Motion Coupling of Earable Devices in Camera View

Christopher Clarke
Lancaster University
Lancaster, United Kingdom
chris.clarke@lancaster.ac.uk

Peter Ehrich
Ludwig-Maximilians Universität
Munich, Germany
p.ehrich@campus.lmu.de

Hans Gellersen
Aarhus University
Aarhus, Denmark
hwg@cs.au.dk

ABSTRACT

Earables, earphones augmented with inertial sensors and real-time data accessibility, provide the opportunity for private audio channels in public settings. One of the main challenges of achieving this goal is to correctly associate which device belongs to which user without prior information. In this paper, we explore how motion of an earable, as measured by the on-board accelerometer, can be correlated against detected faces from a webcam to accurately match which user is wearing the device. We conduct a data collection and explore which type of user movement can be accurately detected using this approach, and investigate how varying the speed of the movement affects detection rates. Our results show that the approach achieves greater detection results for faster movements, and that it can differentiate the same movement across different participants with a detection rate of 86%, increasing to 92% when differentiating a movement against others.

CCS CONCEPTS

• **Human-centered computing** → **Ubiquitous and mobile devices**.

KEYWORDS

Spontaneous device association; earable; motion coupling.

ACM Reference Format:
Christopher Clarke, Peter Ehrich, and Hans Gellersen. 2020. Motion Coupling of Earable Devices in Camera View. In *19th International Conference on Mobile and Ubiquitous Multimedia (MUM 2020), November 22–25, 2020, Essen, Germany*. ACM, New York, NY, USA, 5 pages. https://doi.org/10.1145/3428361.3428470

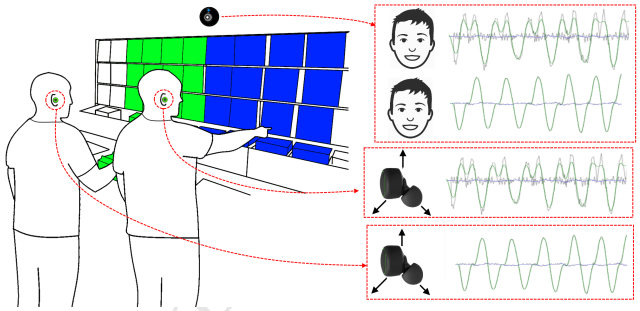

Figure 1: Example application of two users browsing different genres in a music shop, where personalised audio is sent to each user. Spontaneous device association can be achieved by correlating the movement of the faces detected by a camera with the accelerometer data from the earables.

1 INTRODUCTION

By augmenting earphones with established sensing modalities and real-time data accessibility, earables can be leveraged as a socially acceptable sensing platform for ubiquitous applications. Wireless earables can be paired with many devices in a hands-free manner without being tethered to a single device, opening up the opportunity to provide private audio channels for the user based on context or location. For example, consider a music shop where previews of an album can be sent to a user's earphones depending on where in the store they are browsing (Figure 1).

One of the main challenges of realising this vision is the need to spontaneously associate multiple earable devices to multiple users and subsequently track them, without placing the burden on the user to establish an explicit pairing. This is compounded as there may be more users than there are earable devices, and detecting which users are wearing earable devices is non-trivial due to their discreet nature and small form-factor. This rules out conventional camera-based detection techniques due to occlusion, and radio frequency based localisation approaches do not have the required resolution for cases when people are in close proximity without additional infrastructure.

The concept of matching wearable accelerometer data with image features from a camera scene for spontaneous device association has been previously studied as a viable solution (e.g. [4, 8, 12, 16, 19]). In this paper, we investigate how this concept can be applied to earable devices in camera view, and systematically explore the detection rates for different system parameters, and with different types of movement. To investigate this, we implement a sensor fusion approach which looks for correlations between an earable's accelerometer data and faces detected from a web camera (Figure 1). Previous work has investigated how data from an inertial measurement unit can be used to successfully track users with depth sensors in the context of mobile phones [12] and generic devices [19]. However this is complicated for earable devices due to a lack of magnetometer as a result of the magnets in the speakers. The magnetometer is crucial for absolute real-time device orientation, and instead we must rely only on the accelerometer data. We build upon prior work that has demonstrated how only an accelerometer is required for correlation against movement detected from

Unpublished working draft. Not for distribution.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. Copyrights for components of this work owned by others than the author(s) must be honored. Abstracting with credit is permitted. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee. Request permissions from permissions@acm.org.
MUM 2020, November 22–25, 2020, Essen, Germany
© 2020 Copyright held by the owner/author(s). Publication rights licensed to ACM.
ACM ISBN 978-1-4503-8870-2/20/11…$15.00
https://doi.org/10.1145/3428361.3428470

a camera using normalised cross correlation (NCC) and principal component analysis (PCA) [8, 11, 16].

