UNIVERSITY OF SOUTHAMPTON

Informing User Understanding of Smart Systems through Feedback

by

Jacob Kittley-Davies

A thesis submitted in partial fulfillment for the degree of Doctor of Philosophy

in the
Faculty of Engineering and Physical Science
School of Electronics and Computer Science

March 2020

UNIVERSITY OF SOUTHAMPTON

ABSTRACT

FACULTY OF ENGINEERING AND PHYSICAL SCIENCE
SCHOOL OF ELECTRONICS AND COMPUTER SCIENCE

Doctor of Philosophy

by Jacob Kittley-Davies

Recent advances in microprocessing and low power radio technologies have catalyzed the transition of smart technologies from the domain of researchers and enthusiasts to everyday consumers. This new wave of smart devices, and the systems they form, marks a significant step towards Weiser's vision of ubiquitous computing and offers users a wealth of new and exciting opportunities. However, smart technologies are inherently complex and without careful design can prove complicated and confusing for users with no specific knowledge of the underpinning technologies. A poor understanding has the potential to inhibit user experience and may result in the abandonment of technologies which otherwise could bring real benefits to users.

While a considerable body of work exists examining how confusion arising from complexity can be addressed, this work largely focuses on traditional heuristic systems. The non-deterministic nature of some smart technologies and the capacity for the sophisticated interconnected processes they employ to mask the relationship between system inputs and outcomes exacerbate the challenges examined in prior work. There is therefore a need to investigate how these challenges can be overcome for users of smart systems in particular.

This thesis reports a series of five user studies, conducted under both controlled conditions and in the field. In particular, we examine how feedback can be used to inform user understanding of sensor based smart systems. Through qualitative and quantitative analysis we observe and evaluate over 145 participants interacting with sensor based smart systems. From our findings we identify a number of design implications and highlight the pitfalls of poor and uninformed design.

Declaration of Authorship

I, Jacob Kittley-Davies declare that this thesis and the work presented in it are my own and has been generated by me as the result of my own original research.

I confirm that:

- This work was done wholly or mainly while in candidature for a research degree at this University;
- Where any part of this thesis has previously been submitted for a degree or any other qualification at this University or any other institution, this has been clearly stated;
- Where I have consulted the published work of others, this is always clearly attributed;
- Where I have quoted from the work of others, the source is always given. With the exception of such quotations, this thesis is entirely my own work;
- I have acknowledged all main sources of help;
- Where the thesis is based on work done by myself jointly with others, I have made clear exactly what was done by others and what I have contributed myself;
- Parts of this work have been published, please see Section 1.6 for details.

Signed:

Date:

Contents

List of Tables

List of Figures

Acknowledgements

First and foremost, I would like to express my deepest gratitude to my supervisors, Enrico Costanza, Alex Rogers and Sebastian Stein. Without their expert guidance, support and constructive criticism the work contained within this thesis would not have been possible. I would also like to acknowledge here, the funding bodies who financed this research: the University Of Southampton ECS DTA Studentship award, the APERIO project (EP/L024608/1) and the AIoT (EP/N014243/1) project.

I would also like to extend my thanks to my family and friends who have supported me throughout my undergraduate degree and this PhD. In particular, I would like to thank my wife, Jenny and our two children, Harry and Molly, for the sacrifices they have made, their kindness, love and understanding.

To my colleagues at the University of Southampton and University College London, thank you for your friendship, advice and guidance. Without your help many of the challenges I faced during my studies would have been insurmountable.

Chapter 1

Introduction

Since Weiser (1991) first laid out his vision of ubiquitous computing, considerable progress has been made to overcome the associated technical challenges. Most recently, improvements in microprocessor efficiency, faster internet connection speeds, lower power radios and the advent of cloud computing have catalysed the growth of so-called “smart” devices and systems. Smart devices represent the latest evolutionary step towards the goal of truly ubiquitous computing. Despite the popularity of the term “smart”, defining exactly what constitutes a smart device or system has proved to be non-trivial. The next section addresses the issue of defining a smart system. For clarity here a “smart system” is summarised as: a device or collection of connected devices which can sense their environment, communicate with other systems and compute the information they gather.

While clearly defining smart systems is important when discussing the scope of this work, the lack of a consensus has not inhibited researchers and product developers. Thus far smart technologies have been used to endow all manner of previously “dumb” objects with seemingly intelligent functionality (e.g. household lighting which adapts to sustain light levels and heating thermostats which know when you are not at home) and they have facilitated a gamut of novel products such as voice assistants and personal health trackers. While there is still some way to go before they disappear, “weaving their way into the fabric of everyday life” (Weiser, 1991), the potential for smart systems to enrich users’ everyday lives is palpable.

However, despite increasing adoption (Deloitte, 2017), users have been shown to abandon smart systems if they do not prove sufficiently useful in a timely manner (Lazar et al., 2015). While it remains an open question why exactly users abandon smart devices, the Human-Computer Interaction literature contains a wealth of research which demonstrates how poor understanding can be detrimental to user interactions with traditional (i.e. not smart) systems. For example, that misunderstandings can impact users’ capacity to interact effectively (du Boulay et al., 1981; Bayman and Mayer, 1984), operate systems to their satisfaction (Kulesza et al., 2012; Cramer et al., 2008), and perform tasks efficiently (Kieras and Bovair, 1984). User understanding has been at the heart of HCI research for many decades and numerous theories, models and frameworks have been developed and adapted from other disciplines (e.g. activity theory and distributed cognition). This thesis draws heavily on mental models, a conceptual framework originating from cognitive psychology and popularised in HCI by Norman (1983). They have proven to be a useful framework through which user understanding and interactions with systems can be modelled and evaluated (Rogers et al., 2011).

The inherent complexity of smart systems and their perceived non-deterministic nature is likely to only compound the challenges documented for traditional systems, impeding users' ability to develop coherent and useful mental models. Moreover, smart systems commonly employ sensors which until recently have only been used in specialist systems. The environmental factors which can affect a sensor's performance can make reasoning about system behaviour more challenging and gauging the appropriate response to unexpected outcomes more difficult. The rapid adoption of digital cameras in smartphones, for example, has resulted in the increasing use of computer vision-based interaction modalities (Amershi et al., 2019), such as augmenting photographs for entertainment[1] or reverse image searching[2]. While using cameras in this way has many benefits, environmental conditions such as bright sunlight, shadows and reflections can all impact system performance in ways which are not obvious to users. Understanding the implications of these factors on smart system users and how best to design interactions to overcome them is therefore increasingly important. However, literature in this space is sparse.

While complexity is often cited as the problem, it is important to acknowledge that it is not always feasible to reduce it. Complexity is a naturally occurring consequence of functionality and for any system there is a certain amount of complexity which cannot be reduced without removing functionality a.k.a. Tesler's Law, the Conservation of Complexity (Saffer and Tesler, 2007). Instead, the challenge is to overcome the complications and confusion which arise from complexity, a challenge which can be addressed by informing users' understanding through good design, i.e. when we understand something complex, it remains complex, but it is no longer confusing (Norman, 2010). Smart systems and the sophisticated technologies they employ are inherently complex. Ensuring that users develop useful mental models so that their interactions with them are not confusing or complicated is therefore of paramount importance.

Explicitly explaining behaviour to users has been shown to improve user understanding of some smart technologies e.g. Lim et al. (2009) and Koo et al. (2015). However, as smart technologies transition from the domain of experts and enthusiasts to everyday users, the underlying assumption that users are sufficiently interested in learning about

[1] e.g. Instagram's photo filters: https://help.instagram.com/

[2] e.g. Samsung's Bixby search tools: https://www.samsung.com/bixby/

the system cannot be taken for granted (Yang and Newman, 2013), nor should their technical literacy (Edwards and Grinter, 2001).

Historically, feedback has been used to great effect to improve user interactions (e.g. du Boulay et al. (1981); Costanza et al. (2010); Garcia et al. (2016)), but less is known about its efficacy with smart systems and how it might be used to foster greater user understanding of system processes. Feedback design is often discussed in respect of the short-term interaction e.g. how it can be used to acknowledge user actions. While this is arguably its primary purpose, feedback also has longer-term implications and will educate users' understanding by affirming good interactions, highlighting unhelpful ones and suggesting appropriate future courses of action. Nielsen (1994) alludes to this when he describes the role of feedback as going beyond simply notifying users of errors and outcomes, explaining that feedback should keep a user informed of what is happening and why. The theory of action (Norman and Draper, 1986) deconstructs users' interactions with technology into a series of stages. These stages are commonly partitioned into two groups, referred to as gulfs: the gulf of execution - the stages which represent the process through which a user works out what actions can be taken and how to perform them, and the gulf of evaluation - stages representing how users reason about the outcomes of their actions (Hutchins et al., 1985; Norman, 2013). The role of feedback through this lens is one of the most important tools in a designer's toolbox which can be used to overcome the gulf of evaluation, both in the short-term and the long-term formation of useful mental models (Norman, 2013). This thesis examines the role of feedback as a mechanism to inform users' understanding of smart systems while interactions are taking place.

While there is no research in the public domain as yet to support feedback's application to inform users in this way, some comercial smart systems have begun to incorporate feedback which is suggestive of this goal. For example, Amazon's shopping application which allows users to search for products using images captured with their smartphone's camera, overlays the viewfinder with visual markers that signify features of interest identified by the detection algorithms. While it is only possible to speculate as to why Amazon has chosen to incorporate feedback in this way, it does emphasize the need to know more about how feedback can be used to inform user understanding and the potential pitfalls if it is not designed correctly.

To summarise, lay users are now living in a complex digital world where sophisticated smart technologies are embedded in everyday artifacts and portable computers are the norm (e.g. smartphones and tablets). The inter-operational architecture of these devices and the systems they form, allows them to collect, process and feed back information at unprecedented speeds and in ever-increasing volumes (Jennings et al., 2014). Prior work investigating traditional systems has demonstrated the importance of user understanding, showing how misconceptions and misunderstandings can be detrimental to interaction, while coherent mental models improve efficacy, satisfaction and trust. However, how this work relates to smart systems is unclear. This thesis contributes towards addressing a gap in the literature, investigating the role of feedback in informing user understanding of smart systems, highlighting observed implications for design and illustrating how feedback can be used to help users overcome unexpected outcomes without becoming domain experts.

To this end, this thesis reports a series of user studies which investigate how feedback derived from smart systems processes can inform user understanding. Calling on a combination of quantitative and qualitative analysis, over 140 participants were observed and evaluated interacting with sensor-based prototype smart systems. This work reveals and reflects on the design implications of study findings, the challenges of examining user interactions at the boundaries of satisfactory outcomes (e.g. success and failure), and the perils of uninformed design.

1.1 Definitions of Smart Systems

While it is broadly accepted that the term "smart" refers to a device or system's ability to sense, compute and communicate (e.g. Siegemund (2004); Thompson (2005); Lopez et al. (2017)), the rapid proliferation of consumer goods branded as "smart" has resulted in the term becoming less well defined. Definitions within the scientific community are less diverse, yet despite numerous attempts to formalise the term (e.g. Heppelmann et al. (2017) and Streitz et al. (2005)), there remains no definitive agreed-upon definition.

To provide context to this thesis, in Chapter 2.1, definitions used in prior work are reviewed and related literature discussed, from which a definition is derived. It is noted that the inter-operational nature of smart devices makes defining the boundaries between devices and systems equally problematic. To avoid unhelpful differentiation and repetition, in this thesis the term "smart systems" is used to refer to both devices and collections of interconnected devices which act together. In summary, a "smart system" is characterised as a device or collection of devices which is/are able to: sense its environment, communicate with other devices or systems, process information collected or received and directly or indirectly affect an environment. Furthermore, while the word "smart" in everyday language has connotations of intelligence and "smart systems" commonly employ processes which can learn, it is not a necessary attribute. Rather it is smart systems' behaviour which differentiates them from traditional systems i.e. that they appear to be non-deterministic either because they employ technologies such as machine learning or that the relationship between inputs and system behaviour is sufficiently complex that it is challenging for users to reason about it.

1.2 Research Objectives

The overarching objective of this thesis:

> How does the design of smart system feedback impact users' with no specialist knowledge of the underpinning technologies ability to develop useful understandings of system operations such that they can interact more effectively? And what are the implications of bad design?

1.3 Research Requirements

To evaluate the role of feedback in smart systems and how it can be used to improve user understanding, a number of requirements need to be satisfied by this research. This section details and describes them.

For meaningful conclusions to be drawn, it is important to provide a high level of realism. To this end, functional prototype smart systems are necessary. These systems must be sufficiently complex to ensure that users require assistance in overcoming the resulting complication. Prototypes must also encourage user interaction. Studies will require participants to repeatedly interact with the systems for sustained periods of time so that user understanding can develop. Therefore it is important that the tasks are engaging to maintain user attention.

Evaluation of these systems also requires observation of user interactions at the boundaries of success and failure. For the effects of feedback to be measured and so that meaningful comparison can be made, interactions must be controllable such that they can be replicated across participants. However, it is critical that any form of manipulation is not obvious to participants.

In addition to quantifying user interactions in controlled measurable studies, it is also important that findings are validated with in-the-wild observations i.e. when appropriate, field studies should also be conducted to broaden the scope of qualitative data collected.

1.4 Research Challenges

Drawing on prior sections of this introduction, here, the key challenges facing the designers of real-world smart systems are outlined.

- **Understanding when to take action** is a significant problem for users of smart systems, especially those which yield large volumes of information. Studies of recommender and context-aware systems have shown that making users aware of the motivations behind automated decisions through feedback can improve user understanding and trust in the choices made by the system (e.g. Lim and Dey (2011); Bellotti and Edwards (2001); Lim et al. (2009); Herlocker et al. (2000)), however, in these cases the action has already been taken and how these findings relate to invocation of action is unclear. Investigations of software security (Li et al., 2016) have also shown that feedback alerting users of vulnerabilities have a significant effect on the uptake of fixes and patches. Furthermore, messages containing more detailed information are more effective at invoking action. In these studies the need for action is known and the messages are intended to convince rather that inform understanding such that a decision can be made. However, work examining specifically how feedback can invoke action by informing user understanding is still absent from the literature.

- **Understanding which action to take** is equally important for users i.e. identifying which courses of action are most likely to result in satisfactory outcomes. The placement of smart wireless sensors is a good example, with Kazmi et al. (2014), and Fischer et al. (2017) both highlighting the need for research examining how users can be guided to place sensors such that they can reliably communicate with their parent systems, whilst sensing a target with sufficient accuracy. While the HCI community has paid considerable attention to the configuration of rules (e.g. Nandi and Ernst (2016)) and initial system integration of wireless devices (e.g. pairing to the network (Jewell et al., 2015)), less is known about how user understanding can be shaped in this context. Costanza et al. (2010) and Beckmann et al. (2004) have examined the challenges of connectivity and sensor fidelity independently, a matter of increasing importance, as falling hardware costs have also resulted in the burden of installation being passed to end-users (Jakobi et al., 2018).

- **Understanding how to overcome unexpected outcomes:** Systems also have the potential to produce unexpected outcomes. This is especially true for systems which employ pattern machining technologies, where the factors which determine outcomes are not obvious to users (e.g. lighting conditions in camera-based applications). Unrealistic user expectations can also play a role; misunderstandings of system functionality have been reported (Yang and Newman, 2013) to lead to unsatisfactory user interactions. Similarly, studies of procedural computer systems (i.e. not smart) have shown that flawed understanding of system operation can lead to confusion and poor user experience (Kulesza et al., 2015; Tullio et al., 2007). The challenge then is to make the reasons for failures intelligible, without requiring users to become experts. While some work has been conducted into exposing system processes to users (i.e. Patel et al. (2010a); Krause et al. (2016)), their work has focussed on experts. How these findings relate to users with no specialist knowledge of the process is an open question.

In summary, this section outlines three key areas where user understanding can impact users' ability to interact effectively with smart systems, namely: (i) knowing when to take action, (ii) reasoning about which is the correct action, and (iii) understanding how to overcome adversity when exposed to failure.

1.5 Thesis Structure

This thesis details a series of lab and field studies designed to investigate how feedback can be used to foster greater user understanding. In total 145 participants (excluding pilots) were observed across 7 studies. The reporting of this work is divided into the following chapters:

Chapter 2 provides the essential background information required to contextualise the work presented in this thesis. It begins by reviewing the technologies which fall under the "Smart" umbrella and evidences the definition of the terminology used in this document. Finally, in this section, the literature surrounding user understanding and how it can be informed through feedback is reviewed.

Chapter 3 details a user study (n=8) examining how users' impetus to take action is affected by the amount of information contained within the feedback. Participants were tasked with choosing whether action should be taken based on textual feedback messages. Reflecting on the findings, this chapter discusses how user understanding of when to take action is affected by the level of detail encoded in feedback.

Chapter 4 documents a series of investigations examining how feedback designed to inform users' understanding of the technical constraints of wireless sensors affects participants' capacity to select a suitable course of action when they have no specialist knowledge in this domain. This chapter reports three iterations of a controlled user study (n=40) in which participants were tasked with finding suitable installation locations for smart wireless sensors. Through a combination of qualitative and quantitative methods, participant actions and responses are analysed. To further validate the findings, a follow-up field study (n=9) was conducted. This field study assesses the role of feedback in a less constrained and more ecologically valid setting. This chapter concludes by discussing four feedback strategies, reflecting on how they impacted users' ability to select a course of action.

Chapter 5 documents investigations of how users, with no specialist knowledge of pattern matching processes, can be informed through feedback and how this affects their ability to reason about expected outcomes and by proxy the consequences that are actioned. In a controlled lab study, 18 participants were asked to examine a visual form of feedback common to consumer smartphone applications. Through a combination of qualitative and quantitative methods this work reflects on how such visualisations can inform user understanding. In addition to reporting the findings, this chapter describes the limitations of the study design employed and how these might be addressed (work which is reported in the following chapter).

Chapter 6 builds on the work reported in Chapter 5, to better understand how feedback can help users with no specialist knowledge of pattern matching processes better understand which actions are most likely to result in a satisfactory outcome and also how feedback (and the improved understanding they gain from it) can help users overcome unexpected outcomes. To this end, this chapter reports on a series of user studies centred around a novel experimental design. In total, 70 participants were subject to one of seven conditions while creating stop motion animations with a bespoke smartphone or tablet application. Calling again on a combination of qualitative and quantitative methods, user actions and responses were analysed and discussed, focussing on the role of feedback and highlight a series of design implications.

Chapter 7 summarises the work presented in this thesis and reflects on how the research questions outlined in this chapter were addressed. It concludes with the key findings, acknowledgement of the limitations, and possible opportunities for further work in this space.

1.6 Research Contributions

This section outlines the key research contributions reported in this thesis.

- This thesis presents findings from seven user studies in which 145 participants were observed interacting with four different prototype smart systems (two of which have been made publically available for the research community to utilise in future work). Our findings demonstrate that feedback can be an effective means of informing user understanding such that they can take action when necessary, choose a satisfactory course of action when needed and overcome unexpected outcomes.
- This thesis validated existing feedback designs, whilst identifying a number of design implications for future smart system feedback design. For example, that for feedback to be effective, user expectations and the meaning being conveyed must be brought into alignment, otherwise misconceptions can lead to feedback being detrimental to user interaction.
- Finally, this thesis presents methodological contributions in the form of user studies designed to examine user interactions at the boundaries of success and failure.

In addition to the contributions itemised above, the work detailed in this thesis has been published (or is under review) at the following venues:

Chapter 6 is presented in the following CHI conference paper:

Jacob Kittley-Davies, Ahmed Alqaraawi, Rayoung Yang, Enrico Costanza, Alex Rogers, and Sebastian Stein. Evaluating the effect of feedback from different computer vision processing stages: A comparative lab study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 43:1–43:12, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2

The work detailed in Chapter 3 is presented in the following UbiComp workshop paper:

Jhim Kiel M. Verame, Jacob Kittley-Davies, Enrico Costanza, and Kirk Martinez. Designing natural language output for the iot. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 1584–1589, New York, NY, USA, 2016b. ACM. ISBN 978-1-4503-4462-3

Further to the published work above, I have contributed to other research related to wireless sensor-based systems and user interactions, which resulted in the following publications for which I am a co-author:

L. Bourikas, E. Costanza, S. Gauthier, P. A. B. James, J. Kittley-Davies, C. Ornaghi, A. Rogers, E. Saadatian, and Y. Huang. Camera-based window-opening estimation in a naturally ventilated office. *Building Research & Information*, 46(2):148–163, 2018

Carmine Ornaghi, Enrico Costanza, Jacob Kittley-Davies, Leonidas Bourikas, Victoria Aragon, and Patrick A.B. James. The effect of behavioural interventions on energy conservation in naturally ventilated offices. *Energy Economics*, 74:582 – 591, 2018. ISSN 0140-9883

Michael O. Jewell, Enrico Costanza, and Jacob Kittley-Davies. Connecting the things to the internet: An evaluation of four configuration strategies for wi-fi devices with minimal user interfaces. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 767–778, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3574-4

Chapter 2

Background

The work presented in this thesis touches on the boundaries of many research disciplines. As such, in this chapter, the most prominent literature related to user interaction with smart technologies across cognitive science, psychology and human-computer interaction are reviewed. This chapter is organised into the following sections:

- Section 2.1 - Explores how smart systems are defined in the literature and an details an evidenced definition for the context of this thesis.
- Section 2.2 - Describes the various approaches commonly used in HCI to model user understanding of interactive systems and justifies the applications of mental models to this work.
- Section 2.3 - Investigates the literature pertaining to system intelligibility and why it is important for the formation of useful understanding.
- Section 2.4 - Reviews prior work which relates to how users' understanding can be informed and the approaches tested in the literarture.

Concluding this chapter, in Section 2.5, we reflect on how the work presented in this thesis relates to the prior work discussed here.

