Attention, please! Comparing Features for Measuring Audience Attention Towards Pervasive Displays

Florian Alt[a], Andreas Bulling[b], Lukas Mecke[a], Daniel Buschek[a]

[a]LMU Munich
Munich, Germany
firstname.lastname@ifi.lmu.de

[b]Max Planck Institute for Informatics
Saarbrücken, Germany
bulling@mpi-inf.mpg.de

ABSTRACT

Measuring audience attention towards pervasive displays is important but accurate measurement in real time remains a significant sensing challenge. Consequently, researchers and practitioners typically use other features, such as face presence, as a proxy. We provide a principled *comparison* of the performance of six features and their combinations for measuring attention: face presence, movement trajectory, walking speed, shoulder orientation, head pose, and gaze direction. We implemented a prototype that is capable of capturing this rich set of features from video and depth camera data. Using a controlled lab experiment (N=18) we show that as a single feature, face presence is indeed among the most accurate. We further show that accuracy can be increased through a combination of features (+10.3%), knowledge about the audience (+63.8%), as well as user identities (+69.0%). Our findings are valuable for display providers who want to collect data on display effectiveness or build interactive, responsive apps.

Author Keywords
public displays, interaction, audience funnel, phases, zones

ACM Classification Keywords
H.5.1 Multimedia Information Systems — *Evaluation / Methodology*;

INTRODUCTION

User attention is a fundamental prerequisite for public display interactions since displays need to attract the attention of passersby to be used [4, 17]. Knowledge about attention is particularly valuable for interactive applications, since it can be used to guide users through the interaction process: An application can first attract the attention of passersby through a particular stimulus (e.g., motion [12], appearance of new objects [13], moving and looming stimuli [11], or changes in luminance contrast [10]). As soon as attention of passersby is detected, further actions can be taken to encourage interaction [3], for example, by conveying interactivity [18], guiding users to the optimal interaction area [2], or explaining how the offered interaction technique works [1, 22].

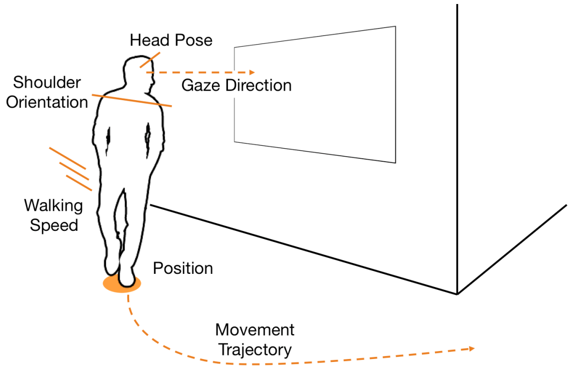

Figure 1. Attention towards displays: We compare the accuracy of different features (face presence, movement trajectory, walking speed, shoulder orientation, head pose, gaze direction) obtained from sensor data.

At the same time, attention is an important metric to assess display effectiveness through conversion rates [17], investigate user behavior in the long term [15], and ultimately to react to user behavior in real-time to support the interaction process. However, reliable attention measurements are challenging and received only little attention in research, and if so, mainly in lab settings requiring accurate user localisation [8, 14, 23]. Prior work claims that if eye contact with the display lasts for more than 800 ms attention can be assumed [6, 16]. Eye contact can be detected using eye tracking [27] but current systems require the user to remain static within a confined area. Hence, deployments often use other solutions, inferring attention from the presence of faces [21] or the mere presence of the user [7]. To the best of our knowledge, there is no research quantifying the accuracy of these features.

To close this gap, we show how attention can be quantified in real-time from features obtained through a camera and a depth sensor. Using state-of-the-art techniques, we provide a comparison of previously proposed features. In particular we compare face presence, movement trajectory, walking speed, shoulder orientation, head pose, and gaze direction. To capture and process the data, our prototype uses multiple Kinects to cover an area of $25\,\mathrm{m}^2$ in front of the display. We conducted a controlled lab experiment (N=18) and show that in cases where users are unknown (e.g., a public square), face presence is indeed among the most accurate features. We also show that accuracy can be further improved by combining face presence with other features (+10.3%), if the audience is known (e.g., in a supermarket) (+63.8%), as well as if people can be identified (e.g., a personal work space) (+69.0%).

