Detecting Physical Collaborations in a Group Task Using Body-Worn Microphones and Accelerometers

Jamie A Ward, Gerald Pirkl, Peter Hevesi and Paul Lukowicz
German Research Center for Artificial Intelligence (DFKI)
Kaiserslautern, Germany
Email: jamie@jamieward.net, gerald.pirkl@dfki.de, Peter.Hevesi@dfki.de, paul.lukowicz@dfki.de

***Abstract*—This paper presents a method of using wearable accelerometers and microphones to detect instances of ad-hoc physical collaborations between members of a group. 4 people are instructed to construct a large video wall and must cooperate to complete the task. The task is loosely structured with minimal outside assistance to better reflect the ad-hoc nature of many real world construction scenarios. Audio data, recorded from chest-worn microphones, is used to reveal information on collocation, i.e. whether or not participants are near one another. Movement data, recorded using 3-axis accelerometers worn on each person's head and wrists, is used to provide information on correlated movements, such as when participants help one another to lift a heavy object. Collocation and correlated movement information is then combined to determine who is working together at any given time. The work shows how data from commonly available sensors can be combined across multiple people using a simple, low power algorithm to detect a range of physical collaborations.**

I. Introduction

Activity modelling and recognition is traditionally concerned with recognising what an individual does and how he or she interacts with their environment. The definition of what is meant by 'activity' can vary in all of these approaches, from low-level 'actions' such as a single movement of the wrist, to higher level situations such as 'dinner'. The long-held goal of many working in this field is to be able to draw together multiple strains of contextual information, including activity, to help develop new context-aware applications. Many ways of handling this information for high-level modelling and decision making have been explored, for example see overviews by Ye *et al.* [1] or Bettini *et al.* [2]).

A question that often arises is how do we sense, and model, activities that are carried out by more than one person? 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 [3]). The idea of recognising group activities using wearable sensors was recently explored in the doctoral thesis of Dawud Gordon [4]. Gordon showed that by sensing the physical movements of multiple individuals, the emergent behaviour of a group could be predicted (e.g. certain team sports like football, or volleyball) [5].

The iGroups project at DFKI aims to explore group activities, structures, and dynamics with a focus on sensing and recognising collaborative activities within groups. This paper describes an initial approach to that work. The main contributions of the paper are (1) to describe a 4 person, 45 minute long, dataset of an ad-hoc group construction task, and (2) to show how physical collaborations can be detected using a relatively simple heuristic based on correlations of features from a small number of body-worn sensors [1].

A. Related Work

a) Vision: Many overviews on activity recognition have been published in recent years, with specific orientation around the particular sensing approach taken. Turaga *et al.* survey the most widely used sense, vision [7]. Vision-based analysis of group-level activities have been studied by several researchers (e.g. Chang [8], and Lan [9]). However problems with occlusion and changes in lighting can inhibit this modality, particularly in a wearable context.

b) Inertial Sensing: Body-worn, inertial sensing is increasingly used as an alternative to vision and is primarily used to detect individual activities like walking, running, and various hand gestures (Bulling *et al.* provide a good tutorial on the topic [10]). But beyond the work of Gordon, inertial sensing has not been well studied for group activies [4].

c) Sound interactions: The PhD thesis of Sumit Basu explored individual speaker detection in groups. The work showed that it was possible to detect conversations using mutual information (MI) between different speakers [11]. Others, like Lian and Hsu [12], tackled the challenge of detecting conversation dynamics using probabilistic modelling.

d) Sound collocation: Eagle and Pentland used multiple streams of wearable audio to identify various social interactions [13]. Although infra-red beacons were used to detect collocation, they suggested using relative audio signal energy to detect whether speakers were in earshot of one another.

A microphone-only based approach to detecting collocation has the advantage that it can provide fine-grained location information without the need for external infrastructure. Our earlier work on wearable activity recognition demonstrated sound-based activity localisation by using the sound intensity differences between pairs of microphones worn at different arm positions to detect whenever a noise was made close to (or by) the hand [14].

e) Combining sound and acceleration: In that same work we also introduced the idea of fusing sound and acceleration

[1] An initial version of this dataset was described in an earlier short paper by the authors, however only a preliminary analysis was carried out [6].