2.1 Smart Systems

Over the last half-century, the term "smart" has firmly established itself in the lexicon of scientific literature and it has been affixed to a wide variety of devices, objects, substances and systems; such as homes Jensen et al. (2018), lighting (Park et al., 2016), textiles (Mikkonen and Townsend, 2019), materials (Vyas et al., 2012), toys (D'Hooge et al., 2000) and media equipment (Basu and Pentland, 2001) to name but a few. This work brings together technologies from a number of once disparate computing disciplines e.g. wireless sensor networks, context-aware computing and machine-learning. As technology in these domains has matured, the boundaries between them have diminished and applications which span them have transitioned from the theoretical to the practical. As new waves of technology emerge in this way, there is a need to differentiate them from their predecessors, both in research and in the public domain. In recent years, the term “smart” has become synonymous with the most recent advances, widely adopted by industry and research alike. However, despite its popularity, a definitive definition of what constitutes “smart” is noticeably absent from the literature (Silverio-Fernández et al., 2018). What characteristics make a system or device smart? In the remainder of this section, we discuss noteworthy definitions and taxonomies, before concluding with a definition, drawn from related work, for the purposes of this thesis.

2.1.1 Definitions

In the early 2000s, the European Commission funded a number of projects to explore how Weisner’s vision of ubiquitous computing (Weiser, 1991) could be realised. The “Disappearing Computer research initiative” (Streitz et al., 2005) was established to explore how technology can be integrated into everyday objects and environments, thus making them smart. When defining the key objectives for project proposals, they detail the following goals:

1. To create information artefacts based on new software and hardware architectures that are integrated into everyday objects.
2. To examine how collections of artefacts can act together, so as to produce new behaviour and new functionality.
3. To investigate designing collections of artefacts in everyday settings, and how to ensure that people’s experience in these new environments is coherent and engaging.

These goals highlight themes common throughout subsequent definitions of smart systems, namely; to embed computation and communication technologies, and support interactions which best suit the application rather than the constraints of technology. For example, Siegemund (2004) defines smart devices as those systems that can perceive their surroundings, and communicate with other systems to create contextually relevant behaviour.

Thompson (2005), in an article discussing the benefits and challenges of the new landscape being forged by smart technologies, extends this technological definition in two key ways. Firstly, Thompson clarifies that in addition to sensing their surroundings, smart systems may also affect it. Secondly, while the word smart is synonymous with intelligence, the capacity to learn is not a necessary characteristic of a smart device.

While these characteristics can be viewed from a purely technological perspective, as sensing technologies mature and the capacity to efficiently compute (both locally and on remote services) meaning from the information gathered increases, the constraints of the technology no longer need to dictate the way in which users interact. Instead, the application can take precedent, with the most coherent interaction modality at the heart of a smart system's design, i.e. what Siegemund (2004) refers to as contextually relevant behaviour.

In a 2014 article in the Harvard Business Review (Heppelmann et al., 2017), Michael Porter, a professor of business administration at Harvard, and James Heppelmann, at the time the CEO of computer software company PTC Inc., echo this sentiment. While they identify computation and communication as being at the core of a technological definition of what constitutes a smart product, they highlight that what makes a smart product fundamentally different is the changing nature of the "things".

Another key issue when defining a "smart" system, is demarcating the boundaries between systems. In the introduction to this thesis, we note that for the purposes of brevity we will refer to all smart devices and the interconnected constellations they form as smart systems. While the motivation is primarily to avoid repetition and improve consistency, the need to constrain the definition of a smart system is important. It is overly simplistic to define the boundaries based on the physical properties of the system, e.g. if a device is defined as a single physical object and a system as a collection of devices, then a contact sensor composed of two parts becomes a system and not a device.

The advent of the Internet of Things (IoT) has helped refine definitions of smart systems by limiting their scope i.e. by defining the broader architecture of IoT systems. Many of the definitions of IoT systems create boundaries based on responsibilities. Lopez et al. (2017) for example, in their work examining the privacy implications of IoT technologies, segment a simplified IoT architecture into three core, but non-trivial, components: smart things, backend servers and communications infrastructure (Figure 2.1). In this

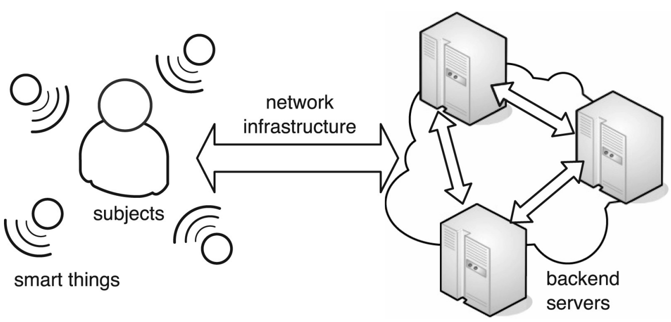


FIGURE 2.1: Simplified internet of things architecture (Lopez et al., 2017).

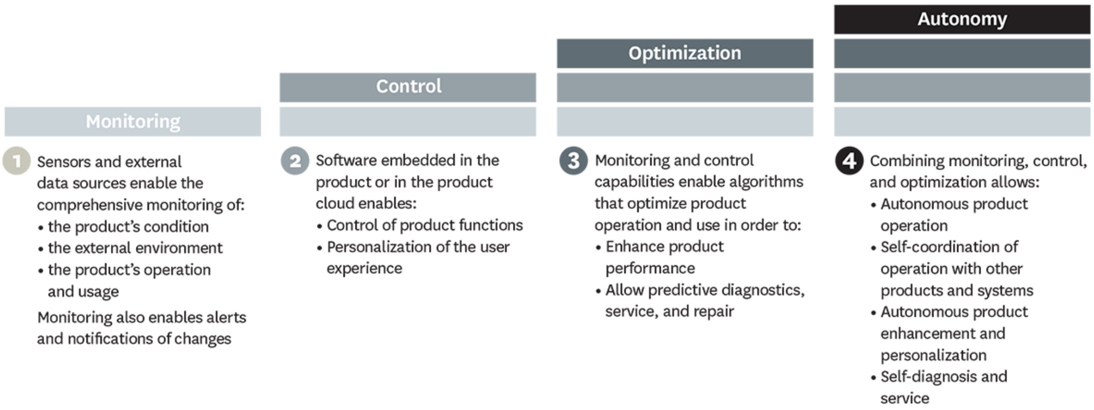


FIGURE 2.2: Proposed grouping of smart products based on their functionality and capabilities (Heppelmann et al., 2017).

definition, smart things can function and interact with users autonomously. While they may or may not yield value to users in this way, their connection to other devices permits enhanced functionality and extends the benefits to users.

2.1.2 Taxonomies

A number of taxonomies have also been proposed. Heppelmann et al. (2017) suggests that smart systems can be grouped into four (non-exclusive) categories, namely: monitoring, control, optimization, and autonomy (see Figure 2.2).

- Monitoring - Systems capable of monitoring their own condition and external environment using sensors and remote data sources. For example, a wireless humidistat which reports environmental health.
- Control - Systems whose behaviour and actions can be controlled through a remote connection of some kind. For example, a smart light bulb which can be remotely configured via a smartphone app.
- Optimization - Systems which combine monitoring and control to self-optimise. For example, a smart thermostat to regulate the indoor climate.
- Autonomy - Systems which go beyond optimization, supporting decision making and autonomous behaviour. For example, a robotic vacuum cleaner which can sense the layout of a room and plan how best to achieve its objective.

Streitz et al. (2005) distinguishes between two types of smart device: those which are system-oriented and those which are people-oriented. System-oriented devices function without the need for human intervention, although they are not exclusively isolated (e.g. a smart thermostat will monitor a building's occupancy and temperature to maintain

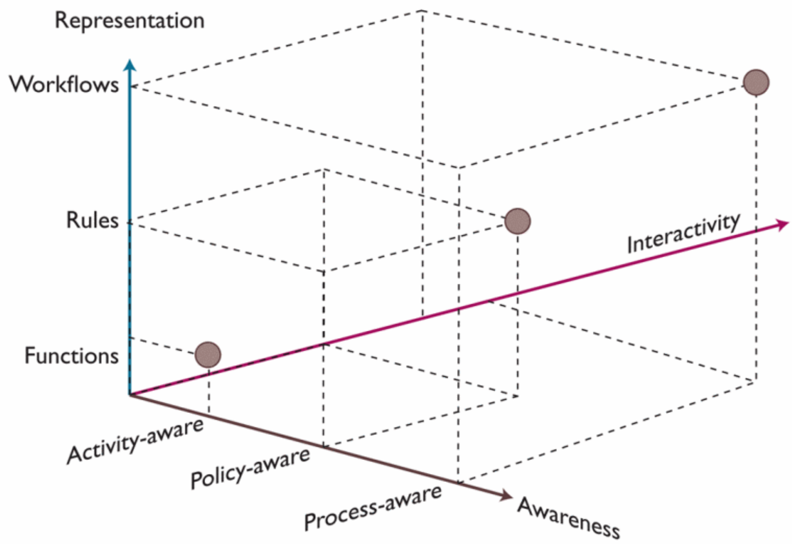


FIGURE 2.3: Smart-object dimensions (Kortuem et al., 2010). The three canonical object types: activity-aware, policy-aware, and process-aware.

a user-specified temperature). While people-oriented artefacts support users in making informed decisions and taking responsible actions (e.g. a system which suggests medical literature for a doctor's consideration, based on patient information).

The taxonomy proposed by Streitz et. al. and Porter and Heppelmann's taxonomy are inherently system-centric, focusing on the behaviours and capabilities of the system. In contrast, Kortuem et al. (2010) classifies industrial smart systems within a three dimensional space (Figure 2.3), the axes of which are:

- Awareness - the ability to make sense of, interpret and react to human activity in their environment.
- Interactivity - the capacity to interact directly or indirectly with users.
- Representation - Programming model i.e. the type of functionality they provide.

Using this space, Kortuem et al. (2010) proposes three classifications.

- Activity-aware objects which can sense their surroundings and record information, but provide little to no interactivity.
- Policy-aware objects which are able to assess how real-world events and activities comply with policies, providing feedback to users based on the application of rules.
- Process-aware objects which can sense real-world events and relate them to a workflow, permitting targeted guidance to be generated and issued to users.

2.1.3 Sensors and Smart Systems

One of the defining characteristics of smart systems, detailed in the definitions discussed above, is their capacity to sense their surroundings. While sensing technologies have been researched extensively for many decades, rapid integration into consumer products has opened many new opportunities for study. Smartphones are a prime example of consumer products increasingly being enhanced with sensors. Barometers, for example, were originally included in smartphones to improve GPS accuracy, but have been exploited by researchers to detect the opening of doors in order to estimate building occupancy (Wu et al., 2015). Similarly, smartphone microphones have been utilised to crowdsource information which otherwise would require an extensive network of bespoke devices. Zilli et al. (2013), for example, proposes a smartphone application which can identify the distinctive audible calls made by endangered animals.

Furthermore, the smartphone market has also contributed to the falling cost of sensor hardware and microprocessors which can be utilised in proprietary devices. Recently, Laput et al. (2017) demonstrated how such hardware in combination with machine learning can be used to great effect, creating a multi-purpose sensor capable of detecting and classifying activities in its local environment. This raises an interesting point. As sensors become higher fidelity and power-efficient microprocessors more capable of supporting processes to interpret the data they yield, users are increasingly being exposed to pattern matching and machine learning processes. While this combination of technologies presents great opportunities, it also raises an important question; what are the implications of poor user understanding? Thus a question which is particularly pertinent for applications such as participatory sensing and crowdsourcing. If a user does not have a good understanding of smart systems' underlying processes, how can they interact effectively? And how will they manage unexpected outcomes?

2.1.4 Summary

The term "smart" has become common parlance in scientific literature and in consumer marketplaces. While it has come to represent the next generation of computing technologies, there is as yet no consensus on a definition. Having reviewed a cross-section of definitions and taxonomies, below we define a smart system for the context of this thesis. In this thesis a "smart system" is considered to be:

- a device or collection of devices which collectivly can compute, communicate and sense and/or affect their environment in a manner befitting their application.
- a system which provides a discrete function which can inform other systems, be informed by other systems, and inform its own behaviour.

- either active, passive or a combination of the two. An active smart system is one which either requires, aids or demands user interaction, whereas a passive system is unobtrusive by nature and monitors its surroundings.
- a system that does not need to employ learning technologies, but may be perceived to have "intelligent" characteristics as a result of its capacity to act autonomously.

In this section, we also highlight the importance of investigating how user understanding of smart system processes impacts users' ability to interact with them. In the following section, we describe some of the most popular frameworks and theories used to do this, before in the subsequent chapter discussing existing work in this space.

2.2 Modelling User Understanding

Being a multidisciplinary field, Human-Computer Interaction has drawn theories and practices from a number of research disciplines, such as computer science, cognitive psychology, sociology, ergonomics and industrial design, to describe, analyse, predict and explain users' interactions with technology (Carroll, 2003). In this section, we discuss the most common methodologies and frameworks utilised in the HCI literature and justify those used in this thesis.

2.2.1 Mental Models

Mental models are one of the most popular theoretical constructs used in HCI research (Moray, 1999; Payne, 2007; Rogers et al., 2011). The concept of a mental model was first proposed by cognitive psychologist Craik (1943), who postulated that people hold small-scale models of how the world works in their minds and use them to reason about interaction. In 1983, two independent publications both titled "Mental Models" were published (Johnson-Laird (1983) and Gentner and Stevens (1983)). Despite being released in the same year, the two publications viewed mental models through very different lenses (Sasse, 1997).

Johnson-Laird (1983) considers mental models as one component of a much larger framework which attempts to capture fully how human beings internally represent real-world phenomena. He describes mental models as encapsulating all the exposed concepts, interrelationships and mappings between concepts and a target domain (i.e. the task which the system is designed to support). These internal representations can then be drawn upon to reason by way of thought experiments (Jones et al., 2011).

Gentner and Stevens (1983) surveyed literature examining user understanding from a range of disciplines and identified a common theme: analogical reasoning. The assertion is that mental models are formed through analogical thinking: when a new phenomenon is encountered, a person draws on existing understandings of phenomena which they perceive to be similar, mapping inferences when reasoning about how to interact (for example, explaining electrical current in a circuit by means of water flowing through pipes). The distinction here is that Johnson-Laird focuses on the structural analogy between a phenomenon and it's internalised representation, whereas Gentner & Stevens consider the analogy of one model to another (Sasse, 1997).

The work of Norman (1983) was instrumental in the adoption of mental models for human-computer interaction and his book "The Design of Everyday Things" (now revised Norman (2013)) first popularised the ideas in the domain of HCI. He defines mental models as internal representations which allow users to explain and predict the actions of a system so that they can reason about interactions before committing to an action. He also described the following core attributes of mental models:

1. Mental models are often incomplete and can only encapsulate those aspects of a system with which the user has experience.
2. People's ability to "run" their mental models are severely limited and can be impacted by poor working memory and existing beliefs and biases.
3. Mental models are unstable and the details will be lost with time.
4. Mental models have no firm boundaries - the edges between similar systems can blur.
5. Mental models are unscientific, people maintain "superstitious" behaviour patterns even when they know they are not needed.
6. Mental models are parsimonious -people often do extra physical operations rather than do the mental planning required to avoid those unnecessary actions.
7. Mental models need only be useful for a user and thus their completeness is irrelevant so long as they remain useful.

Further to this, Rouse and Morris (1986) highlight a key difference between a mental model and knowledge, namely that a mental model has a structural element which is analogous to the target system. In contrast knowledge of the system is linked isomorphically to the model (Moray, 1999).

2.2.2 Conceptual Models

Conceptual models can be difficult to understand (Rogers et al., 2011). To avoid confusion here we outline a definition adapted from Johnson and Henderson (2002):

A conceptual model is a high-level description of how a system is organized and operates. It describes: (i) the major design metaphors employed (if any), (ii) the user-facing concepts, their attributes, and available operations, (iii) the relationships between these concepts, and (iv) the mappings between the concepts and environment.

While mental models and conceptual models describe the idea of a system model, it is helpful to discriminate between the model used to inform the design of a system (i.e. the conceptual model) and the model which users of the system develop through their interactions with it (the mental model). For example, not all systems will have a conceptual model behind their design. Despite this, users will form mental models based on their experience of interacting with it and past experience of other systems. So long as a user mental model aligns with the way the system operates, it will prove useful and one of the key tools to ensure users develop such coherent mental models is a design based on a good conceptual model (Norman, 2013).

2.2.3 Seven Stages of Action

The "seven stages of action" model (Norman, 2013) is closely related to mental models and has been cited extensively in the HCI literature. The model deconstructs a user's interactions with a system into a series of stages, with the intention of facilitating a structured analysis of how a user mental model is informed. Figure 2.4 illustrates the various stages and the relationships between them. The stages are divided into two parts; the stages of execution, which maps the path from goal formation to the action being enacted, and the stages of evaluation, where the user reasons about the consequences of the action and compares the outcome with their original goal.

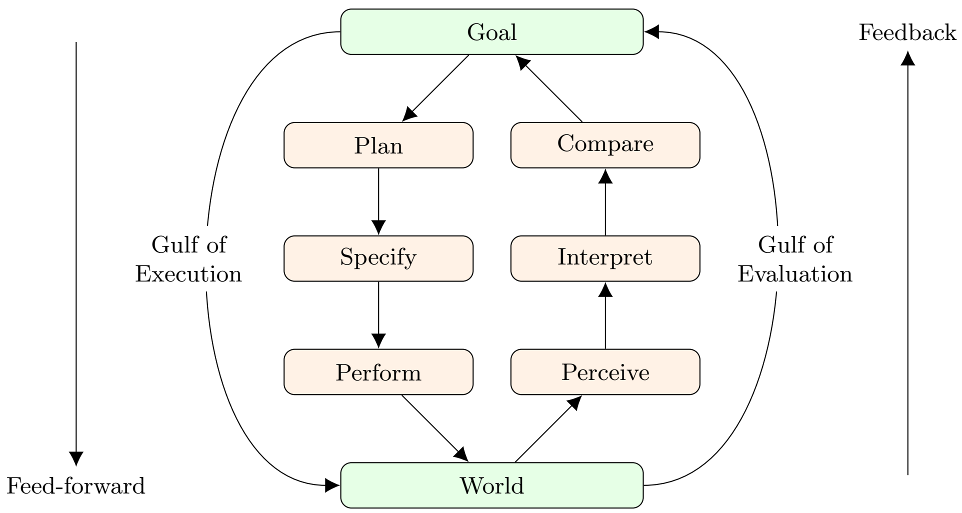


FIGURE 2.4: The seven stages of action cycle (adapted from The Design of Everyday Things - Norman (2013)). Highlighting the gulf of execution, the gulf of evaluation, and the information flows: feedback and feed-forward

The stages are as follows:

Stages of execution: Establish a goal → Plan the action → Specify an action sequence → Perform the action sequence.

Stages of evaluation: Perceive the system state → Interpret the perception → Compare the outcome with the goal.

Although an approximation (e.g. a user may not consciously enact all stages or start with a goal in mind), Norman's model provides a useful reference for analyzing user interactions because it separates the cognitive processes and highlights the links between them that must be supported (Lewis, 1998).

To represent the challenges faced by designers, Norman (2013) associates the stages of execution and the stages of evaluation with a "gulf", highlighting that it is the system designer's job to bridge these gulfs. The "gulf of execution" can be addressed by informing users about what the system can do (discoverability). Making actions discoverable can be achieved through a combination of design and instruction. For example, a flat metal plate on a door affords pushing, while the embossed label "push" signifies the action is appropriate and the absence of a handle that the user can grip constrains the user. Overcoming the "gulf of evaluation" requires the system to feedback information to the user about the outcome of their action. While it is reasonable to think of this solely in the immediate term (e.g. the system acknowledges a button being pressed with an audible beep), the ramifications of feedback are much wider i.e. when coupled with a coherent conceptual model, feedback can be used to foster a useful mental model. In this thesis, we explore how this feedback builds user understanding in the absence of other guidance and how this compares with more explicit instruction.

2.2.4 Summary

In this section, we describe the core HCI frameworks and theories employed in this thesis. Metal models were deemed the most appropriate model for analysis and discussion for much of the work reported in this thesis. However, acknowledging the criticisms of Carroll (2003) and Rogers (2012), who identify mental models misappropriation in some of the HCI literature, we draw on their original definitions to clarify what they represent in the context of this thesis.

In addition, when discussing the seven stages of action model, we highlight the challenges faced by system designers and the opportunity to inform users understanding through feedback. An opportunity which this thesis examines is how feedback can be used to inform users' mental models in the absence of other forms of guidance.

2.3 The Importance of System Intelligibility

The importance of users understanding of computer system operation has received considerable attention from the fields of HCI and psychology. In this section, we discuss the literature in this space.

2.3.1 Effective and Satisfactory Interaction

Over the long arc of research exploring user interaction with computing systems, supporting effective and efficient interactions has, understandably, been at the forefront of much of the literature.

In the 1980s, when examining the challenges of educating novice software developers, du Boulay et al. (1981) comprehensively review prior work which investigates the role of conceptual models in making systems easier to understand and thus impacts users ability to interact effectively. They conclude that novice programmers should be initially exposed to abstract representations of systems to expedite their education. Similarly, Bayman and Mayer (1984) conducted a series of experiments investigating how exposing a conceptual model of a system affected participants' understanding. In this work, they tasked participants with solving a series of mathematical problems using a calculator. Participants were exposed to one of three variations of the calculators UI, each representing the system's inner processes in a different way. Their findings illustrate the breadth of users' interpretations of a system and that conceptual models can greatly benefit the formation of coherent mental models, resulting in improved performance. At a similar time, Kieras and Bovair (1984) investigated whether users with a better understanding of an unfamiliar control panel would be more effective at operating it. Through a series of three experiments, they find how users exposed to a conceptual model of the system were able to learn procedures more quickly, retain them more easily, execute them more quickly and were more able to identify unnecessary steps in deliberately verbose procedures.

More recently, Patel et al. (2008) developed and executed a series of think-aloud studies in which participants with some machine learning experience were tasked with creating an image-based classifier of handwritten digits. In their findings, they highlight that despite the participant pool having a good foundation in software development, the overestimation of statistical machine learning's capabilities impacted participants' ability to successfully implement the technologies correctly. They conclude that the exposure of models, features and the relationships between them will better align user expectations with a system's true capabilities. Similarly, studies have shown that users with more coherent mental models are more likely to make systems operate to their satisfaction. Kulesza et al. (2012) reports an empirical between-groups study of users interacting with a recommendation-engine driven personal radio station, i.e. an application which adjusts the music played based on user preferences. Over the course of five days, they compare users left to form their own understanding of the system's operation with a group who were given a conceptual model via training. They refer to these groups as "with-scaffolding" and "without-scaffolding" respectively. While the two groups received identical training on how to use the app, the with-scaffolding group received training designed to induce a structural mental model of the recommendation engine. They report that participants who were presented with structural knowledge (i.e. those in the "with-scaffolding" group) viewed interacting with the recommendation system in a positive light, while participants of the "without-scaffolding" group more frequently described their experience negatively as "confusing" or "complex". Users with more coherent mental models were also significantly more likely to increase their self-efficacy, which is known to correlate with reduced computer anxiety when tackling complex computer tasks.