DIS'16, June 04 - 08, 2016, Brisbane, QLD, Australia
Copyright is held by the owner/author(s). Publication rights licensed to ACM.
ACM 978-1-4503-4031-1/16/06...$15.0
DOI: http://dx.doi.org/10.1145/2901790.2901897

These findings are valuable for the designers of pervasive display applications who want to either quantify the success (i.e., how many people are paying attention towards a display) or want to build applications that adapt to the user (for example, guidance through the interaction process).

Contribution Statement. Our contributions are two-fold. First, we show how audience attention towards pervasive displays can be measured using state-of-the-art techniques. Second, we present a controlled lab experiment, quantifying and comparing the accuracy of different features. Note, that we do not propose novel features to measure attention. Our work is complemented by discussing implications for the designers of interactive display applications.

MEASURING AUDIENCE ATTENTION

In a first step we identified reliable and easy-to-deploy methods to measure overt attention towards pervasive displays. In particular, we consider features describing user behavior, obtainable from video (v) and depth (d) data.

Face Presence (v) The presence of a face in the display vicinity may already be a good, though not perfect, indicator for attention towards the display. Note that face presence does not necessarily mean the head is directed towards the display. This feature also serves as a baseline.

Walking Speed (d) We assume that if people look at the display they may reduce their walking speed, which could serve as an indicator for attention towards the display.

Position & Movement Trajectory (d) From depth sensor data, the position and walking direction of a passerby can be used to calculate a movement trajectory. This may serve as an indicator whether or not a user is approaching a display and hence may direct his attention towards it [9, 24].

Shoulder Orientation (d) A user's shoulder orientation relative to the display serves as a coarse indicator for attention being directed towards the display. This is motivated by the observation that as a display attracts users, they do not only turn their head but at some point also their upper body, entering an optimal viewing position [5].

Head Pose (v) Prior work showed head pose to be an attention indicator [19], though eye contact cannot be assumed.

Gaze Direction (v) We estimate a user's coarse gaze direction. Two different approaches are considered. Feature-based gaze estimation methods detect prominent facial features, such as eye corners and pupil centers, and use geometric mapping functions to determine the gaze location on the display. In contrast, appearance-based methods directly learn a mapping from eye appearance to on-screen gaze location [26]. While appearance-based methods are typically more robust, e.g. to changing lighting conditions, we opted for a feature-based method because these provide higher gaze estimation accuracy.

IMPLEMENTATION

Based on the skeleton data from the Kinect we track the face position and calculate the user's orientation (*body posture*) based on the shoulder joints. By buffering past positions we also extract the user's movement and *walking speed* over time, resulting in the *movement trajectory*. We refine the determined face position with an OpenCV Viola-Jones face detector on the color image if needed. We detect facial features within the face bounding box using the IntraFace library [25] (*face presence*). If successful, this returns coordinates for several facial feature points as well as the *head pose*.

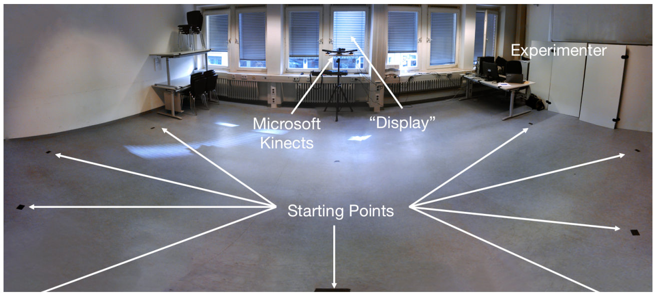

Figure 2. Setup: Subjects started from 7 start points, at different speeds, either looking at the 'Display' or not. They were recorded by 3 Kinects.

In addition, we obtain the center of the pupil for each eye using a gradient-based approach [20]. To determine the eye region we calculate a rectangle from the eye corner points. From the detected eye corners and the pupils for both eyes we estimate the *gaze direction*. We determine the center point between the eye corners and add the head vector normalized to the eye radius to estimate the center of the eye ball. Subtracting the estimated center from the previously determined pupil results in a vector aligned with the user's gaze direction.

EVALUATION

Next, we conducted a lab experiment to determine the overall accuracy as well as the contribution of the different features and feature combinations. We deliberately opted to collect data in a controlled setting to obtain a ground truth, i.e., we needed to determine whether people looked at the display as they passed by. Though the lab setting is a limitation, running the experiment would have been difficult in public, since it would have required either video recording and post-hoc analysis of the data or interviewing each passerby.

Study Design

The study followed a repeated measures design with walking direction, speed, and gaze direction as independent variables.