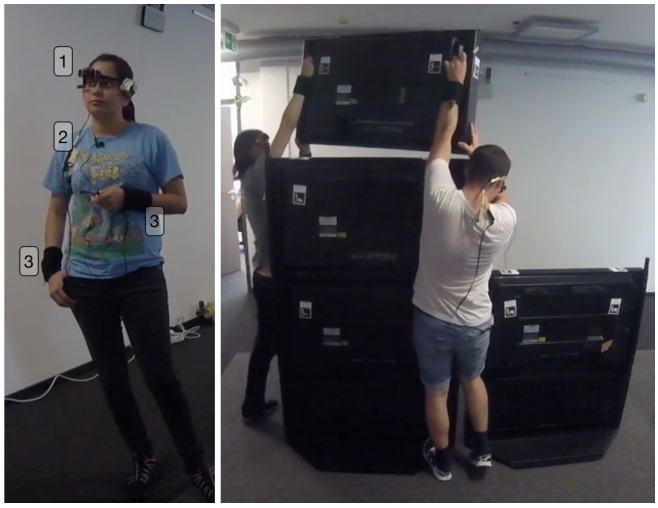

Fig. 1. (Left) sensors include (1) head-worn IMU, eye-tracker, camera, (2) chest-worn microphone, and (3) wrist-worn IMUs. (Right) two people lift a screen into position on the tv wall

information to improve the classification accuracy of recognising tool-use activities (such as sawing, hammering,etc.). This idea – of fusing information from sound and movement – is brought forward to the current work, but this time sound is used to collocate participants, and correlations of acceleration features are used to detect collaboration.

II. Data Collection

The overall task was for 4 people (3 male, 1 female) to collaborate in assembling, and then dismantling, a large, 2.5 m high, 6 screen, video wall.

A short description and guide to the task was given to the subjects, beyond this the group had to organise 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 (as shown in Figure 1), 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 15 minutes of set-up and synchronisation, it took 45 minutes for the subjects to complete the task.

A. Sensor Setup

Each participant wore inertial measurement units (IMUs) on their wrists and head, as shown in Figure 1. The IMU devices record 3-axis acceleration, gyro, and magnetic field – however only acceleration, recorded at approximately 40 Hz, is used in this work. Audio was recorded using a chest-worn microphone connected to an iPhone5 (running the Voice Memos app) in each person's pocket. To aid with synchronisation, a series of clapping and jumping gestures were performed by participants both at the beginning and half-way through the task.

Eye-movement and egocentric video was also captured for each participant. Analysis of this data is outside the scope of the current work but will be explored at length in a future study.

Fixed cameras were mounted in each of the two rooms and in the connecting corridor to assist annotation. In this initial study only two broad classes were considered for annotation: 'collaboration', if two or more people were lifting, carrying, or putting down a screen; or 'no collaboration' to record everything else. An additional annotation on location was made to record which room the participants are in at any given time. This is used to help assess collocation.

Due to sensor failure some of the data had to be discounted. Notably the right-wrist IMU for P2 failed and could not be used. The audio recording for the same subject also failed, but thankfully only the final 5 minutes were lost. The following work is based on the intact 40 minute subset of remaining data (using chest mic, head IMU and left-wrist IMU).

III. Detecting Collocation using Sound

Environmental sounds recorded from two nearby microphones will show a high correlation both in frequency and, assuming a similar distance between source and microphones, intensity. This section describes a method of detecting whether two sound sources are collocated or not by applying a correlation-based classifier on selected frequency and energy features calculated from the sound sources. The method is then evaluated over sound data from all the pairwise combinations of participants in the dataset.

A. Collocation Dataset

A collocation ground truth is annotated for each possible combination and number of participants, indicating whenever people are near one another. In total everyone is together 44% of the time, with participants spending between 9–16% of time alone. The most frequent pairings are P1 & P4 and P2 & P3 (in separate pairs 30% of the time).