This work suggests that user's understanding can have a profound effect on effective and satisfactory interactions. Literature examining these findings relating to smart systems specifically, however, is less mature and future work is needed in this space. Furthermore, there are other factors to consider when discussing systems with agency e.g. trust. In the next section, we explore the role of trust and its relationship to intelligibility.

2.3.2 User Trust

As smart systems become more prevalent, the black-box nature of their underlying processes has led to the topics of trust and accountability gaining considerable attention in the media and research community.

Research investigating user interactions with smart thermostats in their own homes (Yang and Newman, 2013), reports observations which demonstrate how poor understanding of system operation led to unrealistic expectations, which when violated negatively impacted participants trust of the system. This propensity to overestimation abilities was also demonstrated in a longitudinal study of users of a smart energy billing system designed to foster more efficient energy consumption (Alan et al., 2016a), resulting in waste.

In an online study of a peer-assessment system, Kizilcec (2016) reports similar findings of distrust. Participants of the study were tasked with submitting an essay to an online system and grading a subset of their cohorts essays in exchange. The online system then collected the graded essays and derived a grade using an algorithm proven to effectively remove bias. Participants were randomly divided into three groups, with each group receiving a differing degree of transparency relating to the grading process. In the lowest transparency group participants were shown a single sentence describing the approach used to derive their grade “Your computed grade is X, which is the grade you received from your peers”. In the medium transparency group, this sentence was appended with “and adjusted for their bias and accuracy in grading. The accuracy and bias are estimated using a statistical procedure that employs an expectation maximization algorithm with a prior for class grades. This adjusts your grade for easy/harsh graders and grader proficiency”. In the high transparency condition, the explanation was supplemented with the original grades assigned by each peer and how these were adjusted to arrive at their final grade. The findings of this work show how individuals who received lower grades than expected demonstrated less trust in the system. However, this can be reversed by making the system more transparent. This is a finding which is in line with procedural justice theory (Tyler, 1989), which postulates that an unsatisfactory outcome can be judged to be satisfactory so long as there is sufficient evidence for it to be considered just. Interestingly, Kizilcec (2016) also reports that providing too much information eroded this trust and that while providing aggregate information and an explanation of the process to derive a grade was beneficial, exposing the raw grade information was detrimental.

In contrast, Dzindolet et al. (2003) and Yang et al. (2014) document the perils of users trust in automated systems - how over-reliance can lead to users becoming less vigilant to unexpected outcomes and system failures. Dzindolet et al. (2003) reports a series of three user studies in which participants were asked to classify 200 images based on the presents (or not) of a soldier in varying degrees of camouflage (in half a soldier was present and in the other half a soldier was not). After examining each image, participants were asked to indicate if they felt a soldier was present and, using a Likert scale, report how confident they were in their decision. Participants were also introduced to a mock computer program described as an automated assistant which is capable, although not perfect, of identifying soldiers in images such as these and that it would concurrently analyse the images as they classified them. Across the three user studies they found that, after a short exposure to the automated assistant demonstrating accuracy through simple examples, participants were significantly more likely to rate their trust in the system highly. After sustained interactions, however, even automated assistants that make half as many errors as the participant were considered less trustworthy than average. However, providing a reason why the automated assistant might err, reversed this effect and trust increased.

In a later qualitative study comparing user interactions with conventional thermostats with a smart thermostat, Yang et al. (2014) document how some users trusted the smart thermostat to automatically save them energy and as such rarely engaged with it. Furthermore, of those who did identify the need to assist the thermostat, many lacked motivation and commonly reported that their trust in the system had been negatively affected.

This work highlights the risk of short positive experiences and how preconceived notions of how a smart system will operate can jeopardize user interactions with such systems. It would appear then that addressing users preconceived ideas is of the utmost importance, in order to align users' expectations with the realities of these systems. However, there are other opportunities to address misconceptions, namely when they encounter unexpected outcomes. In the next section, we discuss the work in this space.

2.3.3 Mistakes and Misunderstandings

Flawed mental models have been shown to result in confusion and erroneous interactions (Kulesza et al., 2015; Tullio et al., 2007). There are many factors which can lead to such misunderstandings, some as simple as the misinterpretation of terminology (Alan et al., 2016b) and others as complex as subconscious bias. A recent study of floor cleaning robots (Garcia et al., 2016) describes how participants rated a floor cleaning robot's performance more favourably when they had witnessed it in motion.

As we note at the end of the previous section, correcting these biases and misunderstandings is an important consideration for designers. du Boulay et al. (1981) highlight the importance of feedback in this regard, stating that "error messages are a crucial part of the commentary and they form an important window into the machine". More recently, Kulesza et al. (2015) work with "Explanatory Debugging", an approach to help users effectively and efficiently personalize machine learning systems. They showed that bidirectional explanations (i.e. allowing the system to explain its predictions and allow the user to explain corrections to the system) increased participants' understanding of the learning system by 52% and allowed participants to correct mistakes up to twice as efficiently as participants using a traditional learning system.

The work above discusses systems, where the complications to be overcome, are internal i.e. born out of the processes they employ. Systems, such as wireless sensor networks, add an extra dimension. These wireless devices also expose users to complications arising from the systems' interaction with the environment, that of radio propagation. Installing a sensor, for example, requires the movement of a smart device through a physical space, which has implications for the sensor, the underlying processes and the systems' connectivity.

Beckmann et al. (2004) report a field study examining how end-users install a set of mock sensors (electricity consumption, motion, vibration, cameras and microphones) in their own homes. They observe a variety of mistakes, often caused by end-users' uncertainty of the sensor's needs. Among the recommendations they propose for system designers, they highlight the need for sensor systems to detect, alert and guide users such that they can be corrected. Similarly, Hu et al. (2016) examined the viability of non-technical users installing a sensor system which was originally designed for expert installation. Assessments of user installations of the system reveal numerous mistakes made by users, the majority of which are related to connectivity. Approximately 14% of the mistakes made were caused by radio repeater units not being installed correctly. Without connectivity, feedback participants were not able to identify this issue and assumed they had completed the installation successfully.

This work illustrates the potential for users to make errors. However, the threshold for success is binary i.e. can the device communicate and can the device sense a target. Smart systems commonly employ sophisticated processes which are dependent on the quality of the sensors' output (i.e. sensor fidelity). An issue the literature fails to address is how this work relates to quality beyond the baseline of success.

2.3.4 Summary

The evidence supporting the need to inform users of underlying processes is compelling. However, as Yang and Newman (2013) point out, in their work with smart thermostat users, the majority of work in this domain has focussed on providing explanations for how a given system works and why it behaves the way it does, approaches which assume that users are interested in understanding the system and willing to invest time to learn. Their observations of Nest thermostat[1] users indicate that the desire to understand the system arises infrequently (only when something goes wrong) and that their motivation to learn independently is lacking. Yang and Newman conclude that there is, therefore, a need to find means to inform users understanding without explicit instruction. In the next section, we discuss existing work in this space.

2.4 Informing User Understanding

While the vision of ubiquitous computing describes, what we might now consider to be smart, technology weaving its way into the fabric of everyday life, this is not to say that users of the technology will not need to develop an understanding of how it works, only that computational aspects of technology will be indistinguishable from the physical. Having explored in the last section the reasons why it is important to make systems understandable, in this section, we discuss the literature which has investigated the way in which user understanding can be fostered.

2.4.1 Exposing System Processes

In some of the earliest work in this space, du Boulay et al. (1981) when reviewing existing literature relating to the education of novice software developers, develops a notion of "commentary". They describe commentary as the exposure of a systems conceptual

[1]More information relating to Nest thermostats can be found at: https://nest.com/

model delivered through graphical and textual feedback in the user interface. Drawing on prior work, they evidence their hypothesis that exposing abstract representations of a computer system will encourage better understanding.

Shortly after, Bayman and Mayer (1984) experimentally tested this idea, investigating how exposing the inner workings of a calculator impacted users' (with no former programming experience) capacity to form useful mental models. In a between-groups study, users were tasked with solving a series of mathematical problems. During the study, participants were subject to one of three conditions, each visualising how the calculator's underlying processes operate in a different way. They report that despite participants having very similar experiences of the calculator, i.e. solving the same problems with the same input mechanism, the mental models' participants developed differed considerably, remarking: "Hands-on experience does not guarantee that the user will understand what he or she is doing". Furthermore, participants who were exposed to visualisations were significantly more likely to employ appropriate strategies, noting that "mental models can be explicitly taught to users of electronic computing devices".

Similar to Bayman and Mayer (1984), Kieras and Bovair (1984) conducted a series of experiments to test the hypothesis that explicitly providing users with a conceptual model of the system would positively affect their ability to interact effectively. To this end, they designed a series of experiments centred around the operation of a bespoke control panel. In the first of three experiments, participants were tasked with learning set procedures, which detailed the order in which various dials and buttons should be operated to make an indicator lamp on the control panel flash. After a training period, they were tested on their ability to recall the procedure. This test was also repeated one week later. While all participants received basic training on how to functionally operate the simulated control panel, half received a written conceptual model of the system and a schematic diagram outlining how the controls were connected. Their findings indicate that participants who received an explanation, in the form of a conceptual model, were significantly more able to learn and recall the procedures. In two follow-up studies (iterations of the first), Kieras and Bovair refine their study and conclude that exposing users to the inner workings of a system can enhance their ability to infer appropriate interactions and that this information does not need to be complete in order for it to be useful. However, they identify that any representation of a system is vulnerable to misinterpretation and great care is needed.

While these studies appear to demonstrate the importance of exposing system processes to users, the systems tested are relatively easy to conceptualise. The advent of smart systems has led to an increase in the integration of pattern matching and learning technologies. These processes can be considerably more challenging to explain (Dove et al., 2017).

The creators of Gestalt, (Patel et al., 2010a), an integrated development environment

(IDE) designed to assist programmers in creating software which makes use of machine learning technologies, demonstrate that exposing data at various stages of a process significantly improves programmers' ability to identify and correct errors in their code. Similarly, Prospector (Krause et al., 2016), which facilitates the probing of the predictive models by data scientists, has shown that exposing processes can help users understand how features affect the outcome. They report that by allowing users to adjust input variables and see-through visualisations to how the model responds, helped users gain deeper insights into how the model worked and thus more better able to interact more effectively.

In addition to machine learning, the increased availability of powerful mobile processors and high-resolution cameras has resulted in the rise of camera-based interaction modalities (e.g. intelligent filters used for entertainment in social media). While there is no work specifically examining user interaction with smart systems which use these technologies, there has been work for other application spaces. Kato et al. (2012) developed a system designed to expose domain-expert programmers to computer vision technologies, with the ambition of aiding code debugging. The system allows images passing through the various stages of processing to be inspected and an interactive timeline interface lets users record and examine data flow temporarily. Although a small user study was conducted, the focus was system functionality rather than the user experience.

This work, while alluding to the benefits of exposing system processes to users, needs much more robust examinations before any meaningful conclusions can be drawn. The most extensive investigation in this space to date was conducted with context-aware systems. Lim et al. (2009) explore how the messages yielded by such systems impact their intelligibility. Through a series of user studies, with over two hundred online participants, they demonstrate the benefits of making the motivations behind automated decisions more salient to users. Furthermore, they show how explanations of why system behaviour occurred, result in better user understanding in comparison to explanations of why the alternative behaviours did not. This is an observation supported in a later study (Koo et al., 2015) of simulated driver assistants. Koo et. al. demonstrate how the content of verbalized messages, announced to explain the actions of an autonomous braking system, can affect drivers' attitude and safety performance. They find that messages describing why an action was deemed necessary led to better driving performance, concluding that designers need to carefully consider not only the design of autonomous actions but also the appropriate way to explain these actions.

While this work demonstrates the benefits of exposing users to systems' inner workings, via a conceptual model, more exhaustive empirical evaluations are needed to better understand how to maximise their design. In particular, smart systems commonly employ multiple complex processes (e.g. pattern matching, wireless radios and sophisticated sensors) which collectively can obscure the relationship between inputs and outcomes. How then can users be informed of the relationships between their actions and systems outcomes?

In contrast to work investigating the effects of exposing system processes to users, there are a few examples of systems which have attempted to improve interaction by deliberately hiding system complexity. In the next section, we discuss this work.

2.4.2 Hiding System Processes

Designed to aid software developers with no specific knowledge of pattern matching and image processing technologies, Crayons (Fails and Olsen, 2003) is a tool for creating new camera-based interfaces. The tool they describe permits software developers to build classifiers using a painting metaphor, effectively drawing on images to highlight the regions of interest and abstracting away the complex process of feature selection. Similarly, the developers of Eyepatch (Maynes-Aminzade et al., 2007) set out to build and evaluate a prototype system which would permit novice programmers to utilise complex computer vision processes (e.g. optical flow and feature extraction) to inform other rapid prototyping tools, such as Adobe Flash. While both user evaluations demonstrated the system's merits, the small sample size and absence of rigorous analysis prevents any strong conclusions from being drawn. This work does, however, highlight a skills gap between the interface designers and experts in the underlying processes being used, a factor which Dove et al. (2017) have more recently reported in relation to machine-learning technologies in general, asserting that one of the challenges facing systems which employ machine learning technologies is the lack of domain-specific knowledge in the field of interface design.

There is, therefore, a need to explore how abstracting away the complexities of systems impacts users' ability to interact effectively and develop useful mental models. Designing conceptual models which strike a balance between providing sufficient detail so that users can reason about interactions and ensuring there is not a need to become an expert is likely to be non-trivial. Therefore, more robust evaluations in this space are needed to inform smart system design.

2.4.3 Informing through Feedback

Earlier in this chapter, we refer to the notion of "commentary" (du Boulay et al., 1981) which describes informing user understanding thought text and graphics presented in the UI. This approach is suggestive of the capacity to educate user understanding through interaction. In this section, we discuss literature which examines the viability of using feedback in this way.

Costanza et al. (2010) explore the role of feedback in users' search for reliable connectivity when setting up a multi-hop network of wireless nodes. They evaluate an audible interface designed to provide connectivity feedback (analogous to the tuning of a radio), comparing it with a traditional GUI in a controlled lab study. While they demonstrate the effectiveness of this alternative modality, they also discuss a number of search strategies exhibited by participants, which were shaped by users' understanding, namely: (i) aggressive - where participants first test the limits of connectivity before refining potential node installation locations, and (ii) progressive - where participants stop as soon as they notice a small degradation in connectivity.

Zhao et al. (2016) conducted a study examining user interactions with an augmented reality pattern recognition system called CueSee. This bespoke application was designed to assist users with low vision when searching for products in a shopping environment. Using augmented reality (AR) technologies, the application highlights regions of interest (i.e. products) via a head-mounted display. In addition to demonstrating the effectiveness of AR in the application, they also explore the design of this feedback, reporting how feedback, if badly designed, can frustrate and annoy users.

In a user study examining a non-invasive sensing method for electricity consumption, Patel et al. (2010b) describe sensor feedback as critical in helping participants to install the sensor. Patel et al.'s study was conducted in participants' own homes and tasked them with placing a sensor node on the outer casing of their home's consumer unit (a.k.a. fuse box or breaker board). For the sensor to work effectively it must be positioned immediately over a pair of terminals hidden behind the outer case. To guide participants, two lamps indicate when the sensor is correctly located. They report that all participants were able to install the sensor correctly and that the indicator lamps' feedback was critical in helping users achieve this. However, they also report that a small lag between the users' physical actions and the resulting feedback was challenging to participants.

While this research demonstrates the capacity for feedback to assist user interactions and improve efficiency, it does not directly investigate the relationship between the effects of feedback and users' understanding of systems. Furthermore, this work highlights the challenges of designing feedback of this kind, illustrating that it can be detrimental if not well designed. Therefore there is also a need to better understand the many characteristics of feedback design and their implications.

2.4.4 Summary

There has been extensive work investigating how user understanding can be informed and encouraged. In this section, we have discussed the most prominent literature pertaining to exposing system processes, explaining system behaviour and actions, the role of feedback to educate and concealing complexity. Furthermore, we highlight a number of opportunities to improve our understanding of smart systems.

2.5 Chapter Summary

This chapter has reviewed the application of mental models as a theoretical construct to embody user understanding in the HCI literature. Evidencing that deeper user understanding of heuristic systems can enhance efficiency, trust and satisfaction, while also inversely demonstrating that the formation of unsound mental models can be detrimental to user interactions. The work reported in this chapter also highlights a number of opportunities to improve the scientific literature in this space.

Firstly, while research has shown that exposing conceptual models of system processes can be effective for fostering understanding, a balance needs to be struck between being sufficiently informative and overwhelming users, expecting them to become expert. More research is needed to understand how guidance can be shaped to have the greatest effect.

Secondly, as Yang and Newman (2013) point out, it cannot be assumed that users of smart systems are willing to invest time in educating themselves. Therefore, there is a need to find means to inform users' understanding without explicit instruction. Feedback has been shown to be an effective mechanism in research exploring the individual technologies common to smart systems (e.g. wireless connectivity). However, smart systems often bring together many different technologies which can compound the challenges faced by users when reasoning about system behaviour. Therefore, more work is needed to explore the relationship between feedback and user understanding, specifically with systems which link together complex processes prior to yielding outcomes.

Finally, research has shown how poorly designed feedback can be detrimental to interactions. However, the literature does not explicitly examine how feedback relates to the fostering of useful mental models. Understanding the implications of feedback design will, therefore, better inform future work in this domain.

Chapter 3

Natural Language Feedback to Inform Users of Smart Systems

While some smart systems operate autonomously, quietly working away in the background (e.g. gathering temperature data to inform a climate control system), other systems actively require human intervention e.g. alerting a homeowner that motion has been detected by their smart security camera. As smart devices become increasingly common, the number of notifications being presented to users is also rising. The HCI community has previously highlighted the perils of bombarding users of smartphones with push notifications and proposed numerous mechanisms to manage user workload (e.g. Mehrotra et al. (2016); Pradhan et al. (2017); Auda et al. (2018)). However, this work assumes (justifiably for this application) that not all notifications are of equal importance, something which cannot be guaranteed in other application spaces.

In the manufacturing sector for example, smart systems are being utilised to identify and alert maintenance crews to the early warning signs of production line equipment failure (Sezer et al., 2018) with the ambition of addressing mechanical failures before they interrupt production. For these systems to function effectively, all notifications must be read and acted upon appropriately. The challenge then is to design notifications which encode sufficient information that action is taken when necessary and wasted time and effort is avoided.

Smart system notifications are commonly delivered through natural language messages. While other modalities such as audible tones, haptic actuators and more novel approaches (e.g. smart home lighting (Komninos et al., 2018)) have been considered, natural language textual messages remain prevalent as the information carrier i.e. while other modalities are used to gain the users' attention, the message is delivered in textual form. The question then, is how to author messages such that there is a balance between brevity (i.e. they can be consumed quickly) and conveying sufficient information that reasoned judgements can be made about any action executed in response.

In this chapter, we present an exploratory investigation[1] in this space. In a controlled between-group lab study we evaluate how users' perception of the need to take action is affected by the information richness (i.e. the amount of detail) encoded into textual feedback. We begin by describing the motivation for this work, then detail the study design and experimental apparatus, before reporting and discussing the findings and design implications.

3.1 Motivation

A number of studies have looked at providing explanations to users of how improving system intelligibility can improve user understanding (e.g. Buchanan and Shortliffe (1984); Herlocker et al. (2000); Lim and Dey (2011); Kulesza et al. (2013)). Lim et al. (2009) work with autonomous decision making systems, for example, has shown that explaining *why* a system behaved in a certain way helped users understand the systems "reasoning" and consequently they were more accepting of its choices. Similarly, Herlocker et al. (2000)'s study of how explanations can improve the acceptance of automated collaborative filtering systems, suggest that participants value having the explanations. While these studies demonstrate the value of providing information-rich messages, their work focuses on explaining decisions which have already been made. In contrast, this work tests how feedback can be used to inform a user's decision to take action or not. Researchers have also explored the effects of different levels of system transparency on user's trust (e.g. Cramer et al. (2008); Helldin et al. (2014)). In an online experiment (Kizilcec, 2016) of a peer assessment task (i.e. marking others' academic work), transparency (i.e. explanations of decisions) has been shown to impede the erosion of trust, but providing too much information can reverse this effect. This work highlights the challenges of designing feedback and the importance of providing the correct level of information.

We identify two research opportunities for natural language feedback; (i) how applicable is existing research to smart systems and (ii) how can it be used to convey complex information succinctly, such that the recipient's response is appropriate, i.e. invoke the appropriate action.

3.2 Study Design

To better understand what level of detail is required in a natural language message system, so that message recipients (i.e. the users) can make an informed decision about whether to take action or not, we designed and conducted the following study.

[1]Ethics approval granted by the University of Southampton (ref: 18425)

3.2.1 Experimental Task

Eight messages were generated pertaining to a subject matter which was familiar to all participants, namely server maintenance. This topic was selected as it was deemed sufficiently complex to yield queries for which a human actor could realistically be required to decide on whether or not to take action, e.g. taking inspiration and building on Li et al. (2016)'s work with security notifications, highlighting issues which might indicate a security threat.

3.2.2 Conditions

Three textual variations were formed for each of the generated messages, with each variation providing more detailed information than the last. The first variation, "Format A", represents the most information-poor message format, containing a very high-level summary of the notification (examples below). The information was limited to one of three predefined categories: User behaviour, System Performance and System Failure. The second message format, "Format B" increases the amount of encoded information, supplementing the information contained within Format A with a high-level summary of the insights which triggered the message's creation. The final message format, "Format C", expands on Format B, detailing all of the triggers which invoked the messages creations. An example of the format can be seen below:

- Format A: "There is a risk of system failure on server *beta*."
- Format B: "There is a risk of system failure on server *beta* because 1 core metric: temperature, is higher than average."
- Format C: "There is a risk of system failure on server *beta* because the temperature is 50°C and the average temperature is 9°C."

