Analyzing Visual Attention During Whole Body Interaction with Public Displays

Robert Walter[1] robert.walter@tu-berlin.de
Andreas Bulling[3] bulling@mpi-inf.mpg.de
David Lindlbauer[2] david.lindlbauer@tu-berlin.de
Martin Schuessler[1] schuesslerm@acm.org
Jörg Müller[4] joerg.mueller@acm.org

[1]Quality and Usability Lab, Telekom Innovation Laboratories, TU Berlin, Berlin, Germany [2]TU Berlin, Berlin, Germany
[3]Max Planck Institute for Informatics, Saarbrücken, Germany [4]Aarhus University, Aarhus, Denmark

ABSTRACT
While whole body interaction can enrich user experience on public displays, it remains unclear how common visualizations of user representations impact users' ability to perceive content on the display. In this work we use a head-mounted eye tracker to record visual behavior of 25 users interacting with a public display game that uses a silhouette user representation, mirroring the users' movements. Results from visual attention analysis as well as post-hoc recall and recognition tasks on display contents reveal that visual attention is mostly on users' silhouette while peripheral screen elements remain largely unattended. In our experiment, content attached to the user representation attracted significantly more attention than other screen contents, while content placed at the top and bottom of the screen attracted significantly less. Screen contents attached to the user representation were also significantly better remembered than those at the top and bottom of the screen.

ACM Classification Keywords
H.5.m. Information Interfaces and Presentation (e.g. HCI): Miscellaneous

Author Keywords
Public Displays; Whole Body Interaction; User Representation; Visual Attention; Mobile Eye Tracking

INTRODUCTION
In recent years, displays have been widely deployed in public spaces, such as airports and shopping malls. Public displays become increasingly interactive and physical as they are equipped with a variety of sensors, such as depth cameras or eye trackers. Such sensors allow for robust tracking of users in front of the display and enable interactions using full body movements [6], gestures [12], or eye gaze [13, 14, 9].

UbiComp '15, September 7–11, 2015, Osaka, Japan.

ACM 978-1-4503-3574-4/15/09…$15.00.
http://dx.doi.org/10.1145/2750858.2804255

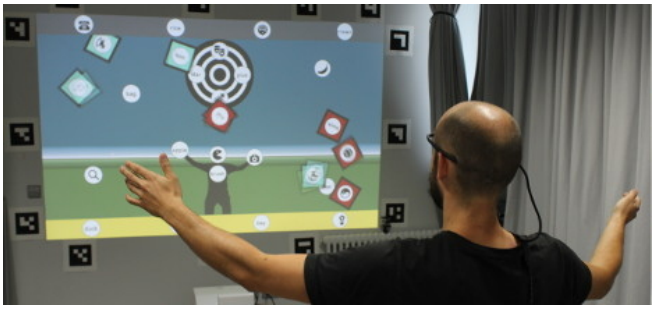

Figure 1. We used a head-mounted eye tracker to analyze visual behavior of users of interactive public displays. We show that people mostly attend the silhouette representation on the screen, especially during the first few minutes of interaction.

Previous work illustrated the advantages of visual user representations that mirror body movements, e.g., in the form a silhouette, or a virtual avatar [6, 11, 12, 7, 8]. A large body of work studied gaze for interaction with public displays and the development of attention-aware interfaces (see [13, 9, 3, 10] for examples). While the benefits of physical interactions on user experience are well explored, few previous work studied the effects of such visual feedback mechanisms on user behavior and performance [11, 14]. Most importantly, it still remains unclear if and how these mechanisms impact visual attention and thus users' ability to perceive content shown on other parts of the display.

To answer this question, we study human visual attention during whole body interaction with a public display game. We observe both spatial distribution of visual attention across different screen elements as well as temporal development of visual behavior over time. While playing the game, users' eye movements are tracked using a head-mounted eye tracker. After playing the game, users perform recall and recognition tasks in which they have to remember screen contents and layout of the interface.