Walking Direction. We tested seven directions from which users could approach or pass by the display (90°, 120°, 150°, 180°, 210°, 240°, 270° to the display). We assumed the system to work best as users walk directly towards the display, compared to cases where users walk parallel to the display.

Walking Speed. To account for both situations where users are hurrying past the display as well as situations in which users are walking in a casual manner, we instructed participants to either walk at normal or fast pace past the display.

Gaze Direction. In the final condition users were asked to look at the display, whereas in the other condition users were asked to not look at the display with no particular instructions where to look. We use this information as a ground truth.

In total, this resulted in seven walking directions × two gaze directions × two walking speeds = 28 conditions.

Setup

We setup the study in a large seminar room in our lab using three Kinects to cover the entire space in front of the display (Figure 2). The room was large enough so that the starting and end points (black squares on the floor) for users' walks were outside the tracking area – similar to a real-world setting. Note that including non-parallel walking directions also allowed measurements from different distances to the display (cf. Figure 3). We marked the starting points for the walking tasks through labels on the ground. The Kinects were placed on a rectangular wooden panel placed in front of a window that at the same time represented the display.

Procedure

We first asked participants to fill in an introductory questionnaire, assessing gender and age. The experimenter provided a brief introduction to the trial and explained them that their task would be to walk past the display on predefined trajectories, at different speeds, and either looking at the display or somewhere else. Before each walk the experimenter told them the condition. To minimize errors, we grouped the walks by speed, by gaze direction, and by walking direction. Participants then performed the walks for all 28 conditions, followed by a short break. This was repeated four times (112 walks).

RESULTS

Data Preprocessing

For the following analyses, we excluded data points in between trials, for example, when participants were captured while walking from one trial's end point to the next start point.

We smoothed feature values with a rolling window of 30 frames. Thus, we base predictions on observations aggregated from approximately one second. This time period was motivated as a slightly conservative version of the 800 ms of eye-contact assumed to indicate attention in related work [6, 16].

Evaluation Schemes

We use a Random Forest classifier[1] with default hyperparameters and three different evaluation schemes to analyze the data. The schemes reflect different assumptions about the practical context of a system using the evaluated features for attention detection. The three schemes are:

- **Identified User**: Here, we conduct ten-fold stratified cross-validation on the data of a single user at a time. Reported mean values in this scheme are averaged over all users and reported standard deviations thus describe the spread of the user-specific means. This scheme reflects performance of a system that has identified the user before predicting attention (e.g. personal computing system).
- **Known User**: This scheme conducts ten-fold stratified cross-validation on the pooled data of all users. Reported mean values are averaged over all folds and thus standard deviations describe the spread of the fold-specific means. This scheme reflects performance of a system that can assume the user to be known, but without identification (e.g. office space or supermarket).

Feature	Identified User		Known User		Unknown User	
	μ	σ	μ	σ	μ	σ
Movement Trajectory	0.86	0.02	0.71	0.00	0.52	0.01
User Position	0.69	0.04	0.58	0.00	0.51	0.01
Face Presence	0.61	0.07	0.60	0.00	0.60	0.07
Head Pose	0.60	0.04	0.57	0.00	0.53	0.02
Gaze Direction	0.58	0.02	0.56	0.00	0.53	0.02
Shoulder Orientation	0.54	0.02	0.51	0.00	0.51	0.01
Walking Speed	0.52	0.01	0.50	0.00	0.50	0.00

Table 1. Comparison of features, sorted by mean accuracy in the "identified user" evaluation scheme.

- **Unknown User**: In this scheme, we use the data of all but one user for training, then test on the data of the remaining user. This procedure is repeated so that each user is the one to test on exactly once. Reported mean values in this scheme are averaged over all these repetitions and thus standard deviations describe the spread of the repetition-specific means. This scheme reflects performance of a system that does not assume to have seen the user before (e.g. a display in a public space).

Basic Statistics

Our dataset has 243,324 data points from 18 participants (9 male, 9 female; all West-Europeans). Hence, each person contributed 13,518 data points on average. Since about 53% of our data points are samples from the "look at display" condition, we use this number as the baseline accuracy. This would be achieved by a system that simply constantly predicts user attention, and never not-attending. We did not find differences caused by walking speed. Walking speed followed a normal distribution with a mean of 1.25 m/s.

Evaluation of Single Features

Table 1 shows a comparison of the different kinds of features for the three different evaluation schemes.