B. Sound Features

Many daily sounds can be regarded as stationary at small time frames (e.g. 10-30 ms is typically used in speech recognition). Two features are calculated over moving short-time frames: the zero-crossing count, ZC (the number of times the locally standardized signal crosses zero); and the short-time energy, E, defined by $log_{10}(\sum_n^w s(n)^2)$, for sample $s(n)$ in short-time frame w. The choice of ZC and E was inspired by their use in Bachu *et al.* as a simple way of characterising speech signals [15].

C. Collocation Algorithm

Pearson's correlation (ρ) gives a measure of the linear dependence between two variables, giving an output of 0 if uncorrelated, 1 if highly correlated, and -1 if negatively

correlated. For pairwise feature frames X and Y, it is defined as $\rho^{snd}_{X,Y} = \frac{cov(X,Y)}{\sigma_X,\sigma_Y}$, where cov is the covariance and σ is the standard deviation of the features.

Correlations are calculated pairwise between participant's data for both ZC and E features. To better capture longer-term temporal information in the signals, and to help avoid short disturbances in the signal, these correlations are calculated over a large moving window of W^{snd} frames.

The correlation streams for each feature, ZC and E, are combined by weighted summing: $\rho^{snd}_{X,Y} = \alpha_{ZC} * \rho^{ZC}_{X,Y} + \alpha_E * \rho^{E}_{X,Y}$, for weights α_{ZC}, α_E (both weights are fixed at 0.5 in this work).

Finally, a binary decision (collocation or not) is made by thresholding the output of $\rho^{snd}_{X,Y}$. To smooth over any noise in the output the final classification is determined by a hysteresis with upper (τ^{snd}_u) and lower(τ^{snd}_l) thresholds.

D. Collocation Evaluation

By exploring different windowing parameters, a short-time feature frame of $w = 300$ ms was found to allow sufficient signal information while reducing the need for high-bandwidth communications between participants. A large correlation window of W^{snd} 20 frames (6 seconds) was chosen to capture the longer-term dynamics, while still being short enough to detect location shifts in a timely way.

Evaluation of the collocation algorithm is first carried out across a sweep of classifier thresholds, with each output compared and evaluated against the ground truth. The ground truth in this instance records whether or not two people are collocated at any one time. Frame-based precision ($\frac{true\ positives}{returned\ positives}$) measures how relevant the returned result is, and recall ($\frac{true\ positives}{actual\ positives}$) measures the fraction of the collocations correctly detected.

For each of the participant combinations (P3&P2, P4&P2, etc.), a precision-recall (PR) curve was plotted. The area under curve (AUC) metric, which gives an overall parameter-independent measure of classifier performance, is also calculated. The results are shown in Figure 2. Overall the sound based collocation classifier performs well with an AUC of 0.907. Fixing the parameters with thresholds $\tau_u snd$ =0.7 (large correlation needed to trigger start of collocation) and τ^{snd}_l ==0.1 (low correlation to trigger end of collocation) results in an overall algorithm performance of precision 86.5%, recall 92.4%, and F1-score (balanced mean of precision and recall) 89.3%.

E. Analysis of Collocation Results

Results in the 90% region are good, but there are several issues with the way in which the data is annotated that might impact this result. For example, when participants are walking along the corridor they are labelled as being collocated, even if they are at opposite ends of the (long) corridor. There are instances in the results where this distance is detected by the algorithm (which is arguably correct), but is treated as an error leading to lower recall. Equally there are a handful of instances where very similar sounds occur in separate rooms around the same time, thus giving a false impression of collocation (lower precision).