We show that the silhouette user representation attracts significantly more visual attention than all other screen elements. In addition, we found that elements placed at the top or bottom of the screen received significantly less attention than all other elements and were remembered significantly less well than items placed on the silhouette. Moreover, an analysis of the temporal development of attention suggests that visual behavior is related to the extend to which users have understood the different interface elements and the intended interaction.

EXPERIMENT

During earlier deployments of interactive public displays we repeatedly observed that users had difficulties to notice and remember content, hints, and messages shown on the display [6, 11, 12]. Although these messages typically appeared in the center of the display and therefore should have been obvious to spectators, interacting users often reported not having noticed them. To investigate these observations further, we conducted a controlled laboratory study and systematically analyzed visual behavior during public display interaction. We implemented a playful public display application involving users playing with an on-screen silhouette representation of themselves. We particularly focused on the first seconds and minutes of interaction, where users are still novices and have no or little understanding of the game and which items on the screen are important to play it successfully.

Interface Elements

We divided the interface elements into seven categories (see Figure 2): The 1) *user* representation consisted of a silhouette, directly mirroring the user's body movements in real-time. This representation has been shown to be effective for communicating the interactivity of public displays [6]. Similar user representations are commonly used for public displays [12, 2, 1], as well as for many *Microsoft Kinect*-enabled games. The user representation is not fixed to the center of the screen but leverages the entire horizontal screen space. To achieve this, we exaggerated horizontal translation of the user representation on the screen (1m of horizontal user movement maps to the entire screen width). The reason for this is to allow users to reach objects at outer regions of the screen more easily. 2) *Interactive* objects were represented by moving (physically simulated) cubes of a specific color and could be manipulated (tossed around) via the user representation. 3) *Non-interactive* objects were shown as randomly moving cubes of a different color, and therefore could attract users' attention, but could not be manipulated. They were moved by applying random pulse forces, similar to actual hits from the user. A fixed 4) *target* was neither moving, nor could it be manipulated. As interactive objects hit the target, the game score visualized by a progress bar was increased and the object stuck to the target for two seconds. A 5) *top* and 6) *bottom* bar are common locations to position additional status information (e.g., score counter, remaining level time, or text hints in an interactive game). They were used to display the game score in our study. Finally, information may also be presented in the static 7) *background* of the interactive scenery.

In each of these seven categories, four items were displayed ($7 \times 4 = 28$ simultaneously shown items in total). The items could be either an *icon* or *text*, with each category containing two of each type. We randomly altered three parameters between participants: 1) because some items may receive more attention or are remembered better than others, they were randomly picked (from a set of 56 items in total), and shuffled within the seven categories. 2) As attention and recognition may also be influenced by the color of the stimuli, we randomly altered the colors of cubes, top and bottom bars. 3) Finally, either the *top* or the *bottom* bar was randomly picked to represent the score counter bar: the size of the bar increased

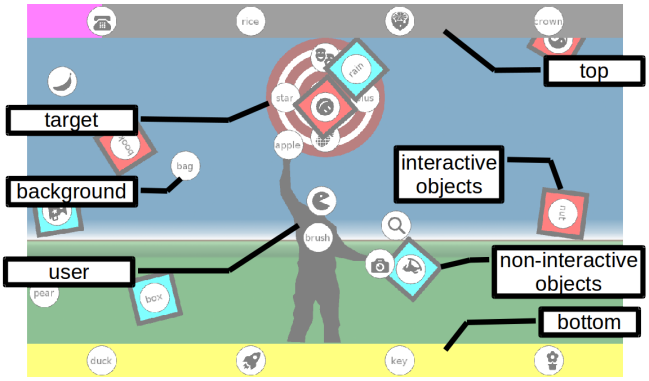

Figure 2. Interactive public display used in the laboratory study. The interface was divided into seven categories: 1) user representation, 2) interactive and 3) non-interactive objects, 4) target, 5) top (with score bar in pink) and 6) bottom bar, as well as 7) background.

as more interactive objects were successfully placed into the target area (see Figure 2). The goal of the public display game was to toss interactive virtual objects into a target area. To better resemble a public scenario, this goal and the game mechanics were not explained to the participants, and had to be explored and understood during interaction.