We contribute a data collection in a controlled environment which provides deeper insights into how earables can be correlated with detected faces based on their relative motion. In particular, we perform a systematic analysis to optimise system parameters and to understand how different factors affect the detection performance. By collecting data on both translational and rotational movements in the three principal axes, as well as compound movements, we show how detection rates of 92% can be achieved when comparing different movement, and 86% when comparing the same movement across participants. Our results reveal how movement in the optical axis is harder to detect, and how faster movements can be more accurately detected.

2 RELATED WORK

Our work builds upon spontaneous device association using sensor fusion of cross-modal sensors. In earlier work, Holmquist et al. introduced the concept of *context proximity* for implicit or explicit connection between devices [3]. The concept is to detect the association between devices when they experience similar conditions or events, and was operationalised by detecting matching accelerometer readings when two devices are explicitly shaken together [3, 9]. Similarly, Hinckley introduced the concept of "synchronous gestures", whereby patterns of similar activities take on specific meaning when they occur together in time [2]. For example, when two tablets are bumped together they record equal but opposite accelerometer peaks that occur at the same time. These earlier works show how explicit user interactions can be detected based on temporal matching of similar signals, however the concepts can be extended to implicit interactions across sensing modalities, such as detecting which phone [14, 15] or body-part [18] touches an interactive display.

The same concept of matching motions has been used for indoor localisation of individuals for ubiquitous applications. Kawai and colleagues first demonstrated how motion sensors combined with camera systems can track individuals with high precision by integrating accelerometer data, or counting the number of steps a user takes, and combining them with marker-based computer vision tracking [4]. However, this approach assumes the coordinate system of the accelerometer is known a priori, and is shared with the camera. To overcome this Shigeta et al. introduced normalised cross-correlation (NCC) based on the norms of the acceleration vectors, which are independent of the coordinate system [16]. This was extended to detect feature points in the camera image as opposed to pre-defined objects of interest by restructuring the calculation of NCC into recursively computable forms [8], and by replacing the calculation of the norms with principal component analysis [11]. More recently, Cabrera-Quiros and Hung demonstrate how multiple devices can be detected in a scene when multiple participants (N=69) wore smart badges with embedded accelerometers which are matched against bounding boxes of users [1].

We utilise Plotz et al.'s approach [11] in this paper due to the lack of magnetometer data, and provide a quantitative analysis and deeper understanding of the performance when correlating accelerometer data from earable devices with detected faces. We chose the latter over generic feature tracking or pixel-based approaches to reduce the search space because PCA is more computationally expensive than calculating the norms, despite providing more robust detections. Processing the faces also has the advantage of reducing the chance of false positive detections from spurious movement.

Previous research has also demonstrated how depth sensors, including the Microsoft Kinect, can be utilised in the matching process. Stein and Mckenna extend Maki et al.'s approach ([8]) by correlating accelerometer data with the Kinect depth sensor, and quantitatively evaluate the performance when KLT image features are tracked using both sparse and dense optical flows [17]. Rofouei et al. developed ShakeID which matches accelerometer data from a phone with a user's hands detected from the Kinect for ultimately determining which users are touching a display [12]. Wilson and Benko build upon this with a more generic approach which combines the accelerometer with gyroscope and magnetometer for absolute orientation tracking, and correlate the movement to pixel-level features using dense optical flow [19]. We contrast these approaches by using a basic RGB camera instead of a depth sensor in the correlation process due to their abundance and low-cost.