Identified User Scheme

This scheme achieved the highest accuracy overall. Here, features from the Kinects' skeleton data worked best: The top features are *Movement Trajectory* (86% accuracy) and *User Position* (69%). Features from the camera stream resulted in worse prediction accuracies: Here, *Face Presence* (61%) worked best, followed by *Head Pose* (60%) and *Gaze Direction* (58%). *Shoulder Orientation* (54%) and *Walking Speed* (52%) did not compare well to the baseline accuracy of 53%.

Known User Scheme

In general, accuracy was lower in this scheme than for identified users, indicating individual differences in user behavior. The top feature was *Movement Trajectory* (71% accuracy), but this time followed by *Face Presence* (60%). *User Position* was the fourth best feature here (58%). *Head Pose* reached 57% accuracy, *Gaze Direction* achieved 56%. Again, *Shoulder Orientation* (51%) and *Walking Speed* (50%) did not compare well to the baseline (53%).

Unknown User Scheme

The final scheme led to the worst accuracy overall. In contrast to the other two schemes, here the top feature (*Face Presence*, 60% accuracy) was based on the camera, not the depth data. Besides this feature, only *Head Pose* and *Gaze Direction* (both 53%) performed comparable to the baseline (53%).

[1] `http://scikit-learn.org/stable/modules/generated/sklearn.ensemble.RandomForestClassifier.html`

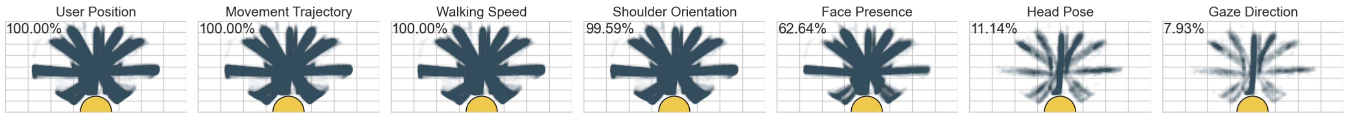

Figure 3. Participants' X-Z-locations at which certain features could be measured. For example, the plot for *Face Presence* shows where our system detected a face. The display is indicated by a half-circle. Percentages show the number of feature detections as a ratio of the whole dataset. The plots show that video-based features are harder to assess robustly than features based on the depth-camera. Not surprisingly, camera-features are mainly assessable when users walk straight towards the display, such that the system can see the face directly from the front for several subsequent frames.

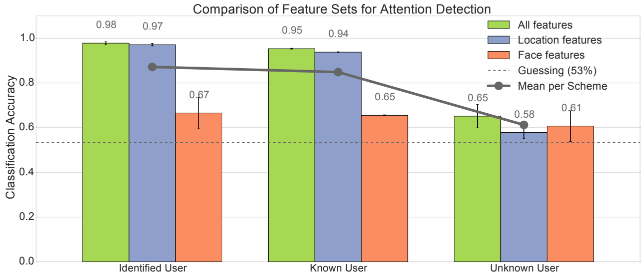

Figure 4. Comparison of feature sets in our three evaluation schemes. These results show a superiority of location features for known users, and better performance of face features for strangers. However, both feature sets can be combined to improve accuracy throughout all schemes.

Feature Locations

We further analyzed at which locations (distances) in front of the sensors certain feature values could be measured. Figure 3 shows the results in a top-down view. While depth-based features were almost always available, faces were only present in about 63% of the frames, either because the system could not recognize the face, or since the user simply had turned away from the system. Features derived from the face data were available in about a tenth of all frames. Looking at the related locations, we found that this worked best when users walked straight towards the system, presumably since this produced multiple subsequent frames with a rather stable face image.

Evaluation of Feature Sets

We further compared three different larger sets of features: 1) all features, 2) features based on location data (i.e. depth camera-based features: *Movement Trajectory*, *User Position*, *Shoulder Orientation*, *Walking Speed*), and 3) features based on face data (i.e. video-based features).

Figure 4 summarizes the results. They match the picture obtained in the single feature evaluation: Location features work best in the *Identified User* scheme and the *Known User* scheme, while face features are superior in the *Unknown User* scheme. Both feature sets outperformed the baseline and their combination resulted in further improvement in all schemes.