Participants and Apparatus

25 paid participants (9 female) with normal or corrected vision, aged between 17 and 36 years ($M = 26.2, SD = 4$), successfully participated in the study. We discarded data of another three participants due to technical issues (e.g., insufficient eye tracking quality). The game application was shown on a 100" (254cm) wall projection at a resolution of 1280×800 pixels. Participants were interacting with the system from a distance of about 3.75m. Body movements were captured using a *Microsoft Kinect* depth camera while visual behavior was captured using a PUPIL head mounted eye tracker [5]. The eye tracker achieves an average gaze estimation accuracy of 0.6 degree of visual angle (0.08 degree precision) according to the manufacturer. This maps to an accuracy of 24 pixels on the screen, which was confirmed during our calibration routines. The target size was 64x64 pixels on the screen. The eye tracker weighs about 100g and allows participants to freely move in front of the screen. We attached 13 visual markers around the screen and used the marker tracking provided by PUPIL to automatically map gaze coordinates to screen coordinates (see Figure 1).

Tasks

The study consisted of three tasks: Participants were first asked to 1) *interact* with the game for five minutes. No further instructions or explanations of the upcoming tasks were provided. Afterwards followed a 2) *recall* task: participants were asked to sit down at a table and to draw the user interface using a graphics tablet. They were supposed to reproduce the game interface from their memory as detailed as possible. The drawing canvas was shown on the same display and from the same distance as during the interaction task. Finally, in a 3) *recognition* task, participants were presented with the test set of 56 items of which only 28 had actually appeared in the application at different locations. They were asked to

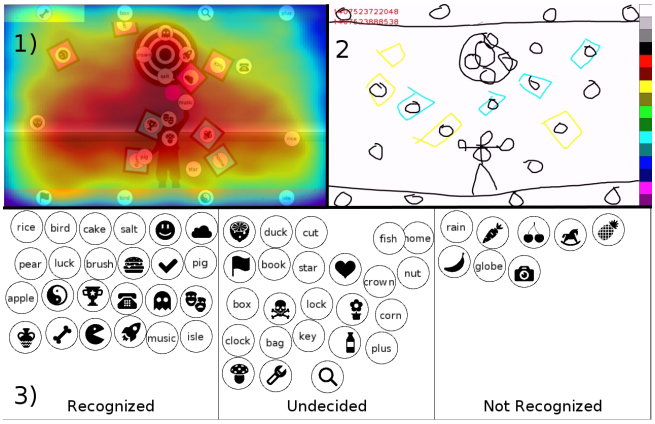

Figure 3. Overview of the three study tasks. 1) *Interaction:* Screenshot of user interacting with the system. Overlaid heat map shows the distribution of visual attention averaged across all participants. 2) *Recall:* Visual reproduction of screen contents by one participant in drawing-task. 3) *Recognition:* Exemplary results of a participant after classification.

classify each item as *recognized*, *undecided* or *not recognized* (see Figure 3).

Procedure and Methodology

Participants were briefly introduced to the experimental equipment, whereas the public display application itself was not explained and had to be explored by participants during the *interaction* task. Participants put on the eye tracker, which was then calibrated using a standard 9-point calibration routine and were then guided through the three study tasks. For both the *recall* and *recognition* task we followed a think-aloud protocol. Because body movements in the interaction task may cause the eye tracker to slightly dislocate, the experimenter constantly monitored the gaze estimation accuracy in real-time on a separate screen. To maintain high accuracy, a recalibration of the eye-tracker was automatically triggered after 90 seconds, or manually at any time on the experimenters demand. During recalibration, the application was paused and hidden. After the participant had finished all tasks, a semi-structured interview was conducted. Questions of the interview included, if participants 1) noticed the difference between interactive and non-interactive cubes, 2) noticed the animated score bar, and 3) think they could remember either *icon* or *text* items better. Our study followed a within-participant design. Our independent variable *category* has seven levels (*user, inter, non-inter, target, top, bottom, bg*). As dependent variables we measured the number of gaze samples on the categories and the recognition rate of items placed in different categories. A gaze sample was considered to be on a category, when the item with the smallest Euclidean distance to the gaze point (< 128 pixels) was in that category.