3.2.3 Message Generation

In total 8 messages were generated, four of which represented situations which required users to take action, while the remaining four were "false alarms". Modifying Format C's example above, a false alarm would look like: "There is a risk of system failure on server *beta* because the temperature is 10°C and the average temperature is 9°C".

3.2.4 Participants

A total of 8 participants (1 female, 7 male) took part in the study. All were members of our University: 7 PhD students and one research assistant. While they represent a broad range of backgrounds and worked in a variety of computing disciplines, all participants had experience of working with networking, server maintenance or system troubleshooting. The study took up to ten minutes to complete and participants were entered into a lottery pool where they could win a £20 shopping voucher as remuneration for their participation.

3.2.5 Procedure

All studies were conducted in the same empty meeting room on a university campus. Two experimenters were present at all times - one to conduct the experiment and the other to observe, take notes and make audio recordings. Participants were shown all messages and their variants on a laptop 15-inch computer. At the beginning of each experiment, participants were given detailed instructions of what was expected of them and how to proceed with the task. The written instructions described a scenario in which they play the role of a system administrator. In this role, they were expected to monitor the operation of a networked computer system and respond when necessary to notifications yielded by an autonomous system designed to identify potential technical issues. During the study, participants were shown all 8 messages in all the formats. After each message, participants were asked how they would respond (i.e. whether they would take action or not). For clarity, they were also asked how they interpreted the message i.e. what they thought it meant.

3.3 Quantitative Findings

The chart presented in Figure 3.1 details the number of messages to which participants reported that they would take action when an action was required (the true positives). For clarity, 4 messages required action was shown to 8 participants, thus there were 32 instances. The figure shows that participants more frequently reported a need for action when shown messages in *Format C*, fewer for *Format A* and the least for *Format C*. However, there is no significant difference across conditions[2]. Conversely, Figure 3.2 reports the number of messages to which participants reported that they would take action when **no** action was required (the false positives). Responses to *Format C* messages most often resulted in unnecessary action being avoided, with *Format C* only slightly improving on *Format A*. A Kruskal-Wallis test revealed a significant difference between conditions (p=.04135, H = 6.3711).

[2] A Kruskal-Wallis Test no revealed a significant difference between conditions (p=.15159, H=3.7732)

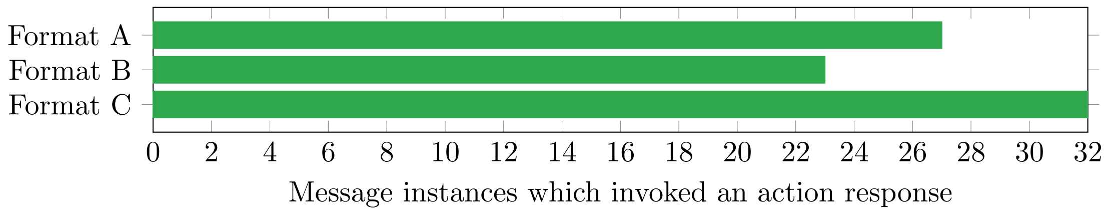


FIGURE 3.1: Responses of action when action was required (true positive)

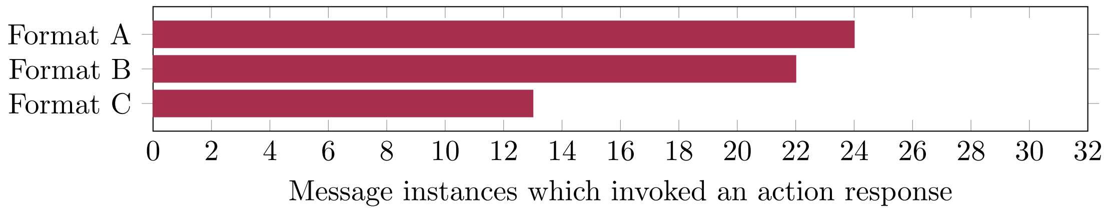


FIGURE 3.2: Responses of action when **no** action required (false positive)

3.4 Qualitative Findings

The vast majority of participants (7 in total) described *Format A* messages as containing too little information for a considered response to be taken e.g. "it is not very detailed" (user 3) and that "it's fairly vague" (user 4). While two largely chose not to act, six participants said that the only reasonable outcome of receiving such a message would be to "try to find out what the risk was" (P5) and "investigate what kind of user behaviour it is" (P1) by for example "checking the user logs" (P2).

One participant (P7) described all notifications to be worthy of investigation and so said that they would take action regardless of the content. Similarly, P6 who chose to take action in nearly, but not all instances, commented that "if you earn money from this job [as a system administrator], you have to check even if there's a message saying oh there's a risk of failure".

When asked which message they thought was most helpful, all participants reported that *Format C* (of any message) helped them better reason about the problem and act accordingly. As one participant said: "I think [the message] is detailed. I mean, it makes my life easier. It tells me what I should look at instead of me trying to find the problem. It automatically suggests me what I should check" (P4). Only one participant made reference to *Format B* (P6) who said that both formats B and C would likely prompt the same response, but *Format B* was more succinct.

3.5 Discussion

While *Format A* proves sufficient to invoke action in this context and so would be effective where all messages require attention e.g. a safety-critical system, these messages do result in a considerable amount of wasted effort with almost as many response to false alerts (see Figures 3.1 and 3.2). *Format B*, on the other hand, appear to reduce unnecessary action, but also appears to inhibit necessary action, suggesting that the proverb “a little information can be dangerous” may be appropriate. *Format C*, the most information-rich condition, however, shows a clear difference in comparison to the other message formats. While Figure 3.1 shows a marked improvement in users’ ability to identify when action is required, the biggest difference can be seen in Figure 3.2, where messages in *Format C* are considerably less likely to lead to unnecessary action. These findings are in line with Kulesza et al. (2013), whose work with recommender systems showed that the most complete explanations resulted in the most coherent mental models.

We then tentatively propose that the information-richness of messages plays an important role in users’ ability to reason about necessary action and that the experiments indicate that the more information is better. However, more research is needed to address the issues of verbosity and how this impacts users.

3.6 Implications

The findings suggest that short non-detailed messages are enough to persuade users to almost always attend to a monitoring system. However, long and detailed messages improves users’ efficiency in terms of when to appropriately attend to such systems. Finally, we discuss future avenues of research which we believe will further our understanding of natural language output’s application in the domain of smart technologies.

3.7 Limitations

The work outlined here was preliminary and further work is needed to better understand this space. Although our findings suggest that detailed messages result in more decisive decision making, the limited scope of the experiment means that these findings must be only considered advisory and that a more detailed investigation must be conducted to gain further insight.

Having designed this experiment and in reviewing the literature we identified a number of potential avenues for future research. We will refrain from discussing the need to develop a logically and grammatically correct scheme for the generation of messages as it is a field with substantial existing work. However, the way in which urgency can be conveyed in the context of inanimate smart devices could be of great interest. For example, a message designed to convey urgency may be misconstrued as low priority. Variation in recipient responses due to the conduit of communication may also be of interest.

Does the delivery method impact the actions or recipients? And finally, can agent-based systems, harnessing two-way communication provide natural language systems for smart devices that not only convey a message but in a way that invokes the necessary action from the recipient? Also in systems where an autonomous agent can suggest actions, it would be interesting to see how the inclusion of confidence information affects human operator's actions. In recent work (Verame et al., 2016a), it was shown that the addition of confidence information i.e. estimated probability that an inference that the agent is correct, can improve the adoption and reliance on agents.

3.8 Summary

In this chapter, we investigate what level of detail is required in the natural language messages produced by smart systems. Our findings suggest different levels of detail have advantages and disadvantages. Displaying short and simple messages are enough to persuade users to attend to the system, which can be more appropriate for safety-critical systems. However, providing detailed messages helps users to efficiently take action, as it helps them distinguish between false and correct alarms. As such, for systems that are less safety-critical, the implementation of detailed messages is more appropriate. Further work is needed to evaluate long term effects and to test the different levels of information in more varied and realistic scenarios.

Chapter 4

Wireless Smart Sensor - Balancing Sensor Fidelity and Connectivity

The recent deluge of smart devices in consumer markets can in part be attributed to new more power efficient radios and longer lasting batteries. One application which has benefited greatly from these advancements is Wireless Sensor Networks (WSNs) which have been used extensively in research and consumer products to, amongst other things, monitor health (e.g. Sull and Lim (2018)), reduce energy waste (e.g. Costanza et al. (2016)), and inform automation (e.g. Brush et al. (2011)). While this new wave of wireless battery powered devices has simplified installation and extended the potential installation locations, by permiting far greater freedom of movement in comparison to their fixed-wire predecessors, they are still subject to two key constraints which determine successful sensor installation:

- Connectivity - devices must be installed where they can reliably communicate with other system components e.g. a base station or the next “hop” in a mesh network.
- Sensor fidelity - sensors must be placed where they can monitor a target with a sufficient degree of accuracy. Following Laput et al. (2017) we call this constraint sensor fidelity.

Historically, these smart systems would have been installed and configured by professionals, but the falling cost of smart technologies, coupled with there no longer being a need for fixed wiring has made expert installation uneconomical for non-critical systems. So the burden of installation has been passed to end-users (Jakobi et al., 2018) who, unlike professional installers, are not necessarily technically literate (Edwards and Grinter, 2001).

While users of these systems might have a good understanding of the system's application, they are less likely to have a knowledge of radio propagation and the sensors' performance characteristics. Choosing an installation location where connectivity is present and sensor fidelity is sufficient (sites we refer to as *suitable installation locations*) can therefore be challenging. Prior work has explored user understanding of sensor fidelity (Beckmann et al., 2004) and connectivity (Costanza et al., 2010), however the constraints we considered in isolation and the question of how feedback can be utilised to inform user understanding so that a *suitable installation location* can be identified remains open (Fischer et al., 2017; Kazmi et al., 2014). In this chapter we report two user studies[1], one conducted under controlled conditions and the other in the field, examining how feedback can be used to foster greater user understanding and assist in the identification of *suitable installation locations*.

The controlled study was comprised of a series of iterations and was conducted with a total of 40 participants. Following a user-centred iterative approach, we began with 20 participants, investigating how visual feedback representing current sensor connectivity and fidelity values (via a simple bar meter display, akin to those used on most mobile phones to indicate signal strength) compared to receiving no feedback at all. While feedback considerably improved participants' ability to find *suitable installation locations*, only half of the ten feedback condition participants were able to complete the task. Observations led us to hypothesize that participants were unsure what was the best course of action. Drawing on our findings in Chapter 3, we tested a new condition: using the same bar-meter feedback as before, we added a short textual message advising participants to "try moving closer to the base station" when no connectivity is present. For comparison we also added a fourth condition, designed to provide an explicit course of action in the form of directions. In this condition the bar-meter feedback was presented alongside an augmented reality arrow which pointed towards the most suitable location detected thus far. Both of these two additional conditions proved effective, with all but one participant completing all tasks.

The second study was designed to further validate our findings, examining one of the most effective feedback strategies from the first study with eight additional participants in their own homes. Drawing on the qualitative analysis we discuss the observed interactions and how they relate to the first study and prior work.

We begin this chapter by describing the motivation for this work, then detail the study apparatus we developed to facilitate both user studies, before reporting the individual studies and finally discussing the findings from both studies in relation to prior work.

[1]Ethics approval granted by the University of Southampton (ref: 44874)

4.1 Motivation

Similar to Jewell et al. (2015), our work is motivated by scenarios where end-users install devices themselves, or a non-technical "facilitator" does it on their behalf e.g. an energy poverty advisor deploying environmental sensors to better understand a client's routines (as proposed by Fischer et al. (2017)), or a caregiver setting up a smart pill box to improve a client's medication compliance (as suggested by de Oliveira et al. (2010)).

Choosing where to install sensor nodes can be challenging for users. First they must place the sensor so that they can monitor their target with accuracy, and second, they must have reliable connectivity i.e. within radio range of their base station or repeater node. As we note in the introduction, the technically literate of users in this regard cannot be assumed, thus it is unlikely that they can realistically estimate the capabilities of the system.

Prior work of smart sensor systems has largely focused on the later stages of configuration and management, for example Castelli et al. (2017) who examines how best to visual smart home sensor data, or Clark et al. (2017) who using a hypothetical system investigates the effects of system abstractions on end users' understanding. While there have been some studies which shows that providing feedback can help users position sensors correctly (e.g. Patel et al. (2010b)) and that without this guidance errors can be made (e.g. (Hu et al., 2016; Beckmann et al., 2004; Costanza et al., 2010)), less is known about how users approach sensor installation when constrained by both sensor fidelity and connectivity (Fischer et al., 2017; Kazmi et al., 2014). Addressing this gap in the literature is the main motivation for the work presented in this chapter.

Further to the constraints of connectivity and sensor fidelity, the means of delivering feedback is also constrained. Demands for power efficiency, sustainability and cost-effective manufacture, makes the inclusion of even display technologies uncommon, especially when they are used solely for initial configuration (Costanza et al., 2010; Jewell et al., 2015). A common approach employed is to utilize the screen real estate of smartphones and tablet computers. Building on this prior work, in this chapter we investigate how feedback can inform user understanding, such that they can reason effectively about these constraints and find *suitable installation locations*.

4.2 Setting up Wireless Sensors

To design a valid user study, it was necessary to develop a prototype wireless sensor system. In this section, we detail the system we implemented using off-the-shelf hardware and describe the rationale behind the design choices we made.

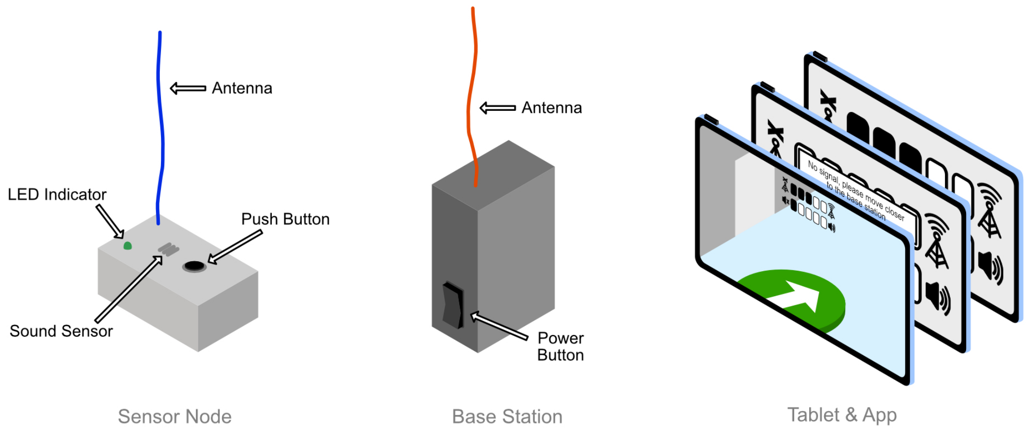


FIGURE 4.1: Diagram of the prototype wireless sensor system. On the left, a sensor node. In the centre, the base station. On the right, tablet computers displaying the three versions of the app.

4.2.1 Prototype Architecture, Operation and Overview

The system is composed of 3 component types: (i) wireless battery powered sensor nodes, (ii) a base station, to which the sensor nodes send readings, and (iii) a smartphone/tablet configuration app to present feedback (see Figure 4.1). The feedback is displayed on a mobile device to mirror commercial products which do not incorporate displays on sensor nodes for economic and power efficiency reasons.

Under normal operation, each sensor node transmits a reading to the base station every 2 minutes. The base station responds to each message with an acknowledgement. When a button mounted on the sensor node is pressed, the node enters configuration mode, reducing the interval between sending readings to approximately 200ms and establishing a Bluetooth link between the sensor node and the configuration app. This direct channel of communication allows the sensor node to update the configuration interface even when connection to the base station is not present. Encoded within every acknowledgement sent by the base station is connectivity metrics as perceived by the base station. It is this information, along with sensor fidelity measurements which are transmitted to the configuration app and inform the feedback. If no acknowledgement is received, then a message of no-connectivity is sent. At any given time, only one sensor node is allowed to be in configuration mode, when a sensor enters configuration mode all other sensor nodes are commanded to revert to normal operation.

To better explain the operation of this system, below we describe the procedure of setting up wireless sensors with this prototype system from a user's perspective (Figure 4.2 presents the process of setting up the system from a user's perspective as a flow diagram): (1) Connect the base station to a power outlet and an internet router, (2) launch the configuration app, (3) select a sensor node to configure by pressing the button on the sensor node, which also establishes a connection between the node and the app, (4) using

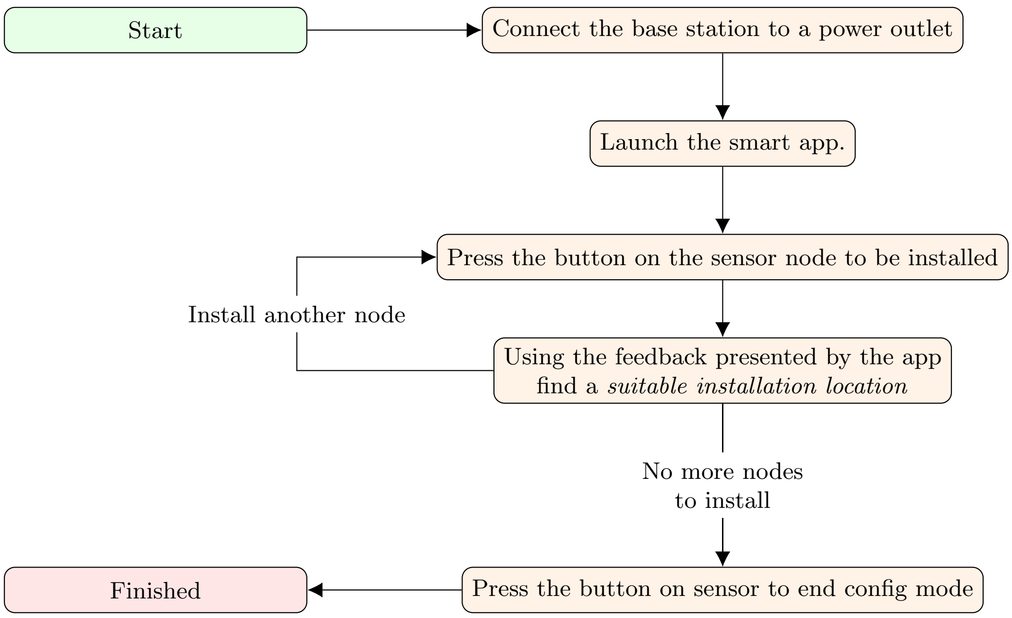


FIGURE 4.2: Flow diagram of prototype system setup from a user's perspective

the feedback presented by the app find a *suitable installation location*, (5) repeat steps 3 to 4 for all additional sensor nodes, and (6) after the last sensor node is installed, press the nodes button to exit configuration mode.

4.2.2 Prototype Design Rationale

The broad range of commercially available prototyping hardware offers considerable freedom when developing wireless sensor systems. In the remainder of this section we discuss the rationale behind our design choices.

4.2.2.1 Power

While there are many ways to power wireless sensors, we chose to use rechargeable Li-Po batteries. They are one of the most common solutions utilised in commercial applications, and the most reliable and maintainable option for the short duration of our studies.

4.2.2.2 Sensor Modality

Selecting an appropriate sensor modality for our prototype proved challenging. We ruled out sensors which relied on correct orientation (e.g. cameras, ultrasonic proximity sensors, directional microphones) to minimise possible confusion and maintain participants' focus on proximity alone. We also avoided sensors which overly constrained installation locations e.g. magnetic contact switches. While these challenges warrant further investigation, they are at this time beyond the intended scope of this study. After careful consideration, sound sensing was selected because: (i) literature Laput et al. (2017); Fogarty et al. (2006) and commercial applications[2] have demonstrated that it is effective at activity detection; (ii) in contrast to other sensor types, sound sensing has fewer limitations and external influences which would be undesirable for the study (e.g. light sensing can be affected by time of day); and (iii) they can yield feedback in real time (in contrast e.g. temperature might be influenced by thermal inertia); and finally, (iv) the ubiquity of sound in everyday life can make sound a more relatable topic for users to discuss.

4.2.2.3 Feedback Modality

While prior work explored the advantages and limitations of different feedback modalities (e.g. audio Costanza et al. (2010); Politis et al. (2015)), we focus on the *content* of the feedback. To this end, we selected a visual medium (i.e. the smartphone/tablet screen) to simplify prototyping, and because users are likely to be more familiar with it (e.g. similar to cellular and Wi-Fi signal strength on many common devices).

4.2.2.4 Radio

Wireless sensor nodes are in general constrained by power limitations, thus they commonly employ radios technologies with low power consumption. For our prototype we selected RFM69[3] radio modules. RFM69 is an increasingly popular low-cost and low-energy ISM band radio, variants of which have been used in commercial applications. We judged RFM69 to provide the best compromise between power consumption, range, availability and cost.

[2]For example: Leeo (`https://www.leeo.com/product`) a product which relays audio alarms to smartphones).

[3]For full specifications see `https://www.hoperf.com/`

4.3 User Study 1 - Office Building (Controlled)

In this first study, we designed and conducted a series of between-group user study. All of the studies centre on the same experimental task: finding *suitable installation location* while constrained by connectivity and sensor fidelity.

4.3.1 Study Design

Developing a controllable and yet ecologically valid study was particularly challenging because it requires observation of interactions at the boundaries of connectivity and sensor fidelity. Through an iterative process of design, piloting and evaluation we developed an experiment in which we asked participants to install 3 sensor nodes on the top 3 floors of a 6-storey office building (one sensor node per floor), communicating with a base station installed in the basement. On each floor, participants were tasked with finding a *suitable installation location* (where reliable connectivity with the base station was present and a target could be sensed with the greatest fidelity). To aid participants, a tablet computer was used (by those participants who received feedback) to inspect real-time connectivity and sensor fidelity feedback. Figure 4.3 shows a cross-section of the building used in this study, with labels highlighting the floors on which sensors were placed and the location of the base station. This particular building was selected because it offered several near-identical floors on which participants could repeatedly execute the sensor placement task, but with varying degrees of connectivity, due to increasing distance from the base station. Prior to the study commencing, the signal strength throughout the building was audited and the corridors of the top floors were shown to offer the greatest challenge in terms of connectivity. Through pilot studies it was concluded that 3 floors and 3 sensor nodes provided sufficiently valuable results whilst maintaining participants' interest and avoiding fatigue.