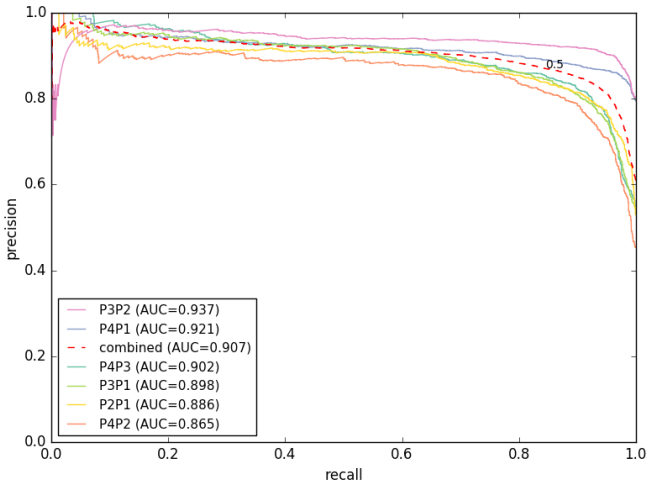

Fig. 2. Precision-recall (PR) curves for audio-based collocation between different couples, and for all combined. Area Under Curve (AUC) is a high .9 for the combined case. Perfect results would be AUC 1 (i.e. at the top right corner).

TABLE I
% OF TIME PARTICIPANTS SPEND IN COLLABORATION

P1P2	P1P3	P1P4	P2P3	P2P4	P3P4	any 3	all 4
12	6	29	18	3	5	1	2

IV. Detecting Physical Collaboration

When two people move a heavy object together, their movements are necessarily correlated in some way. Similarly when walking alongside one another, there will be similarities in their movements. This section describes a method for detecting physical collaborations by classifying on the correlation of acceleration-derived features across multiple participants. The approach is evaluated, (1) across 4 different features and combinations, (2) 3 sensor locations, and (3) in combination with collocation.

A. Collaboration Dataset

The dataset annotations are reduced down to a single class of interest: 'collaborate' representing all instances of large physical activities involving two or more people helping each other, like 'walking together','carry tv','put down tv','lift tv'. All other activities, such as 'using screwdriver', 'finding tools', 'cleaning room' are labelled Null.

The percentage of time participants spend in collaboration is shown in Table I.

B. Collaboration Algorithm

In a similar fashion to the sound based collocation, Pearson's correlation (ρ) is applied across a large rolling window (W_{acc}) between feature data from different participants. Correlation combination is applied as follows: $\rho^{acc}_{X,Y} = \sum^{F}_{i} \alpha_i \rho^{i}_{X,Y}$,

for weight α_i, and F the number of correlations to be combined. (All α weights here are equal.) A decision (collaboration or not) is made by thresholding the output of ρ^{acc}, with final classification using a hysteresis with upper (τ_u^{acc}) and lower(τ_l^{acc}) thresholds.

C. Acceleration Features

3-axis accelerometers were located on the head and wrists of each participant. Due to the failure of P2's right wrist sensor, the analysis that follows is based on the combined correlations from the head and left wrist data.

For each sensor the 3-axis acceleration signals, x,y,z are combined as $S = \sqrt{x^2 + y^2 + z^2}$. This helps improve robustness by making the sensors essentially rotation invariant (as demonstrated by Kunze *et al.* [16]). Four commonly used[2] features – mean, standard deviation (σ), zero crossing count (ZC), and window energy (E) – are then calculated over a range of short-time rolling feature windows, w_{acc}.

D. Evaluation of Features

A short frame of length w_{acc} =300 ms was found to prove the best results in combination with a large correlation window of W_{acc} = 20 (6 seconds). These window settings are also convenient because they correspond to those of the sound analysis, and that they have a similarly low computation overhead.

The PR curves for the 4 features are shown on the top graph of Figure 3. Clearly the ZC (AUC 0.522) outperforms the others, closely followed by short-time energy E (0.447). Correlations based on changing mean perform very poorly (0.26). Even when all features are combined, as shown in the bottom graph of Figure 3, the effects of mean drag down performance. The best PR performance is achieved by combining ZC & E (AUC 0.569). These two features are therefore used for the remainder of the work.

E. Evaluation of Sensor Location

Figure 4 confirms that the strongest performing sensor combination is between head + left wrist. The combined results reflect the intuition that where correlation between single sensors on different people still leaves the possibility of false positives (e.g. chance correlations in hand movement), correlation between two or more sensors on different people is incredibly unlikely (unless the two people are moving together.)