RESULTS

We first analyzed the distribution of visual attention (measured by the number of gaze samples) across the different interface element categories (see Figure 4). We subsequently analyzed participants' performance in the recall and recognition tests (measured by the number of recalled and recognized items).

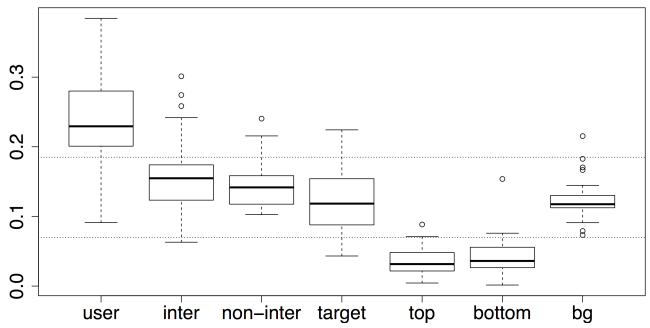

Figure 4. Visual attention on the different categories (ratios). *User* receives significantly more attention than all the other categories, while *top* and *bottom* receive significantly less than all the others. Horizontal lines separate significantly different clusters.

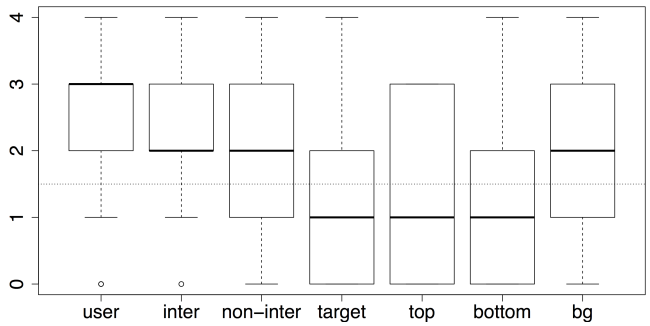

Figure 5. Number of recognized items (out of four) that were shown at different interface categories. Items attached to the user representation were recognized significantly more often than those placed in the top and bottom bars, as well as on the target. The horizontal line separates significantly different clusters.

Visual Attention

A Kruskal-Wallis test revealed a significant effect of screen category on visual attention ($\chi^2(6) = 111.3, p < 0.001$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between *user* vs. all other categories ($p < 0.01$), and *top*/*bottom* vs. all other categories ($p < 0.001$) and *interactive* vs. *background* ($p < 0.05$). The silhouette representation drew most attention, while top and bottom bars drew least attention. Interactive objects drew significantly more attention than background objects.

Looking in more detail at the user representation, a Kruskal-Wallis test revealed a significant effect of the body part in user representation on visual attention ($\chi^2(3) = 34.4, p < 0.001$). A post-hoc test using Mann-Whitney tests with Bonferroni correction showed the significant differences between *torso* vs. all other body parts ($p < 0.01$), with torso drawing the least attention.

Recognition and Recall

A Kruskal-Wallis test revealed a significant effect of the screen category on the number of recognized items ($\chi^2(6) = 23.9, p < 0.001$). A post-hoc Mann-Whitney test with Bonferroni correction showed the significant differences between *user* and *top*/*bottom*/*target* ($p < 0.05$). Items attached to the user representation were recognized more often than items placed in the top and bottom bars, as well as on the target.

For the recall tasks, the user representation was drawn only by 22 of 25 participants in the recall task, although it received most visual attention and the highest values in the recognition task. Only the interactive objects, non-interactive objects and goal were drawn by all users. The top- and bottom bar were drawn by 19 and 18 participants, respectively, while background objects were drawn by 22 participants. The score counter progress bar was drawn only 11 times.

Visual behavior over time

We observed a shift in visual attention from the user representation towards the interactive objects over the time of interaction (see Figure 6). This transition occurred for most participants, but at different points of time between the 1st and the 5th minute. A series of repeated measures ANOVAs (adjusted $\alpha = .007$) revealed a significant difference in attention between the first and the last minute. Specifically, fixations on user representation items decreased (30% to 19%, $p < .001$ pairwise, main effect $F(4, 96) = 7.14, p < .001$), while fixations on interactive objects increased (9% to 22% $p < .001$ pairwise, main effect $F(2.27, 54.49), p < .001$).