3 SYSTEM DESIGN

In order to assess the feasibility of spontaneously associating an earable device with faces from a camera, we developed a system inspired by prior work. The goal of the system was to simultaneously record data from an earable device and images from a camera feed, which could be later post-processed for analysis. The main stages of the processing pipeline are as follows:

(1) Extract earable accelerations: the earable devices transmit the accelerometer data which is filtered, and reduced to a one-dimensional vector using PCA;
(2) Extract video accelerations: faces are detected from the camera's data stream, and the centroid of each face is used to calculate the acceleration, which is filtered and transformed into a one-dimensional vector using PCA;
(3) Similarity detection: NCC is used to find the time shift at which the signals are maximally correlated, and a match is determined based on the variability of the time shifts.

In the remainder of this section, we describe the details of each stage and the practical considerations.

3.1 Extract Earable Accelerations

For the earable devices we used the eSense developed by Nokia Bell Lab [5, 10], which feature a True Wireless Stereo (TWS) earbud augmented with a 6-axis inertial motion unit, a microphone, and dual mode Bluetooth (Bluetooth Classic and Bluetooth Low Energy). The eSense transmits the 3-axis accelerometer data at a sampling rate of 50 Hz to a phone over Bluetooth. The phone adds a timestamp to the packet because the eSense has no internal storage or real-time clock, before forwarding onto a local server for storage. The acceleration data is then filtered using a 2nd order zero-lag Savitzky-Golay filter with a window size of 17. Following Plotz et al.'s approach [11], each axis of the accelerometer is standardised prior to performing dimensionality reduction using PCA. We discard all but the principal component, resulting in a one-dimensional array which represents the earable's acceleration.

Unpublished working draft. Not for distribution.

3.2 Extract Video Accelerations

As the headphones are located in the ear we opted to use face tracking because it provides reliable tracking and is likely to correspond to the headphone movement, assuming very little in-ear movement of the earable device. For face detection we used dlib's implementation of the 68-point landmark detection model [6] which was trained on the iBUG 300-W face landmark dataset [13]. To calculate the acceleration of each user we extract the centroid for each face by calculating the mean position of all 68 landmarks. At each stage (position, velocity, acceleration) we filter the signal with a 2nd order zero-lag Savitzky-Golay filter with window sizes of 5, 11, and 17 respectively. We then perform PCA on the standardised array to reduce the accelerometer vector to a one-dimensional array.

It is important to maintain persistent tracking of faces in the scene to create trajectories for comparison against the earables' accelerometer data. This is complicated by tracking failures and occlusions. One way to achieve this would be to use facial recognition to assign faces between images, however this raises privacy concerns, and instead we opt for a simpler approach. For every frame, we create a matrix of existing trajectories and new face centroids and compute the sum of square distance between them as a cost function. This is then used to assign and correct face detections to motion trajectories using the Hungarian Algorithm. Two factors are used to decide whether a tracking point is still valid or not – the cost function representing the distance, and a counter for frames where the tracking point was unassigned. If the cost function for a particular tracking point is too high, or when the skipped frames counter becomes too large, we create a new trajectory.

3.3 Similarity Detection

The timestamps of data from the earable and images from the camera are not synchronised. Using Shigeta et al.'s approach [16], we use normalised cross correlation (NCC) to determine the similarity between the signals by detecting the time shift between principal components of the accelerations. For every frame we compute the NCC for each combination of earable device and face trajectory after resampling the earable accelerometer vector using the Fourier method to match the camera's sampling rate. The index of the maximum value of the NCC reveals the time shift required to maximise the similarity between the signals for each frame, and in theory the value of the NCC encodes the amount of similarity. However, in practice there may be pertubations in the maximal value due to temporarily coincidental motions or sensor disturbances.

Instead of the instantaneous NCC value, Shigeta et al. proposed to look at the consistency of the maximal value (Figure 2). For similar motions this delay should be consistent across frames because it is physically grounded and manifests itself because of various delays of Bluetooth and network transfer [8, 16]. Shigeta and colleagues originally proposed to use sequential Bayesian analysis to remove outliers, and to use the standard deviation measured over a window to determine the consistency of the detected maxima [16]. In addition to the standard deviation, we implemented the mean absolute deviation and the median absolute deviation which are more robust measures to outliers than the standard deviation [7]. We also investigate how effective the Bayesian filtering is.