The printer room doors were chosen as a sensing target because: (i) the doors are vertically aligned within the building (as illustrated in Figure 4.3), making the task more repeatable across floors, (ii) they have sprung door closers, causing them to audibly slam shut, making them a suitable target for sound sensing and (iii) the application was considered a realistic use case for wireless sensor nodes.

The location of the base station was chosen so that on the first floor explored by participants (Floor 3) signal was present in the immediate vicinity of the printer room door, making the task easy. Starting with an easy task allowed participants to become familiar with the sensor nodes' operation whilst demonstrating the correct functionality of the system. On the next floor up (floor 4) the radio signal did not reach as far and a suitable location for the sensor node could only be found 3m from the printer room door (the sensing target). At this distance, the sound of the door was clearly detectable

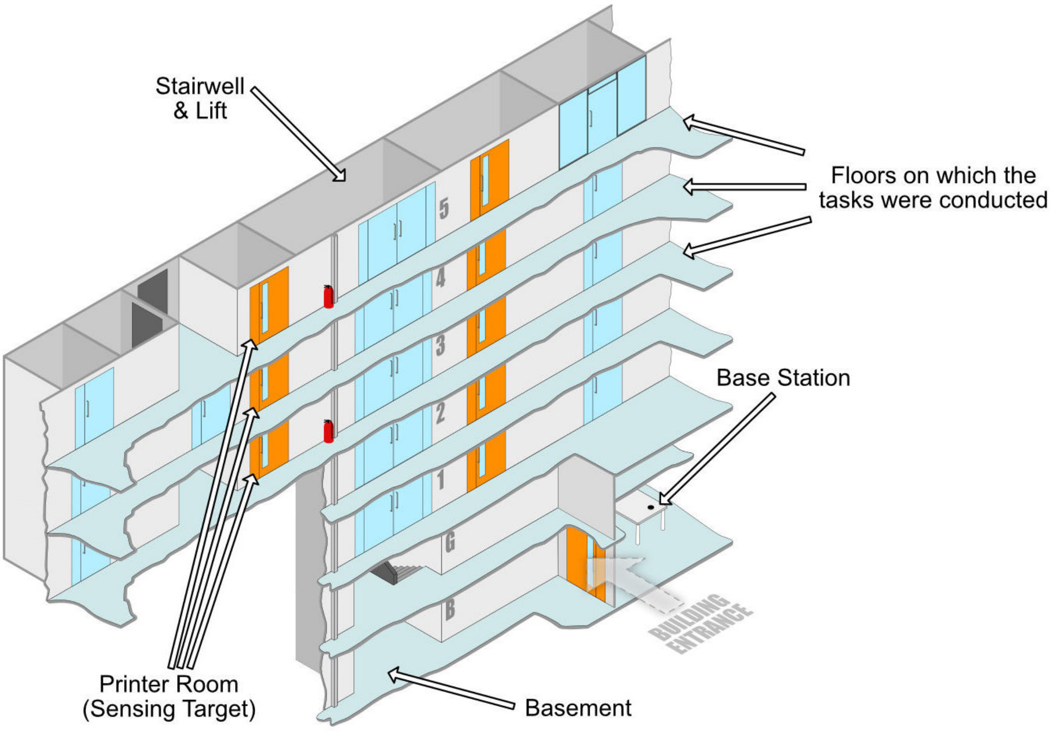


FIGURE 4.3: Approximate layout of the building, showing the position of the base station in relation to the floors on which the sensors were placed.

by the sound sensor. On the final floor (Floor 5) the radio signal was weaker still and connection could be established only around 6m from the printer room door. At this distance, the sound sensing was still acceptable (requiring a trade-off).

4.3.1.1 Procedure

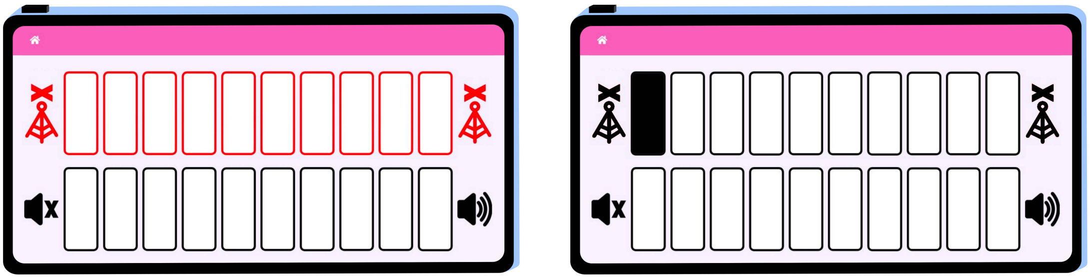
Participants were met in the building's lobby and taken to the basement offices where the base station was situated. Once each participant signed the informed consent form, they were asked to read a short instruction sheet which (i) outlined the system's components and high-level operation, (ii) detailed the nature of the sound sensors - describing it as smart and capable of differentiating between the target door and any other sound source, (iii) highlighted the need for sensors to communicate reliably with the base station, and (iv) specified their objective on each floor i.e. to position the sensors where the sensor could "hear" the target door whilst reliably communicating with the base station. Inline with other sensor systems of this type, participants were not told what constituted an acceptable number of bars for either factor. Then the experimenter physically demonstrated the location of the base station before escorting the participant to Floor 3 where the first sensor node placement task would be conducted. On arrival, the experimenter demonstrated the operation of the sensor node, reiterated the participant's objective (as written in the instructions) and highlighted the target door, opening it so the participant could hear it operate, and for all conditions excluding No-Feedback, see the effect on the screen. Participants were also reminded that if they wanted to test the sensor fidelity, they could open and close the door at any time. Participants were then left to position the sensor node freely while the experimenter stood at a distance to observe and take notes. Once participants indicated that they had selected a final installation position, or that they were not able to find a suitable location, they were asked to explain their reasoning before proceeding to the next floor. Prior to moving to the next floor, participants were handed the next sensor node so that they could see the signal strength change as they moved through the building. This process was repeated for all floors. After all the sensor nodes had been placed, a semi-structured interview was conducted to better understand participants' thought processes, motivations and understanding of the system.

4.3.1.2 Data Collection

Throughout the experiment, audio recordings and notes were recorded by the experimenter. In addition, connectivity and sensor fidelity data were logged by the base station and app so that the sensor placement locations selected by participants could be assessed for their suitability. For clarity, here we define a *suitable installation location* as a physical place where at least one bar of connectivity feedback is consistently present and the sensing target invokes a fluctuation in the sensor fidelity feedback of at least one bar.

Researchers' notes included: (i) Approximate task completion times (rounded to the nearest minute[4]), (ii) the starting point of participants' search, (iii) the maximum distance from the sensor target participants explored and in which directions i.e. towards or away from the base station, (iv) the locations tested, and (v) general observations of their behaviour. The final selected sensor locations were also photographed for later verification of recorded data. To avoid bias, measurements of search directions and distance were done covertly. Discrete markers and fixed artefacts of known distances from the sensor target were used to measure the distances of travel to the nearest meter. Researchers also took note of the regions explored by participants.

[4]Measured with limited accuracy to avoid participants feeling time-pressured. It was therefore hard for the experimenters to define precise start and stop times of each task



(A) Without Connectivity (B) With Connectivity

FIGURE 4.4: Screenshots of the Bar-Only feedback condition tested in Study 1A.

4.3.2 Study 1A: Bar-Meter Vs No Feedback

Our exploration of feedback and its role in aiding users began with an investigation of a common signal strength visualisation; the bar-meter. This visualisation was selected because of its ubiquity in everyday applications (e.g. cellular connections on mobile phones and Wifi indicators on desktop computers) and as such would be familiar to participants and so easier to interpret. Furthermore, bar-meters are also commonly used to represent sound levels (e.g. volume controls on TVs and input levels for microphones), making them suitable for both connectivity and sensor fidelity.

Calling on the aforementioned study design, we compare participants ability to identify *suitable installation locations* when subject to one of two conditions: Bar-Only and No-Feedback (10 participants per condition). Participants of the Bar-Only condition were presented with two ten segment bar-meter visualisations, one bar-meter representing connectivity and the second, sensor fidelity (see Figure 4.4). When no connectivity is present, the connectivity bar-meter (and associated icons) turns red. For contrast, No-Feedback condition participants were not issued with the tablet and received no feedback. Instead, participants were asked to place the sensor nodes where they felt was most suitable in terms of connectivity and of sensor fidelity. This condition represents the status quo for many wireless sensor devices, where the only feedback available to end-users is through checking if the system has received data from the node and if so, what data has been collected (e.g. through a control panel).

4.3.2.1 Participants

Twenty non-technical participants (9F, 11M) were recruited from our university participants pool (composed of the general public, university staff and students). Anyone who expressed interest was allowed to participate in the study, so long as they did not have technical hobbies or interests (e.g. computer programming), were not in technical employment (e.g. lab assistant) and were not technically educated (e.g. no degree in computing or engineering-related subjects). Each participant received £10 for their

TABLE 4.1: Age range and background of participants across all three studies.

Background	Count
Social science	15
Business, Economics and law	9
Art, media and design	7
Health and Biology	8
Education	1

Age Range	Count
60-69	2
50-59	3
40-49	2
30-39	8
20-29	25

participation. Table 4.1 reports the age range and background of participants (to avoid repetition, all three user studies are reported together).

4.3.2.2 Results

All participants in the No-Feedback condition were able to find a *suitable installation location* on Floor 3, while none were able on Floors 4 and 5. Bar-Only condition participants had more success, with all participants succeeding on Floor 3, seven succeeding on Floor 4, and five on Floor 5 (see Figure 4.5). To avoid repetition we detail the full qualitative analysis and quantitative findings for all user studies in Sections 4.3.6 and 4.3.5 respectively.

4.3.2.3 Findings

When connectivity was readily available (Floor 3) feedback was not necessary. However, on subsequent floors where connectivity was more challenging to locate, a marked improvement in Bar-Only condition participants' ability to identify *suitable installation locations* can be seen. However, only half of the 10 Bar-Only condition participants were able to complete the most challenging task (Floor 5). Observations of participants actions and qualitative data derived from interview transcripts and audio recordings suggest that while participants understood what needed to be done, they were unsure of the best course of action, lead us to develop new conditions which we test in the next user study.

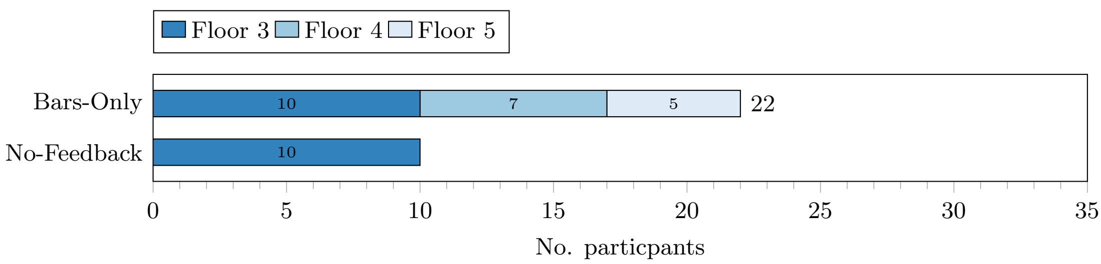
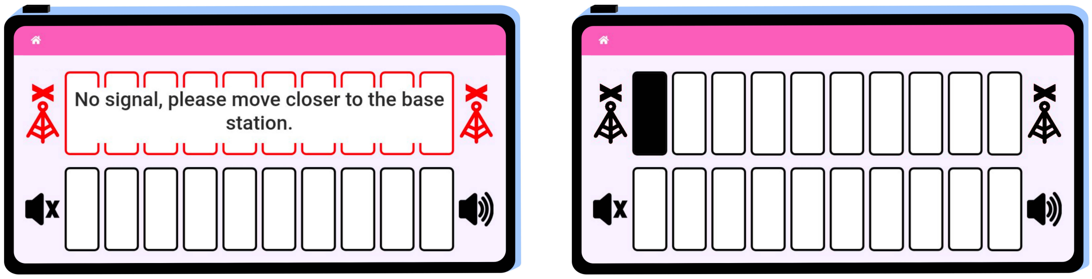


FIGURE 4.5: Successful task completion per condition and per floor



(A) Without Connectivity (B) With Connectivity

Figure 4.6: Screenshots of Bar+Message condition feedback introduced in Study 1B.

4.3.3 Study 1B: Suggesting a Course of Action

Drawing on the findings of Study 1A, we developed and conducted a second iteration of the feedback. Using the same study design, we tested a new condition (Bar+Msg) with 10 further participants. Participants of this condition were presented with the same feedback interface as the Bar-Only condition participants of Study 1A, with one fundamental difference: when no connectivity was present, the connectivity bar-meter was overlaid with a text box suggesting that participants "try moving closer to the base station" (see Figure 4.6). The rationale for adding this message is that it can be used to test if users' performance is affected by any uncertainty of what might be their best course of action.

4.3.3.1 Participants

Ten new participants (7F, 3M) were recruited and remunerated in the same way as the previous study and Table 4.1 reports the general age range and background of participants.

4.3.3.2 Results

All Bar+Msg participants except one were successful, finding *suitable installation locations* across all floors (Figure 4.7). Furthermore, Bar+Msg participants were significantly quicker in doing so (Figure 4.8). Details of the statistical tests can be found in Section 4.3.5.

4.3.3.3 Findings

In contrast to the Bar-Only condition examined in Study 1A, Bar+Msg participants were significantly more able to identify *suitable installation locations* and were markedly quicker. Only one participant failed to complete the task on Floor 5 and average study

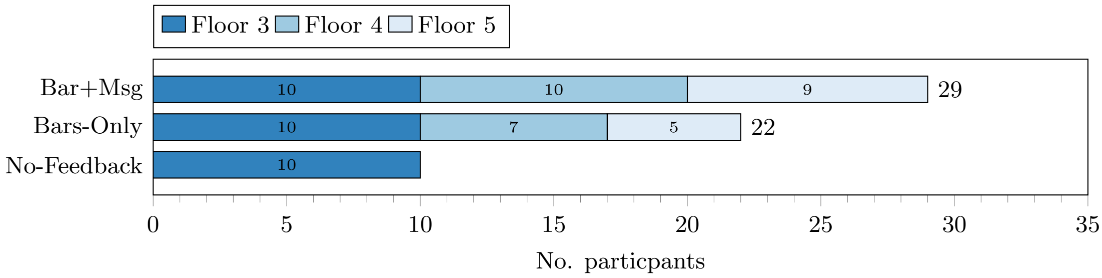


FIGURE 4.7: Successful task completion per condition and per floor

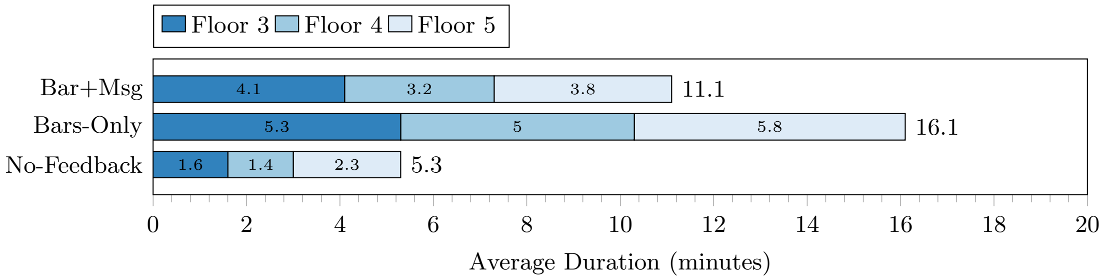
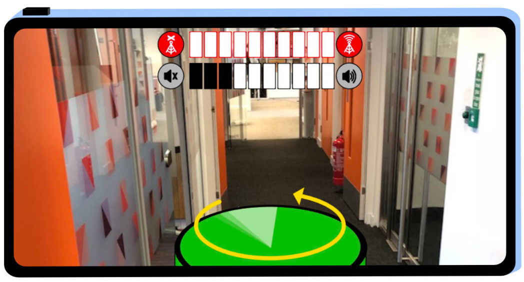


FIGURE 4.8: Task completion time

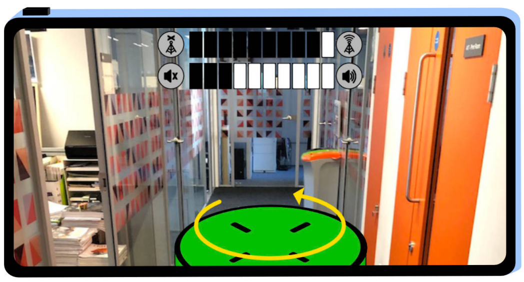
completion times were reduced from 16.1 minutes to 11.1 minutes. However, participants of Bar+Msg condition were commonly observed testing locations beyond the sensing target, further from the base station (in much the same way as Bar-Only participants). This highlighted an opportunity to further guide users. While a homeowner would be familiar with the installation environment, non-technical facilitators such as the energy advisors described by Fischer et al. (2017) may not.

4.3.4 Study 1C: Showing the Way

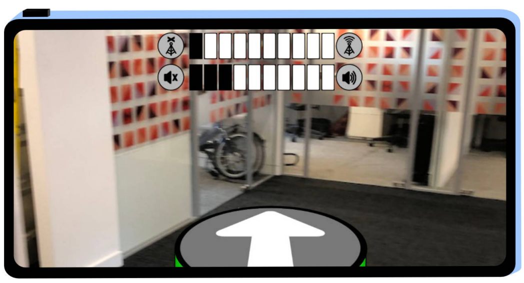
While the Bar+Msg feedback examined in the previous study proved to be significantly more effective at helping users find *suitable installation locations*, observations suggest there was room to improve the speed with which such locations were found. In this third and final iteration of the user study outlined in Section 4.3.1, we examine an additional and more directive condition: Bar+Arrow with 10 new participants. Participants of this new condition were presented, in conjunction with the bar-meter feedback of previous conditions, with a see-through display interface which harnessing AR technologies (now common in smartphones and tablet computers) to overlay an arrow visualisation which rotates such that it always points towards the location where the best connectivity thus far has been identified. If no connectivity has been detected the arrow is replaced by a slowly spinning cone, similar to popular depictions of radar screen (see Figure 4.9). We chose to adopt AR technologies because they provide a convenient way to map internal spaces without the need for additional hardware, and AR interfaces have been demonstrated in the literature (albeit through limited user studies) as effective navigational aids for pedestrians (e.g. Noreikis et al. (2017); Gerstweiler et al. (2015); Hesenius et al. (2018)).



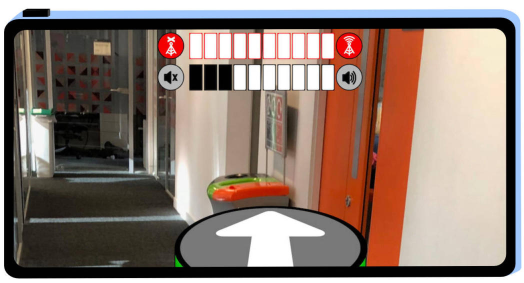
(A) Before connectivity for the current sensor has been detected - Spins at constant speed.



(B) The location where the best connectivity thus far was detected - Spins at constant speed.



(C) With connectivity, but less than that best detected thus far - Arrow points in suggested direction of travel.



(D) Without connectivity, but less than that best detected thus far - Arrow points in suggested direction of travel.

FIGURE 4.9: Screenshots of the Bar+Arrow feedback condition introduced in Study 1C.

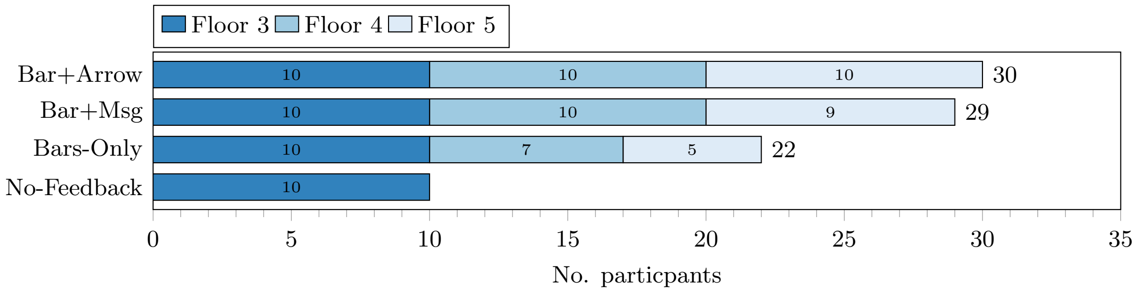


FIGURE 4.10: Successful task completion per condition and per floor

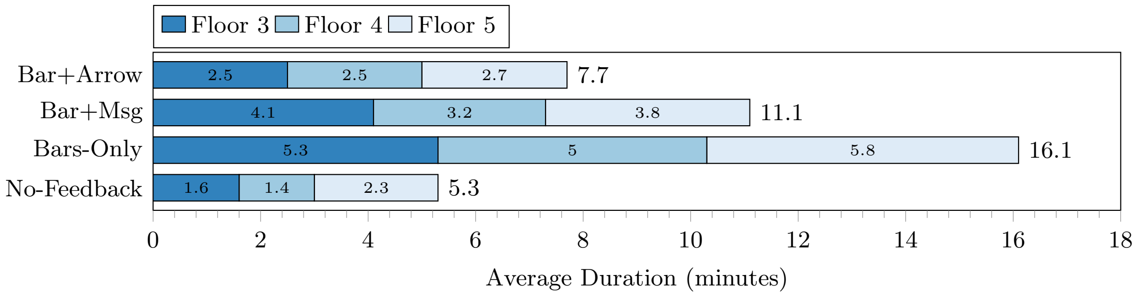


FIGURE 4.11: Task completion time

4.3.4.1 Participants

Ten new participants (6F, 4M) were recruited following the same procedure and criteria as in the previous study and Table 4.1 reports the general age range and background of participants.