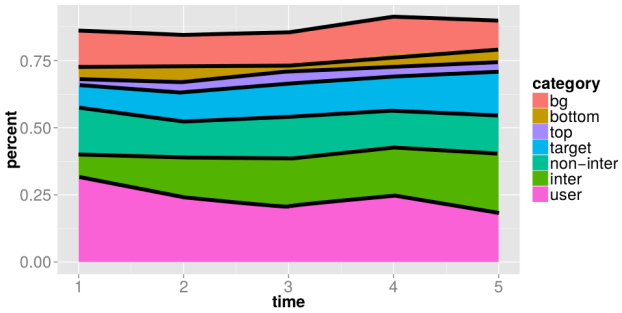

Figure 6. Averaged temporal development of visual behavior over time of all participants. It can be seen that the user representation draws the most attention especially in the beginning of the interaction. Towards the end, it shifts to the interactive objects.

Qualitative Findings

We observed different exploration strategies among users. While most continuously discovered the game while interacting, others paused to get an overview of which objects were presented to them. As users immediately identified themselves with the displayed silhouette, items attached to it were of much interest: "Why am I an *apple* - why is my head a *flower*?" Two users even lifted their feet because they were intrigued whether they also had items attached. Items placed at the users hands were sometimes pushed against other items (that where believed to have a matching or opposing property) in order to combine them: "I tried to combine the smiley with the skull, because one seemed to be sad and the other happy". We observed that users often hesitated to draw the silhouette in the *recall* task: "Am I also supposed to draw *myself*?". Three users did not draw their silhouette at all.

DISCUSSION

We believe that the observed shift in visual attention is related to learning of control: as users learned how to control the game through their silhouette, they focussed less on their user representation, and more on what they want to control (the interactive objects). Thus, we think that visual attention patterns can potentially be used to estimate the level of experience of users.

While the *torso* is the body part that attracted least visual attention, we could not find a significant difference between the other body parts (*head*, *left hand*, *right hand*) among all participants. However, when investigating the distribution of visual attention of individual participants, it appears that they mostly focused on one particular body part. This could either be the *left hand* or *right hand*, or in some cases also the *head*. We believe that people focus on the object that represents their tool to *manipulate* the virtual environment. Thus future work should investigate if other user representations (e.g., a hand cursor) attract a similar amount of attention, and if stimuli specifically *designed* to attract attention (e.g. looming stimuli [4]) can direct the attention away from the silhouette without causing distraction.

We repeatedly observed that participants did not consider their silhouette representation as an interface element. In the think-aloud tasks and the post-hoc interview they mostly referred to it as *me*. It appears that this kind of representation creates a cognitive impression of presence to users. Future work should investigate the relation between degree of abstraction of user representations and degree of immersion.

Limitations

Due to current technical limitations, we were only able to do the visual attention analysis of the public display application in a laboratory setting. We are not aware of remote eye-tracking solutions with the required accuracy, that do not need a calibration and could thus be seamlessly applied in an ecologically valid field setting. In a field scenario, we would expect users to be more distracted by the environment and to interact for shorter durations [6].

CONCLUSION

In this work we studied human visual attention during whole body interactions with a public display game as well as post-hoc recall and recognition performance of screen elements. We showed that the silhouette user representation draws significantly more visual attention than all other screen elements. In addition, we found that elements placed at the top or bottom of the screen received significantly less attention than all other elements and are remembered significantly less than items placed on the silhouette. These findings provide important guidance for designers of public display applications that rely on silhouettes as a user representation for whole body interaction. Concluding from our study results with respect to the applied stimuli, we propose to attach messages, such as gesture hints, directly on the user's contour instead of placing them at the bottom or top of the screen, to allow users to perceive them more easily.