F. Combining Collocation and Physical Correlation

The sound-based collocation output from III-D is combined with the accelerometer correlations, ρ^{acc}. Rather than combining correlation values, ρ^{acc} is multiplied by the best performing 'hard' result of the collocation algorithm (i.e. 1 or 0). The combination is analysed across a threshold sweep in Figure 5.

The final output of the combined collaboration classifier was calculated with correlation hysteresis thresholds, τ_u^{acc} =0.45

[2]See Bulling et al. for an overview of common features [10]

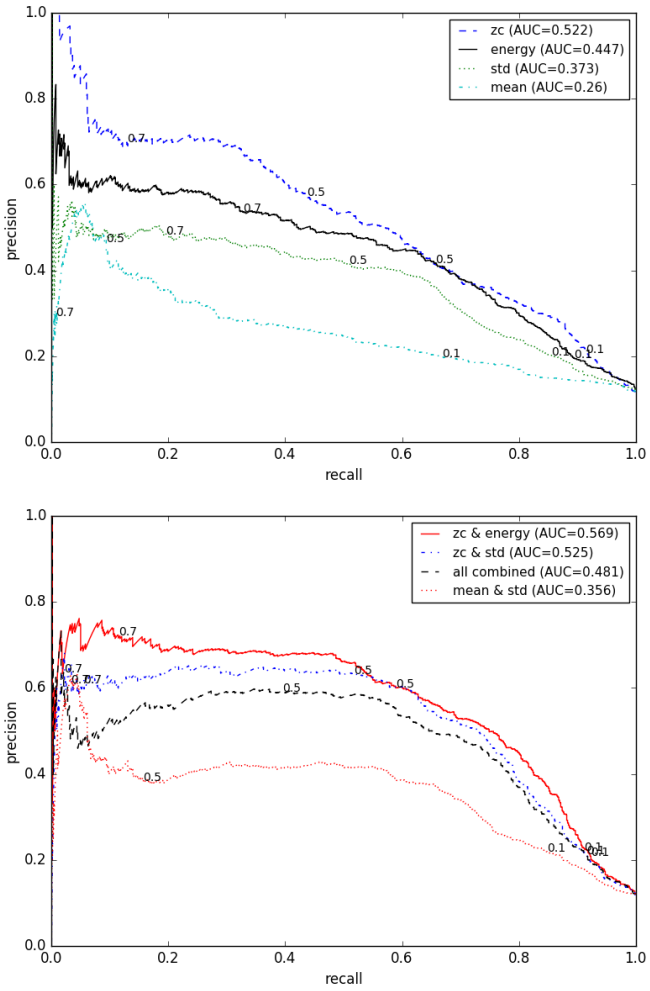

Fig. 3. PR curves comparing collaboration detection using features from head + left wrist accelerometer data; the top graph shows zero-crossing count (zc) and energy outperform standard deviation and mean, while the bottom graph shows a combined zc + energy outperform a combination of all features.

TABLE II
BEST RESULTS (WITH τ_u^{acc} =0.45 AND τ_l^{acc}=0.3)

	Precision	Recall	F1
acceleration only	49.6	74.5	59.6
acceleration with collocation	53.4	70.7	60.1

and τ_l^{acc}=0.3 (picked to capture a range of the best PR values) and is summarised in Table II

Finally, Figure 6 shows the timeline output of this algorithm for each of the participants, P1 to P4, over the experiment duration.

V. Discussion

A. Final results

Overall the results look promising. Figure 6 shows visually how well the majority of collaborations are detected. There are only a handful of false positive events, or insertion errors.