4.3.4.2 Results

Successful task completion for Bar+Arrow participants were slightly improved, but comparable to Bar+Msg participants (30 to 29 respectively, see Figure 4.10). Task completion times were also improved, with the average study duration of Bar+Arrow condition participants reduced to 7.7 minutes (Figure 4.11). Unlike previous conditions, zero Bar+Arrow participants explored regions of the floor beyond the sensing target in a direction away from the base station (Figure 4.13).

4.3.4.3 Findings

While there was little room for improvement in task completion rate, between the Bar+Msg and the Bar+Arrow conditions, study completion times were considerably improved, with Bar+Arrow condition participants being considerable faster at finding *suitable installation locations* than any other condition. This coupled with observations which show that no Bar+Arrow participants ventured beyond the sensor target away from the base station suggests that the participants were guided successfully by this feedback. However, there are other important considerations which our qualitative analysis highlights.

4.3.5 Quantitative Analysis

In this section we report the quantitative findings and general analysis of all three studies. Overall, participants took on average 9.5 minutes (STD=4.70) to complete all 3 floors (see Figure 4.11 for details the sum total time spent by participants searching for suitable locations). A Kruskal-Wallis test revealed a significant effect ($p < 0.001, H = 30.76$) of condition on overall task completion times (i.e. the sum total of the time taken by each participant to select a location at which to place a sensor on Floors 3, 4 and 5, excluding transit time between locations). A post-hoc comparison using the Mann-Whitney U Test (with Bonferroni correction) revealed significant differences between all pairs of conditions[5].

[5] With a Bonferroni correction factor of 6: No-Feedback and Bar-Only ($p = 0.0002 < 0.05/6, U = 0.0, Z = -3.742$), No-Feedback and Bar+Message ($p = 0.0002 < 0.05/6, U = 0.0, Z = -3.742$), No-Feedback and Bar+Arrow ($p = 0.0026 < 0.05/6, U = 12.5, Z = 2.797$), Bar+Message and Bar-Only ($p = 0.0036 < 0.05/6, U = 11.0, Z = -2.903$), Bar+Message and Bar+Arrow ($p = 0.0029 < 0.05/6, U = 13.0, Z = 2.759$) and Bar-Only and Bar+Arrow ($p = 0.0002 < 0.05/6, U = 2.5, Z = 3.553$).

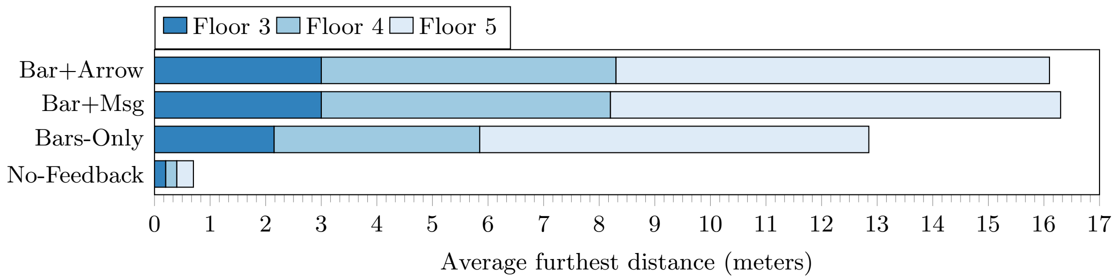


FIGURE 4.12: Average furthest distance explored from the sensing target in either direction

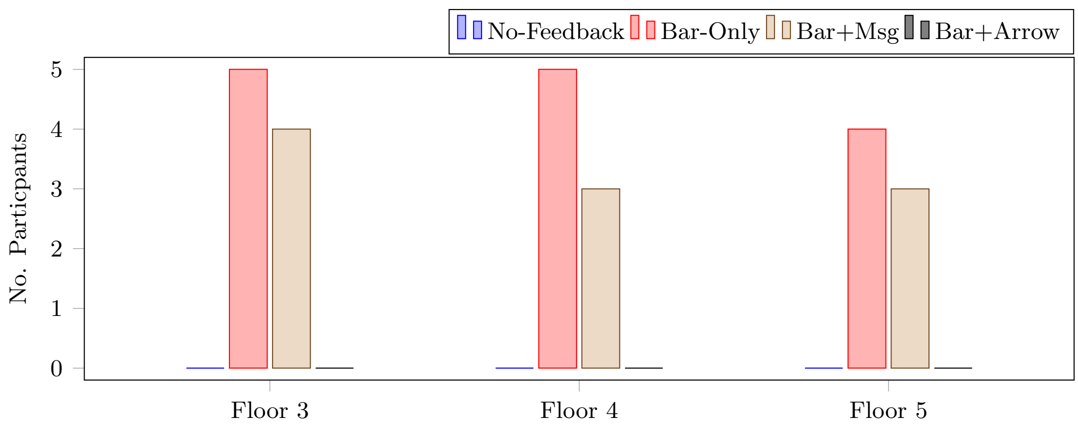


FIGURE 4.13: No. who explored locations further from base station beyond the sensing target

Figure 4.12 details the average furthest distances explored by participants when searching for *suitable installation locations*, and Figure 4.13 the number of participants who explored regions of the corridor beyond the sensing target which is further from the base station (and connectivity). No-Feedback participants explored considerably less of the available search area than feedback condition participants, with Bar+Arrow and Bar+Message participants ventured furthest from the sensing target.

4.3.6 Quantitative Findings

Success rate per condition and by the floor is reported in Figure 4.10. In summary, all participants in the No-Feedback condition were successful on Floor 3, while none of them was successful on Floors 4 and 5. Bar-Only condition participants had more success, with all succeeding on Floor 3, 7 participants on Floor 4 and 5 participants on Floor 5. Bar+Message and Bar+Arrow were most successful overall with only 1 participant of the Bar+Message condition failing to find a *suitable installation location* on Floor 5.

Figure 4.11 details the sum total time spent by participants searching for suitable locations on each floor by the condition. Overall, participants took an average of 9.5 minutes (STD=4.70) to complete all 3 tasks. Participants of the No-Feedback condition took the least time to select what they considered to be a suitable installation location, followed by Bar+Arrow participants, then Bar+Message and finally Bar-Only participants who took the longest. These times do not account for incorrect selections i.e. No-Feedback participants were fastest, but also the least likely to succeed in finding *suitable installation locations* (see Figure 4.10).

A Kruskal-Wallis test revealed a significant effect ($p < 0.001, H = 30.76$) of condition on overall task completion times (i.e. the sum total of the time taken by each participant to select a location at which to place a sensor on Floors 3, 4 and 5, excluding transit time between locations). A post-hoc comparison using the Mann-Whitney U Test (with Bonferroni correction) revealed significant differences between all pairs of conditions[6].

Researchers also recorded information relating to the search area covered by participants. To avoid bias, these measurements were done covertly. Discrete markers and fixed artifacts of known distances from the sensor target were used to measure the distances of travel to the nearest meter. Researchers also took note of the regions explored by participants. Figure 4.12 details the average furthest distances explored by participants when searching for *suitable installation locations*, and Figure 4.13 the number of participants who explored regions of the corridor beyond the sensing target which

[6] With a Bonferroni correction factor of 6: No-Feedback and Bar-Only ($p = 0.0002 < 0.05/6, U = 0.0, Z = -3.742$), No-Feedback and Bar+Message ($p = 0.0002 < 0.05/6, U = 0.0, Z = -3.742$), No-Feedback and Bar+Arrow ($p = 0.0026 < 0.05/6, U = 12.5, Z = 2.797$), Bar+Message and Bar-Only ($p = 0.0036 < 0.05/6, U = 11.0, Z = -2.903$), Bar+Message and Bar+Arrow ($p = 0.0029 < 0.05/6, U = 13.0, Z = 2.759$) and Bar-Only and Bar+Arrow ($p = 0.0002 < 0.05/6, U = 2.5, Z = 3.553$).

is further from the base station (and connectivity). No-Feedback participants explored considerably less of the available search area than feedback condition participants, with Bar+Arrow and Bar+Message participants ventured furthest from the sensing target.

4.3.7 Qualitative Findings

Transcripts of all audio recordings and researchers' notes collected during the studies were independently coded by two researchers. Codes were initially drawn from research questions and then supplemented with those that emerged from the interviews, before being grouped by consensus. In the subsequent subsections, we detail these groups and give example quotations. For brevity, we refer to participants by condition and participant number, for example, N7 was the seventh participants of the No-Feedback condition. Prefixes "B", "M" and "A" refer to the Bar-Only, Bar+Message and Bar+Arrow conditions respectively.

4.3.7.1 Engagement with the Task

Observations and participants' comments demonstrate that the task was sufficiently challenging and forced participants to seek a compromise. For example, A3: "I felt like I should balance how good the signal was and how loud you could hear the printing room door. It was hard to balance those things because when the signal went up the loudness went down". Further to this, participants in the 3 feedback conditions described the system as easy to use and found the configuration process easy to understand e.g. M7: "It was easy. I just had to look at the bars. It's simple", suggesting that the usability of the system did not impede performance.

4.3.7.2 Sensor Fidelity

Participants were asked which factors were most likely to affect sensor fidelity (Table 4.5). Of the factors described, participants most commonly cited distance from the sensing target (31 participants). Interference from our sound sources (11 participants), e.g. "people walking past might dampen the sound" (M9), obstructions impeding the transfer of sound (8 participants) and source-strength i.e. how loud the sensing target is (7 participants) were also frequently reported. Three participants conflated signal and sound sensing (coded as connectivity) e.g. "If you have good signal then it might pick up little sounds better". One described how the device's orientation might impact sensor fidelity, i.e. was the microphone pointing at the sensing target and one the sensor-ability (e.g. M7: "I wasn't sure if the microphone was sensitive [enough]").

TABLE 4.2: Participants' reported motivations

Codes	No-Feedback	Bars-Only	Bars+Msg	Bar+Arrow	Total
Connectivity	3	10	10	10	33
Sensor Fidelity	8	7	8	9	32
Safety	6	1	4	2	13
Other	0	0	1	0	1

TABLE 4.3: Participants' prioritization of motives

Codes	No-Feedback	Bars-Only	Bars+Msg	Bar+Arrow	Total
Connectivity	3	7	5	6	21
Sensor Fidelity	6	1	3	0	10
Both Equally	1	2	2	4	9

TABLE 4.4: Factors which affect connectivity

Codes	No-Feedback	Bars-Only	Bars+Msg	Bar+Arrow	Total
distance	5	8	9	9	31
obstructions	7	3	8	6	24
interference	1	1	0	2	4
orientation	1	0	1	0	2
sensitivity	0	0	0	1	1

TABLE 4.5: Factors which affect sensor fidelity

Codes	No-Feedback	Bars-Only	Bars+Msg	Bar+Arrow	Total
distance	5	6	6	8	25
interference	2	3	4	2	11
obstructions	2	0	2	4	8
source-strength	3	1	3	0	7
connectivity	2	0	1	0	3
sensitivity	0	0	0	1	1
orientation	0	1	0	0	1

TABLE 4.6: Reported connectivity search strategies

Codes	No-Feedback	Bars-Only	Bars+Msg	Bar+Arrow	Total
Follow	0	4	8	10	22
Knowledge	3	3	3	0	9
No-mention	6	0	0	0	6
Elevation	1	4	0	0	5
Random	0	3	1	0	4

TABLE 4.7: Observed search behaviour (No. participants)

	No-Feedback			Bars-Only			Bars+Msg			Bar+Arrow			Total
Floor	3	4	5	3	4	5	3	4	5	3	4	5	
Continuous	0	0	0	1	3	7	7	9	10	10	9	9	65
One Spot	10	10	10	2	0	1	0	0	0	0	0	0	33
Place & Test	0	0	0	7	7	2	3	1	0	0	1	1	22

4.3.7.3 Search for Connectivity

All 30 participants who received feedback started their search for connectivity at the sensing target. When no radio connection was available, they moved away in search of signal (not necessarily in the direction of the base station). Although No-feedback condition participants received no information pertaining to connectivity, four of the ten participants did strategize about where the best signal was likely to be found. All four made reference to the distance between the sensor node and the base station, e.g N8 who said that they placed the sensor "closest to the door [...] but on the side closest to the base station".

Observations of participants as they searched for *suitable installation locations* were classified as one of the following: (i) Continuous - where participants focussed on the feedback as they moved the sensor node through the space, (ii) Place & test - where participants selected a location before checking the feedback to see if it was suitable - if the location was considered unsuitable a new location was sought, and (iii) One Spot - where participants selected a location after contemplating the available options without consulting the feedback (if available), then ended their search. Table 4.7 reports the frequency of these classifications across conditions.

Participants' responses to interview questions pertaining to their search strategy were coded as one or more of the following (see Table 4.6): (i) no-mention - they did not mention making any attempt to find signal, (ii) random - they had no strategy and moved randomly until signal was found e.g. B9: "I just walked around until I saw something come up", (iii) knowledge - they used their knowledge obtained by walking through the building e.g. B2: "[...] I tried to remember where the room was downstairs [where the base station was located], I tried to hover above that so I could get good signal", (iv) elevation - they moved the sensors vertically e.g. N9: "As low as possible so it can reach the [base station in the] basement", and (v) follow - they followed the UI e.g. "When [the bars] fall I stop and move the other way" (B4) or "I just followed the arrow" (A9). We also note an observation common to all feedback condition participants was the tendency to fixate on a location where a fleeting signal had been seen, participants would make repeated small adjustments in attempts to rediscover/stabilize the signal.

4.3.7.4 Comprehension of radio signals

To better understand the motivations of participants, we asked what factors they believed would most likely affect connectivity (Table 4.4) - distance between the base station and the sensor (31 participants) and obstructions such as doors, walls and floors (24 participants) were most common. Interference e.g. electrical interference from HVAC systems (4 participants), antenna/device orientation (2 participants) and sensitivity i.e. the specifications of the radio technology (1 participant) accounting for the remaining reported factors. Examples of statements coded as distance “The closer you are to the base the better the signal is going to be” (M4); obstruction: “if the antenna on top is being blocked” (B1) or “The floor is made of concrete I think, so can block some signal out” (A9); interference: “things like the printer might have an effect on the signal” (N6); orientation: “If I bend it [the antenna] in the direction towards the base station it might be able to pick it up better” (M2); and sensitivity: “how does it get transmitted, what radio is used” (A4).

Given that IoT and environmental sensor systems are relatively new, we asked participants if they could think of similar systems. Twelve participants suggested home WiFi networks, 8 participants home media equipment (e.g. TV, baby monitors), 6 participants IoT lighting systems or remote garage controls, 3 participants wireless burglar alarms, 5 participants mobile phones, 2 participants walkie-talkie radios and 2 participants children’s toys (e.g. M9 who compared it to a “Cup and string phone”: for it to work the string must be under tension for your voice to carry). Ten participants could not think of any similar systems.

4.3.7.5 Placement motivation

When asked what motivated the selection of locations which they considered to be *suitable installation locations*, participants reported good connectivity most often (33 participants), with sensor fidelity a close second (32 participants), safety and security third (13 participants), and 1 participant (B6) reported that they wanted the device to be “hidden from people’s view”, i.e. it should be inconspicuous. Table 4.2 details how these motives were distributed across the conditions. Example comments classified as motivated by connectivity included: M4 “I tried to see where the signal was good and then I opened the door to see if the [sensor] could detect the sound” and N1 “signal was important, but I had no way of knowing”. Example comments classified as motivated by sensor fidelity included: N8 “Basically, [I wanted to stay] closest to the door and then [move] in the direction of the base station” and B10: “I thought the door would make the most noise where it locks, so I aimed to place the sensor as close to where the lock is”. Example comments classified as motivated by safety included: A6: “My first concern was it should not be destroyed” and N10: “where people won’t possibly step on it”. When subsequently asked which took priority, connectivity or sensor fidelity (see Table 4.3), 21 participants reported to prioritize connectivity e.g. “I think the signal is a bit more important for me because I felt if there wasn’t signal for the sensor, then the sensor wouldn’t be able to send the data down to the base station.” (A3), 10 participants said sensor fidelity e.g. “I was looking more at the sounds than the signal, because I thought ok, the sound is what makes [it work]” (M4) and 9 participants said that both were of equal importance e.g. “I didn’t see the point of having one without the other” (M9).

4.3.8 Discussion

The varied level of performance across conditions, both in terms of success rate and task completion time, suggests that our experimental design was successful in providing a challenging setting for our participants to set up a wireless sensor network. Moreover, participant comments and observations suggest that our prototypes were generally usable and did not impede participants.

4.3.8.1 Is Feedback Necessary to Guide Sensor Placement?

All participants in the No-Feedback condition managed to successfully complete the sensor placement task on Floor 3. This result suggests that feedback is not needed when the task is *trivial*, namely setting up a wireless node in an area that has generally good radio connectivity. This might be why some of the systems reported in the literature (e.g. Hu et al. (2016); Patel et al. (2010b)) and many commercially available products do not provide any connectivity feedback. However, care should be taken not to assume that connectivity is present everywhere: even modern mesh networking technologies can be impacted by factors such as interference and obstructive building materials. Participants in the No-Feedback condition were successful *only* on Floor 3: none of them succeeded on Floors 4 and 5 where connectivity is less readily available. In contrast, most participants in the feedback conditions (Bar-Only, Bar+Message and Bar+Arrow) were successful on all floors. The implication, then, is that feedback becomes critical as the radio connectivity becomes more scarce, and the wireless node placement becomes more challenging. The **design implication** then is that real-time feedback should be provided, if at all possible. These results are in line and extend those reported by Hu et al. (2016). Some of their participants failed in setting up a wireless sensor network because they had no feedback and hence were not able to detect that connectivity was not present.

Similarly, when considering sensor fidelity, Beckmann et al. (2004) proposes that system designers should "avoid the use of cameras, microphones and highly directional sensors if possible" because participants of their studies found them difficult to install correctly without feedback (The sound sensor was the sensor most likely to be mis-installed with only a 67% correct installation rate). While the sensors we used are less directional (i.e. they are to some extent Omni-directional), our results suggest that the fidelity feedback made it possible for our participants to identify correct locations. This finding is in line with Patel et al. (2010b), who report feedback being critical in a user's ability to correctly place sensors, albeit as part of a more restricted task.

What strategies did our participants develop in reaction to the feedback information they received? Through qualitative analysis, we identify three search strategy classifications (described in Section 4.3.7.3). In a similar vein, Costanza et al. (2010) identities two strategies: (i) “aggressive” - where participants begin their search moving quickly through a space until the boundary of connectivity is found, then slowing their movements, conducting a more fine-grained search for the most suitable installation location, and (ii) “protective” - where participants stop their search as soon as a small degradation in connectivity is noticed. Most feedback condition participants in our study adopted either a “continuous” or a “place & test” strategy. The “continuous” strategy we observed is similar to Costanza et al.’s “aggressive” strategy. However, the “place and test” strategy departs from that prior work. We propose two possible reasons why, (i) while Costanza et al. (2010) focussed only on connectivity, our participants also had to take into consideration sensor fidelity - as a consequence, they often started their search from the sensing target where connectivity was not always present, and (ii), in our study, participants sometimes stopped monitoring the connection quality until they got close to the sensing target (most notably when climbing up the stairs between one floor and the next).

In contrast, participants in the No-Feedback condition had very little to guide their sensor node placement strategy. The most commonly reported motivation was sensor fidelity (placing the sensor as close as possible to the sound source) and safety (placing the sensors out of harm’s way). The few who did consider connectivity did so based on physical proximity to the base station, e.g. placing the sensor next to the door, but on the side closest to the base station. The interviews also suggest (as summarized in Table 4.4) that without feedback, participants were slightly more inclined to consider obstructions rather than distance, as a factor that influenced the connectivity.

4.3.8.2 The Importance of Directing Users

Bar-Only participants did considerably better than no feedback, in that most (7) of them completed the task on Floor 4, and a half of them (5) on Floor 5 (compared to none on either floor with No-Feedback). As discussed in the previous subsection, these results point to the importance of feedback. However, Bar+Message and Bar+Arrow did considerably better than Bar-Only and on average they took significantly less time, with all participants in the Bar+Arrow condition and all but one in the Bar+Message being able to successfully complete the task on all 3 floors. The maximum distance from the sensing target that participants considered and explored was a crucial factor. The textual message was instrumental in encouraging participants to explore further afield, as also highlighted by the distance data we collected (Figure 4.12). These findings suggested that providing a course of action in addition to feedback is important. These results are in line with recent work from Cognitive Science Harold et al. (2015), which revealed that providing a simple “linguistic warning” (similar to the message displayed in the Bar+Message condition) was effective in improving participants’ interpretation of time-series data visualizations. Similarly, recent HCI work highlighted the dominant effect of titles on users’ interpretation of visualizations Kong et al. (2018).

Furthermore, participants who used the AR-based UI (Bar+Arrow condition) completed the study in significantly less time than those who received a textual message (Bar+Message condition), while having a slightly higher success rate. Such lower task completion time suggests that the AR interface increased usability. As illustrated in Figure 4.12 and Figure 4.13, while the maximum distance from the sensing target was approximately the same for participants in the Bar+Message and Bar+Arrow conditions, those in the last condition explored only in the correct direction. This is probably because the arrow on the UI pointed them in the direction where the signal strength was higher. The differences could also be attributed to the see-through video display, or to the AR arrow being faster to interpret and more directive. Indeed prior studies Noreikis et al. (2017); Gerstweiler et al. (2015); Hesenius et al. (2018) pointed out that AR is an effective means of aiding users when navigating indoor and outdoor environments.

These results highlight two clear **design implications**. The first is that it is useful to include textual prompts in wireless sensor node setup interfaces, as they help users take action on the feedback. The second implication is that AR interfaces should be considered as a mechanism to deliver guidance, given the ubiquity in modern smartphones and the availability of quality toolkits such as Apples ARKit[7], Google's ARCore[8] and the open-source project ARToolkit[9].

4.3.8.3 Reflecting on Users' Understanding of the System

Following Beckmann et al. (2004), we investigated our participants' understanding of the system they were asked to deploy. We asked our participants to draw comparisons between our system and other systems they had been exposed to previously. The vast majority of participants (28 in total) made reference to other wireless systems e.g. home WiFi, baby monitors and remote control toys. This result points to the fact that in the 15 years that have elapsed between the work of Beckmann et al. and our own, wireless technologies have become prominent in everyday life.