ACKNOWLEDGMENTS

This research was funded by EIT, BMBF (grant 01IS12056) and ERC (grant ERC-2010-StG 259550 *XSHAPE*). Special thanks go to Viktor Miruchna and Ines Ben Said for their valuable support.

REFERENCES

1. Imeh Akpan, Paul Marshall, Jon Bird, and Daniel Harrison. 2013. Exploring the Effects of Space and Place on Engagement with an Interactive Installation. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 2213–2222. DOI: **http://dx.doi.org/10.1145/2470654.2481306**
2. Florian Alt, Stefan Schneegass, Michael Girgis, and Albrecht Schmidt. 2013. Cognitive Effects of Interactive Public Display Applications. In *Proceedings of the 2Nd ACM International Symposium on Pervasive Displays (PerDis '13)*. ACM, New York, NY, USA, 13–18. DOI: **http://dx.doi.org/10.1145/2491568.2491572**
3. Jakub Dostal, Uta Hinrichs, Per Ola Kristensson, and Aaron Quigley. 2014. 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. DOI: **http://dx.doi.org/10.1145/2557500.2557541**
4. Steven L. Franconeri and Daniel J. Simons. 2003. Moving and looming stimuli capture attention. *Perception & Psychophysics* (2003), 999–1010.
5. Moritz Kassner, William Patera, and Andreas Bulling. 2014. Pupil: An Open Source Platform for Pervasive Eye Tracking and Mobile Gaze-based Interaction. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct Publication (UbiComp '14 Adjunct)*. ACM, New York, NY, USA, 1151–1160. DOI: **http://dx.doi.org/10.1145/2638728.2641695**
6. 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. DOI: **http://dx.doi.org/10.1145/2207676.2207718**
7. Garth Shoemaker, Anthony Tang, and Kellogg S. Booth. 2007. Shadow Reaching: A New Perspective on Interaction for Large Displays. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology (UIST '07)*. ACM, New York, NY, USA, 53–56. DOI: **http://dx.doi.org/10.1145/1294211.1294221**
8. Garth Shoemaker, Takayuki Tsukitani, Yoshifumi Kitamura, and Kellogg S. Booth. 2010. Body-centric Interaction Techniques for Very Large Wall Displays. In *Proceedings of the 6th Nordic Conference on Human-Computer Interaction: Extending Boundaries (NordiCHI '10)*. ACM, New York, NY, USA, 463–472. DOI:**http://dx.doi.org/10.1145/1868914.1868967**
9. Sophie Stellmach and Raimund Dachselt. 2012. Look & Touch: Gaze-supported Target Acquisition. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '12)*. ACM, New York, NY, USA, 2981–2990. DOI: **http://dx.doi.org/10.1145/2207676.2208709**
10. Roel Vertegaal. 2003. Attentive user interfaces. *Commun. ACM* 46, 3 (2003), 30–33.
11. 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. DOI: **http://dx.doi.org/10.1145/2470654.2470774**
12. Robert Walter, Gilles Bailly, Nina Valkanova, and Jörg Müller. 2014. Cuenesics: Using Mid-air Gestures to Select Items on Interactive Public Displays. In *Proceedings of the 16th International Conference on Human-computer Interaction with Mobile Devices & Services (MobileHCI '14)*. ACM, New York, NY, USA, 299–308. DOI: **http://dx.doi.org/10.1145/2628363.2628368**
13. Yanxia Zhang, Andreas Bulling, and Hans Gellersen. 2013. SideWays: A Gaze Interface for Spontaneous Interaction with Situated Displays. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems (CHI '13)*. ACM, New York, NY, USA, 851–860. DOI: **http://dx.doi.org/10.1145/2470654.2470775**
14. Yanxia Zhang, Jörg Müller, Ming Ki Chong, Andreas Bulling, and Hans Gellersen. 2014. GazeHorizon: Enabling Passers-by to Interact with Public Displays by Gaze. In *Proceedings of the 2014 ACM International Joint Conference on Pervasive and Ubiquitous Computing (UbiComp '14)*. ACM, New York, NY, USA, 559–563. DOI: **http://dx.doi.org/10.1145/2632048.2636071**