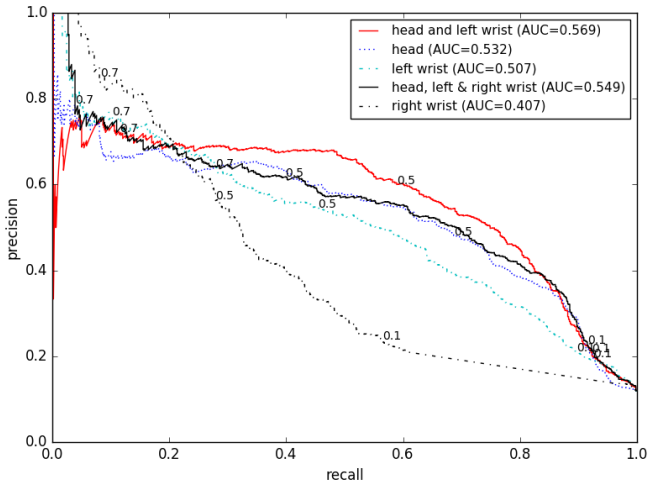

Fig. 4. PR curves comparing collaboration detection using different sensor combinations. Head + left wrist combined outperform individual sensors; the right sensor performs poorly, largely due to a faulty right sensor on one of the participants (P2).

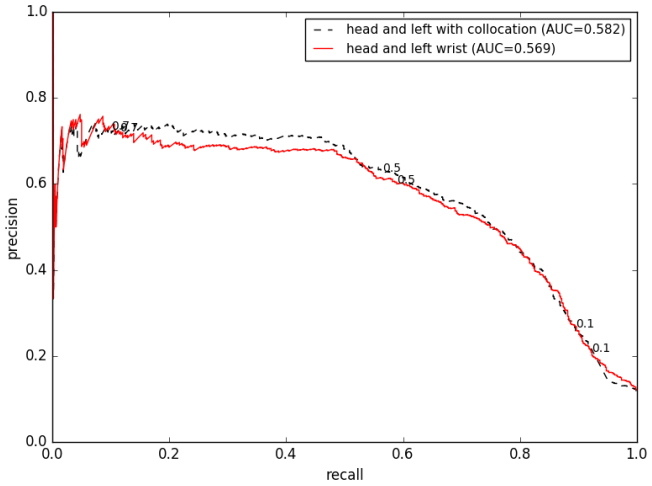

Fig. 5. PR curves showing the slight improvement (from AUC 0.569 to 0.582) in adding collocation to head + left wrist based collaboration.

In fact the largest of these errors, at 00:40 between P1&P2, is actually not an error at all. The video at that time shows P2 moving across the room and then helping P1 from behind the wall. Both of their actions and locations are well correlated, but there was no clear way of annotating the activity, so the ground truth was left blank. However it is clear that this is a physical collaboration.

In addition to this, most of the false positives (accounting for low precision in, for example, Table II) are largely due to inaccuracies in the way 'collaborative activity' is annotated in the ground truth.

B. Missing Data

Simultaneously recording multiple streams of sensor data from 4 people over the course of an hour is non-trivial, particularly when it is not always possible to stop participants in the course of their activities to check the equipment. The loss of data from P2's right wrist, the participant's dominant hand, meant that overall recall rates (relying on left and head data only) are not as high as would be expected – and is likely the reason for the failure to detect some of the P2 collaborations from 00:25 to 00:40. However the results are still good, and it reveals a robustness of the overall approach to real world sensor failure.

C. Future Work

The next stage in this project is to explore more complex collaborations and interactions, for instance activities like one person holding an object while another drives a screw into it, or one person instructing another how to perform a task. This involves a wider range of sensors such as eye-trackers and cameras, and will require more complex recognition methods.

A crucial part of this is in the time-consuming and expensive process of producing accurate annotation. Many components of activity that might be useful in inferring collaboration, such as walking, specific arm movements, or even conversation, need themselves to be recognised. To apply supervised machine learning algorithms to these sub-problems requires annotation at this level too, at least for part of the data.

Annotating activities for multiple, interacting people is a challenging problem. Even with linear, person-specific labelling, as used here, there are many fine-grained issues to be resolved, particularly for overlapping and concurrent tasks. For example, how should annotation be handled if the participants begin on the activity at different times? Or what if one person leaves a collaboration to go help out with another, and then returns again?

VI. Conclusion

The correlation algorithms used to detect collocation and collaboration have the main benefit that they require only a handful of (global) threshold parameters to be set. Because there is no attempt to model activity classes, no expensive training step is needed. They can also be applied cumulatively as a way of fusing different correlation information.

