration, and highlights some of the interesting challenges that lie ahead.

REFERENCES

1. Diane J Cook Geetika Singla and Maureen Schmitter-Edgecombe. 2010. Recognizing independent and joint activities among multiple residents in smart environments. *Journal of ambient intelligence and humanized computing* 1, 1 (2010), 57–63.
2. Dawud Gordon. 2014. *Group Activity Recognition Using Wearable Sensing Devices*. Ph.D. Dissertation. PhD thesis.
3. Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. (April 2014). http://arxiv.org/abs/1405.0006
4. O. D. Lara and M. A. Labrador. 2013. A Survey on Human Activity Recognition using Wearable Sensors. *IEEE Communications Surveys Tutorials* 15, 3 (2013), 1192–1209. DOI:http://dx.doi.org/10.1109/SURV.2012.110112.00192
5. Paul Viola and Michael Jones. 2001. Rapid object detection using a boosted cascade of simple features. In *Computer Vision and Pattern Recognition, 2001. CVPR 2001. Proceedings of the 2001 IEEE Computer Society Conference on*, Vol. 1. IEEE, I–511.