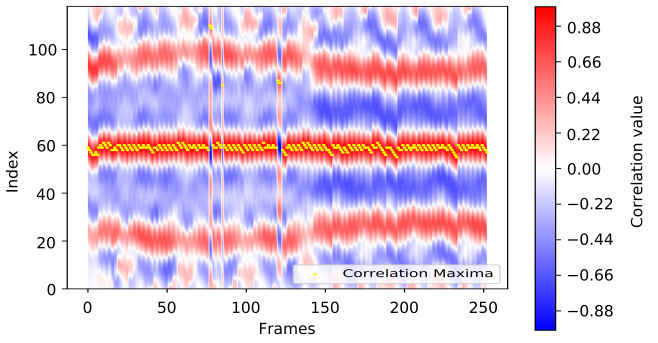

Figure 2: NCC result for a recording using a 2 second window size where the earable data matches the video. One can see the low variability and consistent time shift for the maximum correlation (yellow dots). At some time steps there is a very low correlation at the ground truth point which can be caused by coincidental motions or random noise.

4 DATA COLLECTION

We collected data in a controlled environment to gain deeper insights into how earables can be correlated with detected faces based on their relative motion. We perform a systematic analysis to understand how different factors affect the detection performance. As we perform the correlation using the accelerometer for detecting earable movement and a 2D camera for face detection, we hypothesise that movement in the optical axis or parallel to the ground plane will be harder to detect. To investigate this, we collect data on both translational and rotational movements in the three principal axes. We also investigate how the speed of the movements affect the detection rate, and expect faster movements to be more easily detected due to more pronounced accelerometer readings. Whilst investigating how the type of movement and speed affect the detection rates, we consider how multiple devices can be detected, and also what the optimal system parameters are to maximise detection rates, in particular the window size over which the NCC is calculated (1, 2, and 3 seconds) and the variability measure of the peaks (standard deviation, mean absolute difference, and median absolute difference).

4.1 Participants and Apparatus

Ten participants were recruited to take part in the study (M=25.4, SD=3.7), three of whom identified as female and the remainder as male. A 60" Samsung Smart TV was used to display the study instructions, and an off the shelf Logitech C920 Pro USB Webcam was placed atop, with images captured at 30 frames per second with a resolution of 640×480. A Pixel 3a XL3 was used as a bridge between the eSense and the server for data recording. All data was recorded locally for post-processing because it allowed for more permutations than is practically possible running the system online.

4.2 Procedure

Participants were located in the centre of the display at a distance of 1.5 m. They then performed seven different movements for 10 seconds each on cue of the researcher. These included an idle movement, in which users were asked to stand still, which we used

Unpublished working draft. Not for distribution.

to see if the proposed approach could detect the subtle nuances and micro-movements between participants. Participants then performed three translational movements (left-right, up-down, and forward-backward), and three rotational movements (roll, pitch, and yaw) at whatever speed they felt comfortable for the duration of the 10 seconds. This was then repeated with participants asked to perform the movements slowly and more quickly compared with their original movements (excluding the idle movement). Participants were free to choose how they moved their head as long as it conformed to the study task (e.g. neck movement or full body movement both result in head movement). We then collected data on compound movements which occur simultaneously in multiple axes, including clockwise and counterclockwise circular movement, and random movement from the participant. All movements were performed at three different speeds in the same manner as the first set, however the random movement was recorded for 20 seconds.

We analyse the data as a classification problem and extract the detection rate of the proposed approach. The detection rate is defined as the percentage for which the variability of the peaks detected in the NCC process is the lowest possible for the correct movement combination compared with all others. We analyse the data from two perspectives – the first is across movements in order to gain deeper insight into how different types of movements in the principal axes affect the detection rate. We compare each principal axis movement of a participant against all others performed by that participant (1 idle + 3 speeds * 6 movements = 19), which is analogous to multiple people performing different movements simultaneously. We then look across participants to see how well the approach can differentiate different users performing the same type of movement. This is analogous to multiple people (N=10) performing the same type of movement in the camera scene.

5 RESULTS AND DISCUSSION

For the optimal system parameters, we found that the mean absolute difference outperformed all others when comparing across movements (Std dev.: 88.24%, Mean AD: 91.06%, Median AD: 84.64%), and across participants (Std dev.: 81.95%, Mean AD: 84.09%, Median AD: 81.15%). For the mean absolute deviation we found that 3s was the optimal window size for the NCC calculation both across movements (1s: 90.09%, 2s: 91.29%, 3s: 91.81%) and across participants (1s: 80.67%, 2s: 85.16%, 3s: 86.43%), however we note the small increase in detection rates between 2 and 3 seconds. We found the sequential Bayesian estimation did not improve performance, and for the remainder of the analyses we discuss the results using the mean absolute deviation with the 3 second NCC window size.