The zero-crossing count (ZC) and short-time energy (E) proved versatile features for characterizing both environmental sound, and physical movements. They have the additional advantage of being quick and simple to compute on a low power device.

Detection of collocation was shown to be very effective (with F1-measure 89.3%) using only inter-person correlations based on ZC and E from chest-worn microphone signals.

Combining data from head-worn and left-wrist accelerometers enhanced detection of physical collaborations beyond what either sensor alone could achieve. By combining sound-based collocation with the physical collaboration, recognition of collaborations achieved an F1-measure of 60.1%.

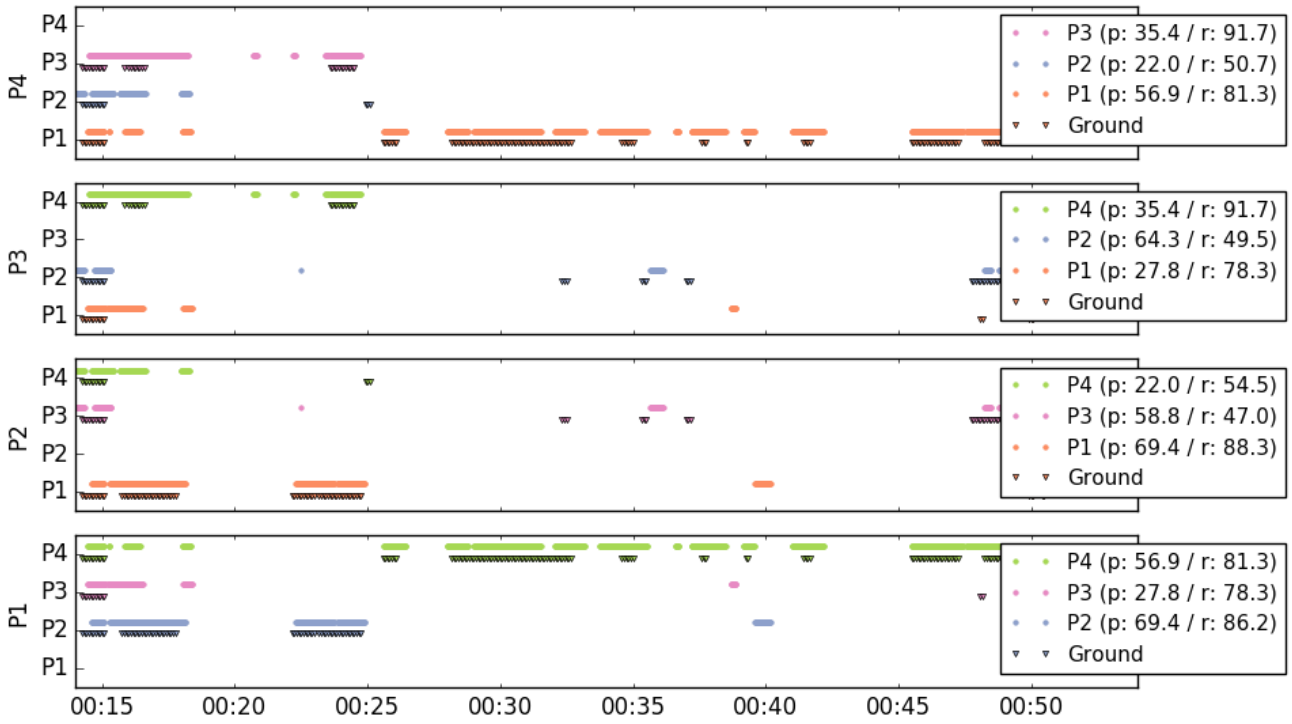

Fig. 6. Physical collaboration detection from the perspective of each participant (P1 to P4) using head + left arm acceleration data combined. Ground indicates when two (or more) participants are collaborating – e.g. at time 00:15 everyone is moving together (walking), while at 00:30 only P4 and P1 are helping one another (lifting and carrying tv screens). Note that around 00:33 P2&P3 help one another, but this is not detected because of the predominant use of P2's right arm in this instance (which had a broken sensor). Visually it can be seen that most collaborative events are captured within an approximate timeframe and there are very few falsely detected or missed events.