Despite the familiarity reported by most of our participants with wireless devices, some of our results indicate that technical issues associated with digital radio should not be underestimated. For example, short lived anomalies in radio signals are common and can be caused by all kinds of environmental conditions e.g. interference from electronic sources. In our study, we observed how study participants commonly reacted to small fluctuations in the connectivity feedback, stopping to examine locations where fleeting connectivity had been seen. This is not dissimilar to the *protective* behaviour reported by Costanza et al. (2010), where small changes in signal immediately affected user behaviour. Rather than move further away from the sensing target, where more reliable

[7] `https://developer.apple.com/arkit/`
[8] `https://developers.google.com/ar/`
[9] `http://www.artoolkitx.org/`

connectivity might be found, participants persisted in trying to rediscover connectivity. Given that our participants were by design non-technical and they were not informed what constitutes acceptable connectivity, they may have believed that if they could find the signal then this location would be reliable. A **design implication** is that feedback should account for such temporary variations in signal and be "smoothed out", to prevent lay-users from focussing on the microscopic and more on the macroscopic scale. Furthermore, providing an indication of what acceptable connectivity looks like may also be of help.

In the interviews, participants were also prompted about their understanding of factors which may affect sensor fidelity. The results (summarised in Table 4.5) indicate that distance is the factor most commonly mentioned directly. Moreover, most of the other factors mentioned (e.g. 'interference,' 'obstructions,' and 'sensitivity') are also *indirectly* related to distance – i.e. they would be addressed by placing the sensor closer to the sensing target. The only exception is the connectivity: three of our participants suggested that the quality of the radio connection between the sensor and the base station would influence the quality of the sensor recordings. Perhaps this misconception could be explained in relation to the way analogue radio works.

The safety of the sensors was one of the most common motives described by participants of our study when selecting *suitable installation locations*. This was expressed in terms of protecting the device from accidental damage, theft and the safety of other occupants. This finding is in line with previous reports of concerns about safety i.e. with children and pets Beckmann et al. (2004). In our study, some participants' concerns about sensors safety might be considered excessive, suggesting that perhaps our prototypes looked more fragile than they really were. A **design implication** then, is that sensors need not only to *be* robust, but also *look* robust.

4.3.9 Summary

In this controlled user study, 40 participants were subject to one of four feedback conditions. These conditions were designed to examine how best to support users in their search for *suitable installation locations* when setting up wireless sensor nodes. Our findings indicate that providing connectivity feedback is significantly more effective at guiding users than no feedback, that providing a suggested course of action significantly expedite users' search in comparison to feedback alone, and that spatially informed suggestions significantly further reduce search times. In addition, drawing on our qualitative and quantitative findings we highlight a series of design implications which we hope will enable better feedback design and improve end-user deployments of wireless sensor node systems.

4.4 User Study 2 - Users' Homes (Field)

To build on the qualitative findings of the previous study (documented in Section 4.3), a follow-up field study was conducted utilising the same prototype apparatus. To improve ecological validity this second study was carried out in participants' own homes, as a result the study design was modified. In the next section, we describe these modifications before reporting our findings.

4.4.1 Study Design

Participants of this study were tasked with placing three sensor nodes in three different rooms within their homes (one per room). In each room, participants were tasked with finding a *suitable installation location*- where reliable connectivity with a base station (which we installed next to their wifi router) was present and a target could be sensed with the greatest fidelity. In contrast to Study 1, the opening and closing of windows was selected as the sensing target. Windows were chosen because: (i) they are also common to users homes; (ii) they can be detected using sound sensing in the same way as doors, and (iii) they have an application which was relatable for participants. Figure 4.14 shows an example layout of a participant's home, the location of the base

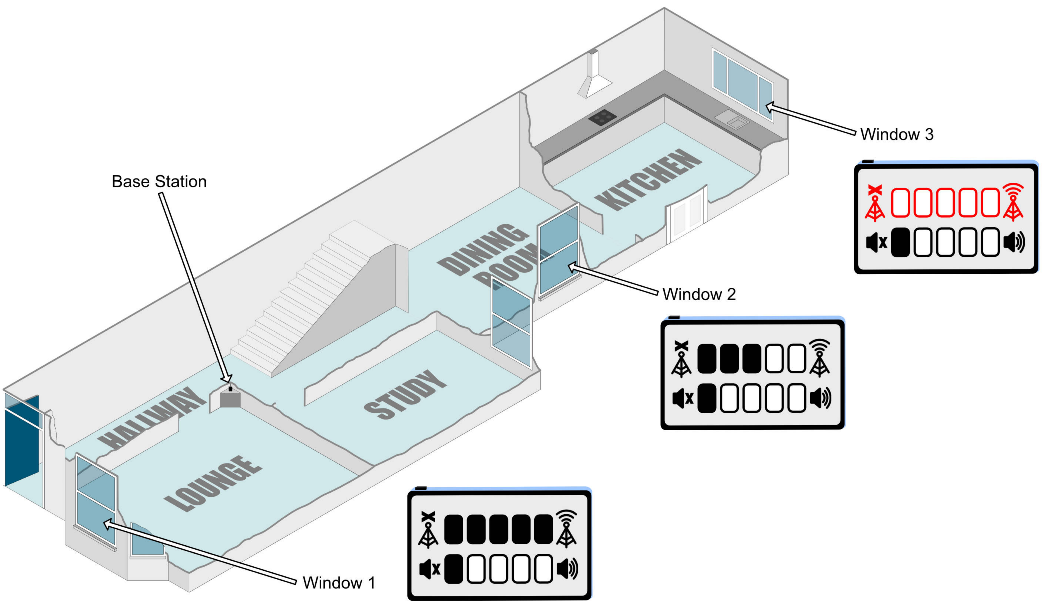


FIGURE 4.14: An example of a participant's home, showing the position of the base station in relation to the windows which the sensors were targeting.

Table 4.8: Age range and background of participants across all 3 studies.

Background	Count
Business and Education	5
Health and Biology	2
Geography	1

Age Range	Count
70-79	1
50-59	1
30-39	6

station and the sensing targets (this example was selected because it is representative of three of the participants' Victorian terrace homes).

After considerable deliberation, we concluded that making meaningful comparisons of data collected in different environments (i.e. participants own homes) was not feasible, as such all participants of this study were subject to the same feedback. Of the feedback conditions tested in Study 1, the Bar+Msg feedback was selected. This feedback condition was chosen because not only had it demonstrated its effectiveness, it was considered more likely to operate correctly in unknown environments.

Prior to the study comensing (and after consent was granted), participants were asked to give a guided tour of the rooms they were willing to use in the study, this was conducted under the pretence that the rooms needed to be checked for interference. However, in reality the experimenter was assessing the signal strength at each location. To ensure that all participants in the study experienced a range of connectivity and sensing challenges, an offset was discreetly applied to all sensors equally so that the window location with the weakest radio signal was reduced to just below zero (e.g. Window 3 in Figure 4.14).

4.4.2 Participants

Eight participants (5F, 3M) were recruited from the researcher's social network. Inline with the first study, anyone who expressed interest was allowed to participate in the study, so long as they: (i) did not have any knowledge of the work; (ii) did not identify as having technical hobbies or interests (e.g. computer programming); (iii) were not in technical employment (e.g. lab assistant); and (iv) were not technically educated (e.g. no degree in computing or engineering related subjects). In a change from the Study 1, participation each received £20 as remuneration for their participation. Table 4.8 reports the age range and background of participants (to avoid repetition, all three user studies are reported together).

4.4.3 Procedure

At the start of the study participants were asked to read an instruction sheet which detailed: (i) how the sensing nodes work i.e. that they detect sound, (ii) that the sensor nodes need to communicate with the base station, and (iii) instructions on how to operate the tablet, app and sensor nodes. To help participants better understand the task, the instructions also ask participants to envision the following scenario:

> You have purchased a sensor system to monitor the windows in your home. The monitoring of windows could for example, be used to prevent energy waste (when a window is open and the heating is turned on) or to improve your home's security. In this study, we want you to place the three sensors in three separate rooms. Your goal is to ensure that the sensors can detect the operation of the windows and communicate with the base station. Please position the sensors as if you were installing them for long term use.

Upon completion, participants were asked if they had any questions, before the key points from the instructions were verbally summarised by the experimenter. The experiment was then conducted in the following three distinct phases:

1. In the first phase participants were asked to place each sensor so that it could hear the target window and communicate with the base station (one sensor per room selected by the participant for the study). Once the sensor was placed and the participants indicated that they had completed the task the participant was advised to move to the next location. So as not to influence participants actions during this phase, no questions were asked. Participants were however reminded prior to each placement that they could operate the window to test the sensor fidelity.

2. In the second phase, each of the sensor nodes was revisited in order. At each sensor the experimenter asked the participants what motivated them to place the sensor node in the chosen location, what they thought the signal reading indicated and how effective they felt the sound sensor would be at detecting the window. The participants were then asked to place the sensor node in two further locations: (i) where they would have placed it if signal was not a concern, and (ii) the furthest position from the target window where they felt the sound sensor could accurately detect the window.

3. In the third and final phase a semi-structured interview was conducted to better understand the motivation and rationale of participants during the study. The questions were structured in the same manner as the Study 1.

4.4.4 Data Collection

Throughout the experiment, audio recordings and notes were recorded by the experimenter. In addition, connectivity and sensor fidelity data was logged by the base station and app so that the sensor placement locations selected by participants could be

Table 4.9: Field study task completion times (minutes)

Codes	P1	P2	P3	P4	P5	P6	P7	P8	Mean	S.dev
Room 1	5	5	5	3	5	3	3	5	4.25	1.04
Room 2	5	4	5	4	5	3	2	5	4.13	1.13
Room 3	15	9	12	8	10	7	5	12	9.75	3.20

assessed for their suitability i.e. were they *suitable installation locations*. Photographs of the selected locations were also captured for verification purposes. Researchers' notes included: (i) Approximate task completion times (rounded to the nearest minute[10]), (ii) the starting point of participants' search, (iii) the locations tested, and (iv) general observations of their behaviour. The final selected sensor locations were also photographed for later verification of recorded data.

4.4.5 Quantitative Findings

In all of the participants homes' it was possible to apply the connectivity offset such that two of the three sensor nodes could establish a connection in the immediate vicinity of the window and one sensor node (the "challenging target") could establish a connection 1 to 2 meters from the window.

All participants successfully managed to find *suitable installation locations* for all sensor nodes. Figure 4.9 presents the task completion times for all participants in all rooms. The mean task completion time for participants in the first room tested was 4.14 (Std.dev 1.07), in the second room: 4.01 (Std.dev 1.15) and in the final room (the challenging target): 9.43 (Std.dev 3.31). This demonstrates the study design was successful in challenging participants to find compromise between signal and sensing.

4.4.6 Qualitative Findings

In the same manner as the controlled user studies (Chapter 4.3), transcripts of audio recordings and researchers notes were coded independently by two researchers. The codes were drawn from the findings of the previous studies and supplemented with those that emerged from the interviews. Here we report findings resulting from this process. To avoid confusion with Study 1, we refer to participants by their subject number prefixed by the letter "F" for field.

[10]As per Study 1, completion times were measured with relatively low accuracy, because we did not want participants to feel time-pressured. It was therefore hard for the experimenters to define precise start and stop times of each task

FIGURE 4.15: Example sensor placements in field study

4.4.6.1 Engagement With The Task

Participants largely reported that the system was simple to use and the feedback was easy to understand e.g. “The icons on the tablet were immediately recognisable” (F7). Observations made during the studies support this, with participants responding quickly to no-signal indicators, suggesting (as per Study 1) that the usability of the system did not impede performance. In addition, participants reported the task to be challenging e.g. “finding somewhere good to put it [the sensor node] in terms of signal wasn’t easy” (F1), further demonstrating the quantitative findings of this study that show the task design was sufficiently challenging.

4.4.6.2 Sensor Fidelity

When asked which factors participants felt would most likely effect sensor fidelity (Table 4.13), the most often common was interference (5 participants) e.g. “the only sensible place to put it was a bit further away and there was nothing in between [the sensor and the door] to back a noise so I thought it would be ok” (F1) and “other noises, wind might be a bit of an issue if the window is open” (F5). This finding is inline with the Study 1. Of the other factors mentioned, only one participant mentioned distance from the sensing target: F3 “close to the window”; one participant suggested obstructions: F1 “so there’s nothing in the way”; and one participant connectivity: F2 “the signal to the box [base station]”.

4.4.6.3 Search for Connectivity

As we observed in the Study 1, all participants began their searches from the sensing target. Then if no connectivity was present, participants would begin their search, working away from the sensing target (not always in the direction of better connectivity).

Observations of participants search behaviours were again coded as (i) Continuous - where participants focussed on the feedback as they moved the sensor node through the space, (ii) Place & Test - where participants selected a location before checking the feedback to see if it was suitable - if the location was considered unsuitable, a new location was sought, and (iii) One Spot - where participants selected a location after contemplating the available options without consulting the feedback. Table 4.15 reports the frequency of these classifications. Continuous was most common (15 total occurrences), with Place & Test second (9 total occurrences) and no participants demonstrated a One-Spot search behaviour in this study.

Table 4.14) details participants' responses to interview questions where they were asked to describe their search strategy. Of the codes identified in Study 1, four participants described a strategy of following the feedback e.g. "started by looking at the signal strength and being close to the window" (F1); 3 participants randomly moving around e.g. "[i] just kept moving around" (F6); 1 participants calling on existing knowledge: F4 "I started as close as possible to the window and work backwards towards the [base station]"; And one participant suggesting elevation e.g. "I would consider putting the [sensor node] up higher" (F4). No new codes were identified.

4.4.6.4 Comprehension of radio signals

When asked to describe factors which the participants felt influenced or affected signal strength (Table 4.12), distance was commonly identified (4 participants) e.g. "distance really, the further away you get the worse it is" (F5) and "I'm guessing that's the distance" (F1). The second most common (3 participants) was obstructions such as "walls being in the way" (F6) and "I tried to avoid the walls and big things like the fridge" (F1). Interestingly, interference was only mentioned by one participant: F6 who said that "other electrical appliances" might be an issue. Two participants conflated distinct facets of the system, suggesting that the speed of their broadband may be a factor e.g. F2 "having better broadband signal may improve [connectivity]".

4.4.6.5 Placement motivation

When asked to describe the factors which motivated their choice of installation location (Table 4.10), signal was most common (6 participants) e.g. "I was thinking about signal first and foremost" (F4). Sensor fidelity was the second most frequently cited motivation (4 participants) e.g. F5 wanted to place the sensor for "optimal clarity in hearing the windows". Safety again proved an important consideration with 6 participants expressing concerns about damaging the device or harm to other e.g. F4 suggested "out of sight out of mind" was a good policy when children are around, or F8 who was

Table 4.10:
Participants' reported motivations

Codes	No. Participants
Connectivity	6
Sensor Fidelity	4
Safety	5
Aesthetics	1

Table 4.11:
Participants' prioritization of motives

Codes	No. Participants
Connectivity	6
Sensor Fidelity	1
Both Equally	1

Table 4.12:
Factors which affect connectivity

Codes	No. Participants
distance	4
obstructions	3
interference	1
broadband	2

Table 4.13:
Factors which affect sensor fidelity

Codes	No. Participants
distance	1
interference	5
obstructions	1
connectivity	1
sensitivity	1

Table 4.14:
Connectivity search strategies

Codes	No. Participants
Follow	4
Knowledge	1
Elevation	1
Random	3

Table 4.15:
Observed search behaviour

	No. Participants			Total
Room	1	2	3	
Continuous	5	5	5	15
Place & Test	3	3	3	9

concerned that they didn't "fall out the window". One participant (F4) also mentioned aesthetics, reporting that they should be "tucked behind the curtain so you cant see it".

When subsequently asked whether signal or sound sensing took priority (Table 4.11) the majority of participants reported signal as being the most important consideration e.g. "Without signal its not going to record anything" (F6) and "I was thinking about signal first and foremost" (F5). These results are also inline with the findings of the Study 1, where participants priorities connectivity over sensor fidelity.

4.4.6.6 What constitutes good signal

In this study, we also asked participants what they felt constituted good signal. Three participants described good signal as being one which was consistent i.e. the bars on the UI did not fluctuate e.g. "It's not as strong as it could be, but it's fairly consistent" (F1). When asked if this was more important than number of bars they explained that although the number of bars is important, the stability gave them confidence that it would work long term. F6 shared this idea, saying that good signal is "when the bar is stable, it's not flickering". A further three participants offered a more simplistic explanation, that the "more bars are better" (F3), "I would be concerned if it was one or two bars, but 50% seems reasonable" (F5).

4.4.7 Discussion

Participants of this study once again described the system as easy to use and did not report any problems. The variation in performance across tasks is inline with the results of Study 1 and supports that the experimental design is fit for purpose. Broadly speaking the findings of this field study support those of Study 1. All eight participants of this field study were able to find *suitable installation locations* despite the absence of connectivity in the immediate vicinity of the sensing target in the final task. This demonstrates (in more ecologically valid setting) that the Bar+Message feedback is an effective means of guiding users when setting up wireless sensors. Furthermore, it validates the **design implications** we proposed: that providing real-time connectivity and sensor fidelity feedback is important and should be made available whenever possible. Participants performance together with observations of participants responding to the suggested course of action also support our conclusion that messages such as these are helpful. Interestingly, we did not observe any participants exhibiting the "fluctuation fixation" behaviour we noted in Study 1, i.e. where small fluctuations in signal led to participants examining those location in greater detail. Given the varied nature of participants homes and limitations of our study design, to draw any firm conclusions as to why this is the case will require further examination. For example, it is possible that participants were more relaxed and less anxious in their own homes, a factor which can change the way people tackle challenges (Norman, 2004).

Participants of both studies reported misconceptions. In Study 1, three participants suggested that the quality of the radio connection between the sensor node and the base station would influence sensor fidelity. Similarly, two participants of this field study (F3, F5) suggested that the speed of their broadband might be detrimental to accuracy of the sound sensing. This further demonstrating the challenge of making such complex systems understandable to non technical users. On a different yet related note, one participant of the field study (F4) moved the tablet rather than the sensor to try and find signal. We are reminded, then, that the monitoring interface increases the chance of confusion, the tablet is an extra device with its own additional challenges. A new **design implication** then is that configuration interfaces need to take into account such additional complexity, and highlight the differences between what is *being monitored* (the sensor nodes and their custom radio connection to the base station) and what is *doing the monitoring* (the tablet).

Finally we also note that safety was also a concern. As per Study 1, participants of this study expressed concern for the protection of the devices and for the safety of occupants. This further validates the **design implication** highlighted in Study 1, that sensor nodes should not only be robust but also look robust.

4.5 Implications

Our user studies demonstrate that feedback is helpful to users setting up wireless sensors and that it plays an increasingly important role as radio connectivity becomes more scarce. Furthermore, we demonstrated that providing a suggested course of action is helpful. In Chapter 3 we showed that displaying concise messages informing users that action was needed were enough to persuade users to act. In this Chapter we show that short and simple messages can also be effective for suggesting a course of action. The simple text prompt we tested significantly improved users ability to find *suitable installation locations*. Given the relatively simplicity of implementing such messages this **design implication** can easily be adopted to improve users experience when installing wireless sensor nodes. In addition, showed that if the suggestion is informed by the current environment significant reductions in search time can be achieved. While the augmented reality interface we tested was effective, the design space is considerably larger and should be explored in future work.

We also note that participants of our studies highlighted concerns over the robustness of the prototype system. The physical fragility of the sensor nodes and fluctuations in the feedback both influenced participants search strategies and *suitable installation location* selection criteria. These **design implications** will inform the design of our prototypes and we hope will guide other researchers conducting studies in this space.

4.6 Limitations

Our goal was to examine users setting up wireless sensor networks. Designing an ecologically valid yet controllable study was a considerable challenge and as such we constrained the scope of our investigation so that meaningful comparisons could be drawn.

Extending prior work Costanza et al. (2010); Beckmann et al. (2004), in this paper we consider the constraints of both connectivity and sensor fidelity. To this end, our studies utilised sound sensors as a generic omni-directional sensor type. However, it should be acknowledged that other types of sensors (e.g. cameras, proximity, contact sensors) may introduce additional constraints increasing the difficulty of installation. For example, camera based sensors may need to be pointed in a specific direction, and a magnetic reed switch may need to be affixed to a door. In these cases, if no connectivity is present at the installation site, radio coverage would need to be extended, e.g. by introducing repeater nodes or moving the base station. Therefore, further work is needed to establish whether the feedback that we proposed and evaluated in this paper would be applicable to these situations, and how it may need to be adapted or extended.

In addition, the prototype sensor nodes developed for our studies utilised battery power.

While common to many commercial products and research prototypes, there are viable alternatives such as solar panels. Installers of devices powered in this way must then also consider the availability and reliability of light sources. In contrast, more "power hungry" devices may still require mains power outlets. Further exploration is needed to assess how our findings can be leveraged in this way.

4.7 Summary

The work reported in this chapter aims to address a challenge identified in prior literature Fischer et al. (2017); Kazmi et al. (2014), namely how best to support users in their search for *suitable installation locations* when setting up networks of wireless sensor nodes. To this end, we designed and developed two user studies, the first conducted under controlled conditions in an office environment and the second in participants homes.

Our findings indicate that: (i) feedback plays an important role in users' search for *suitable installation locations* when setting up wireless sensor systems, (ii) that a textual message suggesting a course of action are significantly more effective than providing real-time connectivity information alone, and (iii) that providing a spatially informed suggested course of action (i.e. using AR) significantly expediates the search for *suitable installation locations*, but overall does not improve users' ability to identify them. Building on these results, the field study evaluates the most effective and easy to implement feedback condition (Bar+Message) in users' homes. The findings of this second study: (i) confirm that providing a textual message suggesting a course of action is an affect approach even in less constrained environments, (ii) reinforces the qualitative results and discussion, and (iii) gives further insight into users' understanding of wireless sensors, the factors they considered when selecting installation locations and participants' motivations.

In the studies we describe in this chapter, participants were shown a primitive measure sensor fidelity, i.e. how well the sound sensor could "hear" its target. While this was sufficient to constrain participants in our studies, we do not consider how this sensor data will be used. Sensor based systems commonly employ sophisticated detection processes to generate useful outcomes. In the next chapter, we take a step further, redefine what is a *suitable installation location* to meet the input requirements of a dependant detection process.