Cross Movement Analysis: We found that the detection rates were higher when participants performed the movements faster (94.26%), compared with both the normal (91.30%) and slow (89.81%) speeds. Figure 3(a) shows the detection rates for each individual movement, averaged across all participants. The up-down movement achieved the highest success rate of 98.33%, which may be due to the fact the movement is parallel to the gravitational vector. The lowest detection rate is the roll movement (86.85%), which suggests that even in the worst case the approach is capable of accurately classifying different movements.

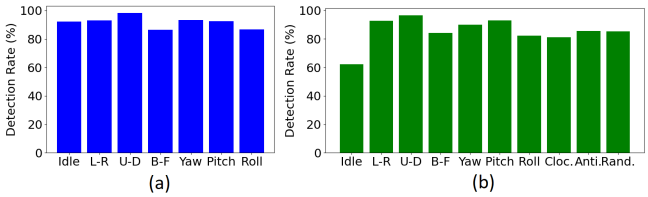

Figure 3: Results when (a) translational movements of left-right (L-R), up-down (U-D), back-forward (B-F), and rotational movements in the principal axes are compared with other movements, and (b) movements compared across participants including compound movements of clockwise and anti-clockwise circular and random movements.

Cross Participant Analysis: Similarly, when we analyse detection rates across participants we notice the faster movements (93.83%) outperform both the normal (86.42%) and slow (81.73%). This is a more challenging classification which is reflected in the lower detection rates compared with comparing across different movements. Figure 3(b) reflects again that the up-down movement was the easiest to correctly detect across participants (96.67%). The idle detection rate is significantly lower (62.22%) which suggests that the approach and sensing devices can not distinguish between the nuances of small micro-movement when users are idle, however it still performs better than random chance.

These results demonstrate that earables can be spontaneously associated with users by only transmitting their accelerometer data, and can be used to provide users with personalised audio without requiring prior information, user identification, or explicit user action. The insights into how movements affect detection rate can be utilised by favouring detections in specific axes, and also suggest that more explicit device association gestures may be more accurate, such as nodding one's head quickly (which achieved 100% detection rates across both movements and participants).

These insights can be generalised to other application contexts in which only an accelerometer is available, and where the camera could take other form factors such as the front-facing camera of a smartphone. The study used a limited set of movements directed towards the camera as a starting point to evaluate such as system, but it shows the potential to work well even in more realistic scenarios such as the random movement task, and invites exploration of the approach in realistic application contexts using the optimal system parameters discussed.

6 CONCLUSION

Earables present a newly emerging ubiquitous platform that can be leveraged for unique applications when users can be correctly assigned to their device. We have demonstrated how earable devices can be spontaneously associated with faces detected by camera movement with detection rates of 86% across different types of movement, and of 92% across participants performing the same type of movements. A deeper understanding of what movements are most accurately detected provides insight into which movements to focus on during the association process, and for design of explicit gestures for device association.

Unpublished working draft. Not for distribution.

REFERENCES

[1] Laura Cabrera-Quiros and Hayley Hung. 2016. Who is where? Matching people in video to wearable acceleration during crowded mingling events. In *Proceedings of the 24th ACM international conference on Multimedia*. 267–271.

[2] Ken Hinckley. 2003. Synchronous Gestures for Multiple Persons and Computers. In *Proceedings of the 16th Annual ACM Symposium on User Interface Software and Technology* (Vancouver, Canada) *(UIST '03)*. ACM, New York, NY, USA, 149–158. https://doi.org/10.1145/964696.964713

[3] Lars Erik Holmquist, Friedemann Mattern, Bernt Schiele, Petteri Alahuhta, Michael Beigl, and Hans-Werner Gellersen. 2001. Smart-Its Friends: A Technique for Users to Easily Establish Connections Between Smart Artefacts. In *Proceedings of the 3rd International Conference on Ubiquitous Computing* (Atlanta, Georgia, USA) *(UbiComp '01)*. Springer-Verlag, London, UK, UK, 116–122. http://dl.acm.org/citation.cfm?id=647987.741340

[4] Jun Kawai, Kimio Shintani, Hirohide Haga, and Shigeo Kaneda. 2005. Identification and positioning based on motion sensors and a video camera. In *IASTED Int. Conf. Web-Based Education*.