DISCUSSION

Preferred Features Depend on Deployment Context

Comparing results between our evaluation schemes, we found that features based on the Kinects' depth cameras greatly outperformed video-based features, if training data from the specific users is available. This indicates that movement trajectories are user-specific, but recognizable across multiple repetitions, as observed in our experiment. This is particularly valuable from a privacy perspective, since location features can usually be collected without the need to record video data. Hence users can stay anonymous at the time of detecting attention. Such a setup is usually desirable in workplaces (e.g., a display in the entrance area of an office or university building) or a supermarket with a known user base (employees, customers) where people should not be identified and no profiles be created (e.g., when somebody arrives at work).

We thus conclude that systems in known environments, such as an office space, should preferably use depth features. If available, these features can be combined with video-based ones to improve accuracy. In contrast, systems in public environments should definitely include camera-based features, which are, however, more privacy-invasive. Note that for known and identified user schemes, places with large user numbers (shopping malls, train stations), scalability might be an issue.

Face Features are Powerful but Less Robust

Our analysis showed that information derived from faces in videos is less robustly available than depth-based features. Faces were detected in about 63% of the frames. While such presence of a face was the best single indicator for attention for unknown users (with 60% accuracy, see Table 1), it also meant that many frames had no face based on which the other features such as gaze could be assessed. For this reason, head pose and gaze information was only available for about a tenth of our data points (see Figure 3).

We conclude that face images are a powerful source of information for attention detection, but they may not always be robustly available. While future systems could aim to improve on this with high-resolution video cameras, such a technical improvement may still not capture cases where users' faces are simply not turned towards the system.

We thus recommend to combine both depth-based features and face features. This is backed by our observation that combining face features with depth-based location features even improved prediction accuracy for unknown users, where location features on their own were less accurate (see Figure 4).

CONCLUSION

In this work we presented the first quantitative comparison of features available from depth and video data with regard to how accurately they can determine user attention towards a display. We found that face presence is the most accurate feature in settings with unknown users, achieving at least 60% accuracy (13.0% better than our baseline). Furthermore, we found that accuracy can be increased in settings with a known user base, in particular with features that preserve the user's privacy (+77.3% accuracy compared to the baseline).

REFERENCES

1. Christopher Ackad, Andrew Clayphan, Martin Tomitsch, and Judy Kay. 2015. An In-the-wild Study of Learning Mid-air Gestures to Browse Hierarchical Information at a Large Interactive Public Display. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '15)*. ACM, New York, NY, USA, 1227–1238.

2. Florian Alt, Andreas Bulling, Gino Gravanis, and Daniel Buschek. 2015. GravitySpot: Guiding Users in Front of Public Displays Using On-Screen Visual Cues. In *Proc. of the 28th ACM Symposium on User Interface Software and Technology (UIST 2015)*.

3. Florian Alt, Jörg Müller, and Albrecht Schmidt. 2012a. Advertising on public display networks. *Computer* (2012), 50–56.

4. Florian Alt, Stefan Schneegaß, Albrecht Schmidt, Jörg Müller, and Nemanja Memarovic. 2012b. How to Evaluate Public Displays. In *Proceedings of the 2012 International Symposium on Pervasive Displays (PerDis '12)*. ACM, New York, NY, USA.

5. Gilbert Beyer, Florian Alt, Jörg Müller, Albrecht Schmidt, Karsten Isakovic, Stefan Klose, Manuel Schiewe, and Ivo Haulsen. 2011. Audience Behavior Around Large Interactive Cylindrical Screens. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '11)*. ACM, New York, NY, USA, 1021–1030.

6. Nicholas S. Dalton, Emily Collins, and Paul Marshall. 2015. Display Blindness?: Looking Again at the Visibility of Situated Displays Using Eye-tracking. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems (CHI '15)*. ACM, New York, NY, USA, 3889–3898.

7. Jörg Auf dem Hövel. 2005. Der mobile Kunde im Visier. *Telepolis* (2005). `http://www.heise.de/tp/artikel/21/21482/1.html`

8. Jakub Dostal, Uta Hinrichs, Per Ola Kristensson, and Aaron Quigley. 2014a. SpiderEyes: Designing Attention- and Proximity-aware Collaborative Interfaces for Wall-sized Displays. In *Proceedings of the 19th International Conference on Intelligent User Interfaces (IUI '14)*. ACM, New York, NY, USA, 143–152.

9. Jakub Dostal, Per Ola Kristensson, and Aaron Quigley. 2014b. Estimating and using absolute and relative viewing distance in interactive systems. *Pervasive and Mobile Computing* (2014), 173–186.