ACKNOWLEDGMENT

The authors would like to thank Orkhan Amiraslanov for the use of his custom-built, miniaturised IMU devices. The work is funded by the German Federal Ministry of Education and Research (BMBF).

REFERENCES

[1] J. Ye, S. Dobson, and S. McKeever, "Situation identification techniques in pervasive computing: A review," *Pervasive and mobile computing*, vol. 8, no. 1, pp. 36–66, 2012.

[2] C. Bettini, O. Brdiczka, K. Henricksen, J. Indulska, D. Nicklas, A. Ranganathan, and D. Riboni, "A survey of context modelling and reasoning techniques," *Pervasive and Mobile Computing*, vol. 6, no. 2, pp. 161–180, 2010. [Online]. Available: http://cs.umd.edu/class/spring2014/cmsc818g/files/bettinisurvey.pdf

[3] D. J. C. Geetika Singla and M. Schmitter-Edgecombe., "Recognizing independent and joint activities among multiple residents in smart environments," *Journal of ambient intelligence and humanized computing*, vol. 1, no. 1, pp. 57–63, 2010.

[4] D. Gordon, "Group activity recognition using wearable sensing devices," Ph.D. dissertation, PhD thesis, 2014.

[5] D. Gordon, M. Scholz, and M. Beigl, "Group activity recognition using belief propagation for wearable devices," in *Proceedings of the 2014 ACM International Symposium on Wearable Computers*. ACM, 2014, pp. 3–10. [Online]. Available: http://www.teco.edu/ gordon/publications/DGAR.pdf

[6] J. A. Ward, G. Pirkl, P. Hevesi, and P. Lukowicz, "Towards recognising collaborative activities using multiple on-body sensors," in *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*. ACM, 2016, pp. 221–224.

[7] P. Turaga, R. Chellappa, V. Subrahmanian, and O. Udrea, "Machine recognition of human activities: A survey," *Circuits and Systems for Video Technology, IEEE Transactions on*, vol. 18, no. 11, pp. 1473–1488, Nov. 2008.

[8] M.-C. Chang, N. Krahnstoever, S. Lim, and T. Yu, "Group level activity recognition in crowded environments across multiple cameras," in *Advanced Video and Signal Based Surveillance (AVSS), 2010 Seventh IEEE International Conference on*. IEEE, 2010, pp. 56–63.

[9] T. Lan, Y. Wang, W. Yang, S. N. Robinovitch, and G. Mori, "Discriminative latent models for recognizing contextual group activities," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 34, no. 8, pp. 1549–1562, 2012.

[10] A. Bulling, U. Blanke, and B. Schiele, "A tutorial on human activity recognition using body-worn inertial sensors," *ACM Computing Surveys (CSUR)*, vol. 46, no. 3, p. 33, 2014.

[11] S. Basu, "Conversational scene analysis," Ph.D. dissertation, Massachusetts Institute of Technology, 2002.

[12] C.-c. Lian and J. Y.-j. Hsu, "Probabilistic models for concurrent chatting activity recognition." in *IJCAI*, 2009, pp. 1138–1143.

[13] N. Eagle and A. S. Pentland, "Social network computing," in *International Conference on Ubiquitous Computing*. Springer, 2003, pp. 289–296.

[14] J. Ward, P. Lukowicz, G. Tröster, and T. Starner, "Activity recognition of assembly tasks using body-worn microphones and accelerometers," *IEEE Trans. Pattern Analysis and Machine Intelligence*, vol. 28, no. 10, pp. 1553–1567, October 2006b.

[15] R. Bachu, S. Kopparthi, B. Adapa, and B. Barkana, "Separation of voiced and unvoiced using zero crossing rate and energy of the speech signal," in *American Society for Engineering Education (ASEE) Zone Conference Proceedings*, 2008, pp. 1–7.

[16] K. Kunze, P. Lukowicz, H. Junker, and G. Troester, "Where am i: Recognizing on-body positions of wearable sensors," in *LoCA*, vol. 1, May 2005.