[5] Fahim Kawsar, Chulhong Min, Akhil Mathur, and Allesandro Montanari. 2018. Earables for personal-scale behavior analytics. *IEEE Pervasive Computing* 17, 3 (2018), 83–89.

[6] Vahid Kazemi and Josephine Sullivan. 2014. One millisecond face alignment with an ensemble of regression trees. In *Proceedings of the IEEE conference on computer vision and pattern recognition*. 1867–1874.

[7] Christophe Leys, Christophe Ley, Olivier Klein, Philippe Bernard, and Laurent Licata. 2013. Detecting outliers: Do not use standard deviation around the mean, use absolute deviation around the median. *Journal of Experimental Social Psychology* 49, 4 (2013), 764–766.

[8] Yuichi Maki, Shingo Kagami, and Koichi Hashimoto. 2010. Accelerometer detection in a camera view based on feature point tracking. In *2010 IEEE/SICE International Symposium on System Integration*. IEEE, 448–453.

[9] Rene Mayrhofer and Hans Gellersen. 2009. Shake well before use: Intuitive and secure pairing of mobile devices. *Mobile Computing, IEEE Transactions on* 8, 6 (2009), 792–806. https://doi.org/10.1109/TMC.2009.51

[10] Chulhong Min, Akhil Mathur, and Fahim Kawsar. 2018. Exploring audio and kinetic sensing on earable devices. In *Proceedings of the 4th ACM Workshop on Wearable Systems and Applications*. 5–10.

[11] Thomas Plotz, Chen Chen, Nils Y Hammerla, and Gregory D Abowd. 2012. Automatic synchronization of wearable sensors and video-cameras for ground truth annotation–a practical approach. In *2012 16th international symposium on wearable computers*. IEEE, 100–103.

[12] Mahsan Rofouei, Andrew Wilson, AJ Brush, and Stewart Tansley. 2012. Your phone or mine? Fusing body, touch and device sensing for multi-user device-display interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*. 1915–1918.

[13] Christos Sagonas, Epameinondas Antonakos, Georgios Tzimiropoulos, Stefanos Zafeiriou, and Maja Pantic. 2016. 300 faces in-the-wild challenge: Database and results. *Image and vision computing* 47 (2016), 3–18.

[14] Dominik Schmidt, Fadi Chehimi, Enrico Rukzio, and Hans Gellersen. 2010. PhoneTouch: A Technique for Direct Phone Interaction on Surfaces. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology* (New York, New York, USA) *(UIST '10)*. ACM, New York, NY, USA, 13–16. https://doi.org/10.1145/1866029.1866034

[15] Dominik Schmidt, Julian Seifert, Enrico Rukzio, and Hans Gellersen. 2012. A Cross-device Interaction Style for Mobiles and Surfaces. In *Proceedings of the Designing Interactive Systems Conference* (Newcastle Upon Tyne, United Kingdom) *(DIS '12)*. ACM, New York, NY, USA, 318–327. https://doi.org/10.1145/2317956.2318005

[16] Osamu Shigeta, Shingo Kagami, and Koichi Hashimoto. 2008. Identifying a moving object with an accelerometer in a camera view. In *2008 IEEE/RSJ International Conference on Intelligent Robots and Systems*. IEEE, 3872–3877.

[17] Sebastian Stein and Stephen J McKenna. 2012. Accelerometer localization in the view of a stationary camera. In *2012 Ninth Conference on Computer and Robot Vision*. IEEE, 109–116.

[18] Andrew M Webb, Michel Pahud, Ken Hinckley, and Bill Buxton. 2016. In *Proceedings of the 29th Annual Symposium on User Interface Software and Technology*. ACM, 287–300.

[19] Andrew D Wilson and Hrvoje Benko. 2014. Crossmotion: fusing device and image motion for user identification, tracking and device association. In *Proceedings of the 16th International Conference on Multimodal Interaction*. ACM, 216–223.

Unpublished working draft. Not for distribution.