10. J.T. Enns, E.L. Austen, V. Di Lollo, R. Rauschenberger, and S. Yantis. 2001. New Objects Dominate Luminance Transients in Setting Attentional Priority. *Journal of Experimental Psychology: Human Perception and Performance* 6 (2001), 1287.

11. S.L. Franconeri and D.J. Simons. 2003. Moving and Looming Stimuli Capture Attention. *Attention, Perception, & Psychophysics* 7 (2003), 999–1010.

12. Laurent Itti, Christof Koch, and Ernst Niebur. 1998. A Model of Saliency-Based Visual Attention for Rapid Scene Analysis. *IEEE Transactions on Pattern Analysis and Machine Intelligence* 11 (1998), 1254–1259.

13. J. Jonides and S. Yantis. 1988. Uniqueness of Abrupt Visual Onset in Capturing Attention. *Attention, Perception, & Psychophysics* 4 (1988), 346–354.

14. Christian Lander, Sven Gehring, Antonio Krger, Sebastian Boring, and Andreas Bulling. 2015. GazeProjector: Accurate Gaze Estimation and Seamless Gaze Interaction Across Multiple Displays. In *Proceedings of the 28th ACM Symposium on User Interface Software and Technology (UIST'15)*. ACM, New York, NY, USA.

15. Nemanja Memarovic, Sarah Clinch, and Florian Alt. 2015. Understanding Display Blindness in Future Display Deployments. In *Proceedings of the 4th International Symposium on Pervasive Displays (PerDis '15)*. ACM, New York, NY, USA, 7–14.

16. Hermann J Müller and Patrick M Rabbitt. 1989. Reflexive and voluntary orienting of visual attention: time course of activation and resistance to interruption. *Journal of Experimental psychology: Human perception and performance* 2 (1989), 315.

17. Jörg Müller, Florian Alt, Daniel Michelis, and Albrecht Schmidt. 2010. Requirements and Design Space for Interactive Public Displays. In *Proceedings of the International Conference on Multimedia (MM '10)*. ACM, New York, NY, USA, 1285–1294.

18. Jörg Müller, Robert Walter, Gilles Bailly, Michael Nischt, and Florian Alt. 2012. Looking Glass: A Field Study on Noticing Interactivity of a Shop Window. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 297–306.

19. Andreas Sippl, Clemens Holzmann, Doris Zachhuber, and Alois Ferscha. 2010. Real-time gaze tracking for public displays. In *Proceedings of the First International Joint Conference on Ambient Intelligence*. Springer, Berlin-Heidelberg, 167–176.

20. Fabian Timm and Erhardt Barth. 2011. Accurate Eye Centre Localisation by Means of Gradients. In *Proceedings of the International Conference on Computer Vision Theory and Applications (VISAPP'11)*. 125–130.

21. Matthew Turk and Alex P Pentland. 1992. Face recognition system. (Nov. 17 1992). US Patent 5,164,992.

22. Robert Walter, Gilles Bailly, and Jörg Müller. 2013. StrikeAPose: Revealing Mid-air Gestures on Public Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 841–850.

23. Miaosen Wang, Sebastian Boring, and Saul Greenberg. 2012. Proxemic Peddler: A Public Advertising Display That Captures and Preserves the Attention of a Passerby. In *Proceedings of the 2012 International Symposium on Pervasive Displays (PerDis '12)*. ACM, New York, NY, USA.

24. Julie R. Williamson and John Williamson. 2014. Analysing Pedestrian Traffic Around Public Displays. In *Proceedings of the International Symposium on Pervasive Displays (PerDis '14)*. ACM, New York, NY, USA.

25. Xuehan Xiong and Fernando De la Torre. 2013. Supervised Descent Method and Its Applications to Face Alignment. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*. IEEE Computer Society, Washington, DC, USA, 532–539.

26. Xucong Zhang, Yusuke Sugano, Mario Fritz, and Andreas Bulling. 2013. Appearance-Based Gaze Estimation in the Wild. In *Proceedings of the 2013 IEEE Conference on Computer Vision and Pattern Recognition (CVPR '13)*. IEEE Computer Society, Washington, DC, USA, 532–539.

27. Yanxia Zhang, Ming Ki Chong, Jörg Müller, Andreas Bulling, and Hans Gellersen. 2015. Eye Tracking for Public Displays in the Wild. *Personal and Ubiquitous Computing* 19, 5 (2015), 967–981.