Chapter 5

Feedback for Computer Vision Based Smart Systems

The data collected by sensors is rarely presented back to users in its raw state. It is much more common for sensor data to provide the background information on which a parent system operates. Take home automation, for example. A smart CCTV camera might analyse the faces of people walking towards the house it protects, so that when the home owner's face is recognised the door will unlock (example taken from (Norman, 2010)). Pattern recognition technologies such as this, are increasingly common (Yang et al., 2018). For any given pattern recognition system, its efficacy depends on how well the inputted material meets the system's requirements. Returning to our example, the likelihood that the face recognition will succeed is largely dependent on how the camera is installed i.e. is it pointing in the correct direction and is there a sufficient level of detail present? The challenge then, is to ensure that smart systems which employ pattern recognition technologies are provided with the most suitable input.

The inclusion of high resolution cameras in modern smartphones and tablet computers has resulted in a deluge of apps utilising pattern recognition, and thus it is one of the most likely places where an everyday user will encounter such technologies. For example, as we mention in the introduction of this thesis, Amazon's mobile shopping app allows users to search their product catalogue using images captured with the host device's camera. While a user of such an application will likely have a good understanding of how to achieve the task of capturing an image of an artifact, they are less likely to have specific knowledge which will help them reason about the most suitable input for the underlying pattern recognition, knowledge which is of particular importance when experiencing difficulties. Factors such as technical limitations (e.g. limited training datasets), environmental challenges (e.g. lighting conditions and shadows), image composition (e.g. "noisy backgrounds" and camera focus) and unrealistic user expectations (Yang and Newman, 2013) can all negatively impact user experience - making it difficult

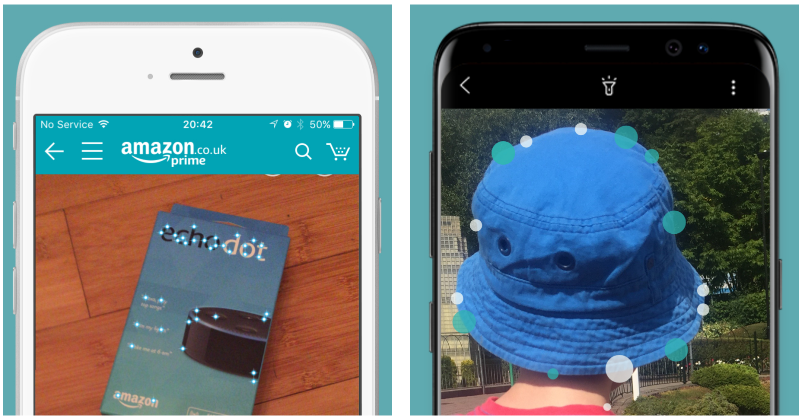


FIGURE 5.1: Smart Camera Apps that display keypoint markers feedback to users: left, Amazon and right, Samsung's Bixby.

to reason about what constitutes a suitable input and what is the best course of action when faced with unexpected outcomes. There is therefore a need to support users in their understanding of system behaviour, so that they can better overcome failures and provide the most suitable input. To this end, designers must find ways of making the reasons for failures intelligible, without requiring users to become experts in pattern recognition.

Perhaps it is to address these challenges that a number of commercial smart camera apps include visual feedback, overlaying the camera's viewfinder with visual aids. Two notable examples are the aforementioned Amazon app's "search by image" feature and Samsung's Bixby, a camera-based search tool[1] (Figure 5.1). Both display feedback in the form of "keypoint markers" - coloured dot visualisations which correspond to features of interest identified by an underlying algorithm. While such visualizations have long been popular as a debugging tool for software developers[2], to date little is known about their effect on end-user interactions. Their inclusion may simply be motivated by a need to convey background activity, however, their presence raises some interesting questions: (i) are they intelligible to lay users? (ii) do they improve usability and aid users' interaction around failures? and conversely (iii) can they mislead users if misunderstood?

In this chapter we report our initial investigations of keypoint marker feedback[3], examining its intelligibility (question 1). In the following chapter (Chapter 6) we build on this work and address questions 2 and 3.

[1] which tries to find matching images from an internet search
[2] e.g. OpenCV https://goo.gl/bX4XEM
[3] Ethics approval granted by the University of Southampton (ref: 27198).

5.1 Study Design

This lab study was designed to assess the effectiveness of keypoint markers as a feedback visualisation. Specifically, whether keypoint markers would make it easier for users to estimate the outcome of an object recognition algorithm.

Participants were asked to examine a series of 44 image pairs. As illustrated in Figure 5.2, each pair comprised of a "reference" image and a "captured" image. The *reference image* depicts an object of interest which is also present in the *captured image*. Participants were asked to estimate whether the object recognition algorithm would be successful in identifying the object from the *reference image* in the *captured image*. All images had previously been processed using the object recognition algorithm and in half (i.e. 22) the object recognition algorithm was successful and in the remaining 50%, it was not. Figure 5.2 demonstrates also how this task was presented to participants via a computer based survey.

5.1.1 Setup and Procedure

The study took place in a neutral private University office space. Participants were asked to sit at a desk where they completed a survey using a 13-inch laptop. At the start of the study participants received written instruction, which (i) provided context to the experiment, describing a consumer application which uses computer vision and explaining that the way humans and computers "see" images is not necessarily the same, (ii) stated that the aim of the study was to examine how feeding back information about how computer vision algorithms "see" can help users use them more effectively, (iii) detailed the task procedure, and (iv) explained the operation of the survey application

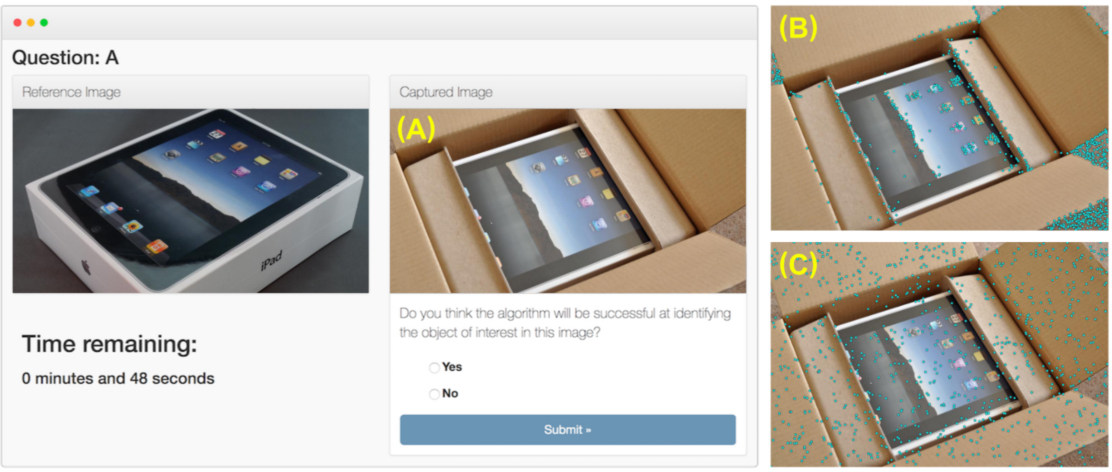
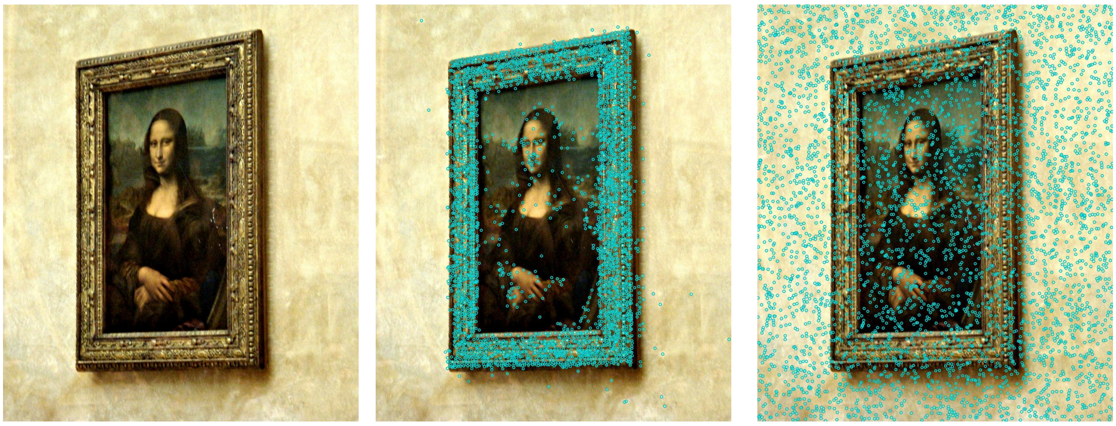


FIGURE 5.2: Screenshot of survey app used in Study 1 and examples of the feedback conditions: A) No-Feedback, B) Keypoint-Markers and C) Random-Markers



(A) No-Feedback (B) Keypoint-Markers (C) Random-Markers

FIGURE 5.3: Examples of feedback shown to participants.

interface. After reading the instructions, participants were asked if they had any questions, before being instructed to begin the survey. The survey application recorded their answers and calculated their selection-accuracy in relation to the known outcome of the algorithm. Pilot studies and practice runs led us to estimate that the study duration was approximately 30 minutes.

5.1.2 Images

All images were sourced from Flickr[4] and selected so to minimise the chance to cause upset or offense to participants.

5.1.3 Data Collection

Selection-accuracy was the key measure of the study. However, at the end of the survey we conducted a semi-structured interview to better understand participants rationale when making judgments. Audio recorded were made throughout the study and the experimenter collected written notes when relevant.

5.1.4 Participants

18 participants (12F, 6M) were recruited from the university participants pool which includes university staff, students and the general public alike. Each participant received £10 for their time. Anyone above 18 years of age who expressed interest was included in the study, so long as they did not have a technical background (i.e. no degree in

[4] https://www.flickr.com/ - Only images published under a Creative Commons license https://creativecommons.org/licenses/ (See acknowledgements for attributions)

computing or engineering and not in technical employment) and had normal or corrected to normal vision. Figure 5.1 details the age ranges of participants.

5.1.5 Conditions

In this study we examined 3 conditions:

- **No-Feedback** Images were displayed without any form of visual feedback. This condition constituted the baseline. (Figure 5.3a).
- **Keypoint-Markers**: Images were overlaid with markers which highlighted the keypoints of interest identified by the object recognition algorithm (Figure 5.3b).
- **Random-Markers**: Images were overlaid with the same number of markers as seen in the Keypoint-Markers condition, however their positions were randomised. This condition was designed to test if the presence of markers alone (not necessarily related to how the system works) would draw more attention to the image and thus improve user performance (Figure 5.3c).

Each participant was assigned two of the three conditions in fully counterbalanced order. This exposed each condition to 12 of the participants, 6 times as the first condition and 6 times as the second.

5.2 Quantitative Findings

Figure 5.4 present the number of correct answers for each condition, when shown to participants as the first of the two conditions, in the order they were asked. Figure 5.5 present the same information when the condition was shown to participants as the second of the two conditions. The results for participants selection-accuracy are reported in Table 5.2. A two-way ANOVA test revealed no significant effects on participants selection-accuracy from the different feedback condition (F=1.254, p=0.300), the task order (F=1.301, p=0.263), nor any interaction of the two (F=0.121, p=0.886).

TABLE 5.1: Age range and background of participants across all 3 studies.

Age Range	50-59	40-49	30-39	20-29	18-19
Count	1	5	2	3	1

TABLE 5.2: Participants' selection-accuracy (correct answers per condition)

Condition	First Condition	Second Condition
No-Feedback	61.36%	56.82%
Keypoint-Markers	65.91%	64.39%
Random-Markers	66.67%	61.36%

5.3 Qualitative Findings

The qualitative data collected during interview was analysed independently by three researchers and then grouped by consensus. Codes were initially drawn from research questions and then supplemented with those that emerged from the interviews before being grouped by consensus. Here we note that no participants reported any issues using the experimental apparatus.

In this section we refer to participants by their condition followed by their index. Because each participant was subject to two of the three possible conditions, there are 6 possible prefixes. For example, NK1 refers to a participant number 1 who was subject to the No-Feedback condition first and the Keypoint-Markers condition second. Below we list the six prefixes:

- NK - No-Feedback first, Keypoint-Markers second
- NR - No-Feedback first, Random-Markers second
- RK - Random-Markers first, Keypoint-Markers second
- RN - Random-Markers first, No-Feedback second
- SN - Keypoint-Markers first, No-Feedback second
- SR - Keypoint-Markers first, Random-Markers second

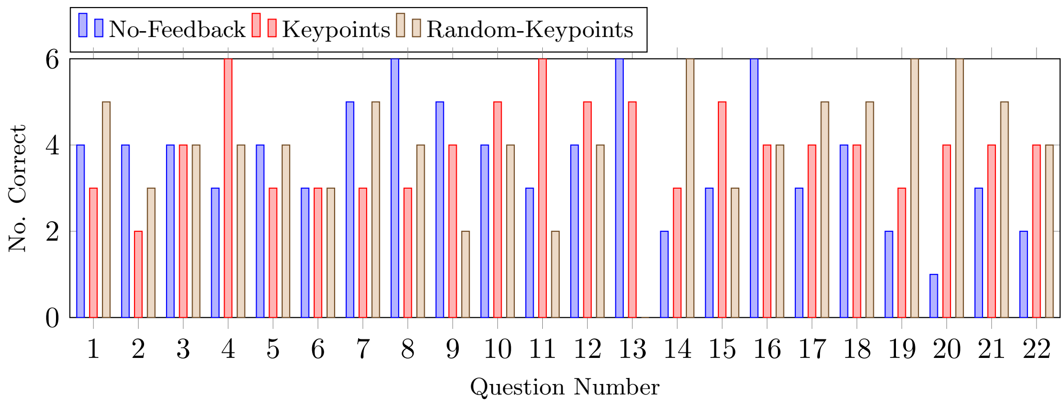


FIGURE 5.4: Number of correct answers - as first condition

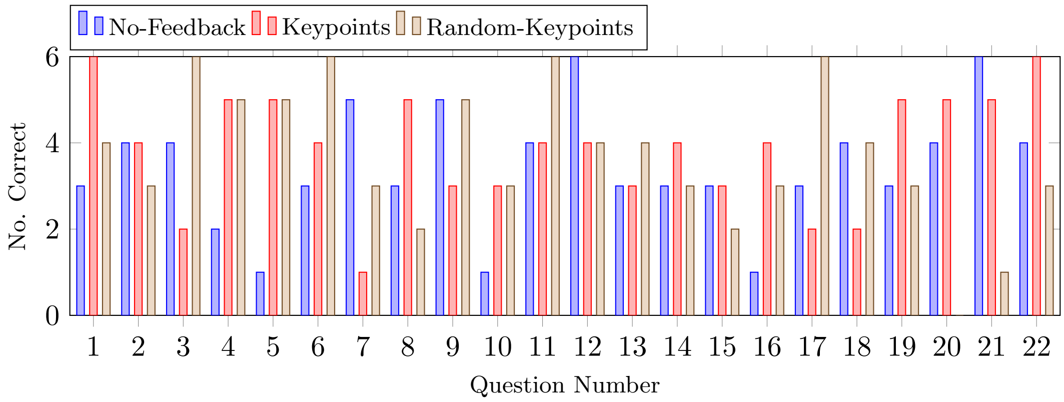


FIGURE 5.5: Number of correct answers - when second condition

TABLE 5.3: Factors in decision making.

	NK	NR	RK	RN	KN	KR	Total
Composition	3	2	2	3	3	3	16
Dominance	0	2	3	3	3	2	13
Learn	1	2	0	0	1	0	4
Feedback	1	0	1	0	0	1	3
Features	1	0	0	2	0	0	3
Colour	0	1	1	0	0	0	2

5.3.0.1 Factors in decision making

The *composition* of images was the most commonly reported factor participants took into consideration when estimating the success of the algorithm (16 participants). Responses coded as *composition* included words such as “cropping”, “orientation” and “angle”. The next most common factor was how *prominent* the object of interest was in the image, reported by 13 participants. NK1 for example said that the “clarify of the object” was important and RN10 said that the “whole image should be there” i.e. not cropped and in frame. *Learning* from experience was reported by 4 participants e.g. NR4 who “learnt [over] the time”. *Feedback* was reported by 3 participants e.g. RK9 who said that the “dots also has an impact”. *Features* such as shape by 3 participants, and finally *colour* was mentioned by 2 participants. Table 5.3 details the distribution of codes across conditions.

TABLE 5.4: Participants’ understanding of feedback

	NK	NR	RK	RN	KN	KR	Total
Correct Understanding	2	0	1	0	0	1	4
Partial Understanding	1	1	0	0	0	2	4
Incorrect Understanding	0	2	2	3	1	0	8
Opposite Understanding	0	0	2	0	0	0	2

TABLE 5.5: Participants who noticed feedback

	NK	NR	RK	RN	KN	KR	Total
Yes	3	3	2	3	1	3	15
No	0	0	1	0	2	0	3

TABLE 5.6: Participants' preferred condition

	NK	NR	RK	RN	KN	KR	Total (Max 12)
No-Feedback	0	3	-	3	1	-	7
Keypoint-Markers	3	0	0	-	0	0	3
Random-Markers	-	-	2	0	-	1	3
No Preference	0	0	1	0	2	2	5

5.3.0.2 Did participants consider feedback

Fifteen participants stated that they were aware of the feedback (Table 5.5), three of whom said they were helpful and reported them as a factor in their decision making (Table 5.3). When asked which condition they preferred i.e. found most helpful (see Table 5.6), No-Feedback was most popular (7 participants), followed by Keypoint-Markers and Random-Markers in joint second (3 participants each). A further 5 participants expressed no preference.

5.3.0.3 What keypoint markers represent

Four participants were able to provide a high level explanation of the keypoint markers meaning (Table 5.4), with a further 4 demonstrating a partial understanding i.e. they were able to relate the markers to the algorithm e.g. NR5 "reference points to the algorithm" but were not sure how to use them. In contrast, eight participants demonstrated a flawed understanding e.g. RN11 "not sure whether it's for me or for the computer', two of whom (RK7 and RK8) believed that the markers highlighted regions where the algorithm was having difficulty e.g. "concentrated in some part of an image is not a good sign." (RK7).

5.4 Discussion

Contrary to our expectation, there was no significant differences between participants' performance across conditions. The results (as detailed in Table 5.2) show only a very small difference between the number of correct answers given by participants in the No-Feedback and Keypoint-Markers conditions. The most sizable difference represents a difference of only 3 correct responses. Nor can a pattern be discerned from

Tables 5.4 and 5.5. Qualitative analysis supports these findings, with only 3 participants reporting feedback to be a motivating factor in decision making, relying more on factors such as how prevalent the object of interest is within the image (Table 5.3). Participants predominantly reported themselves to be motivated by the overall composition of the images (i.e. how aesthetically pleasing it is) and the dominance of the object within the composition. This finding is suggests that feedback is not necessary for applications where the user is required to capture a specific object e.g. Amazon's search by image functionality. Given that only 4 participants demonstrated a correct understanding of the keypoint marker feedback (Table 5.4) the potential for this kind of feedback to be detrimental to user interaction is may outweigh any benefit. Indeed, 2 participants developed understandings which were contrary to the algorithms true operation, highlights the risks. However, both of these participants were shown Random-Keypoints followed by Keypoints, which potentially affected their capacity to develop a correct understanding. On reflection, the study design may have inhibited participants ability to reason about keypoint marker feedback and the lack of interactivity may have prevented experimentation. In the next section we discuss these limitations further, the lessons learnt and outline the implications for our future studies.

5.5 Implications & Limitations

Asking participants to make a judgement based on still images may have inhibited their ability to theorise and experiment. A real word application of this algorithm would most likely be implemented as part of the image capture. In this context, referring back to Norman (2013)'s seven stages of action model (see Chapter 2), the keypoint marker feedback is a response to the act of moving the camera. The capacity of feedback to inform a users understanding is inherently linked to dynamic interaction. While the findings of this study show that keypoint markers are not effective for still images, how this applies to smart systems is unclear and further work is needed where interactivity is an intrinsic part of the study design.

While it was not the intention of this work to qualitatively elicit and assess user understanding, the exposure of multiple feedback conditions to each participant highlights the limitations of within-group comparisons. Consider participants of the KR (Keypoin-Markers first, Random-Markers second) group, participants may have begun to develop a correct understanding during the first condition, but then the theories they developed would not have fitted with the Random-Markers feedback shown in the second half of the study.

5.6 Summary

In this study three visual feedback approaches were tested in a controlled lab study (n=18). Using a counterbalanced distribution, each condition was presented to 12 participants, 6 times as the first condition and 6 times as the second. Our results showed no significant difference between conditions. However, the qualitative data suggests that participants overlooked the feedback and relied primarily on their experience of previous questions. Reflecting on this finding, we speculate that the lack of interaction in the study design inhibited participants capacity to experiment. To address this, in the next chapter we report a follow-up study centered around an interactive task.

Chapter 6

Feedback Derived from Different Stages of Processing

In this chapter, we build on the work of the Chapter 5, which investigated keypoint markers, a common debugging visualizations which are now being presented to users in mainstream consumer applications (e.g. Amazon smartphone app and Samsung's Bixby app). We concluded the previous chapter, reporting that our findings showed, contrary to our expectations, that keypoint marker feedback was ineffective. However, reflecting on the experimental design, a possible failing was identified i.e. that the still images may have inhibited participants ability to theorise and experiment about the meaning of the markers. In this chapter, we report two user studies[1] designed to overcome this limitation and in so doing address whether keypoint maker feedback: (i) is intelligible to users with no experience of how pattern matching technologies function? (ii) can improve usability and aid these users' interaction around failures? and conversely (iii) mislead these users if misunderstood?

Chapter 5 demonstrates how developing a controlled yet ecologically valid study is challenging. Such studies require observations of user interaction at the boundaries of success and failure. The experimental task must also be controllable and repeatable, but in a way which is not obvious to participants. Moreover, the task needs to be engaging and enjoyable to motivate participants, have a clear goal and provide discussion points. To address these issues, we present a novel experimental lab study design enabled by a novel smart camera app that we developed. By so doing, we aim to make a methodological contribution to HCI. Leveraging this experimental design and the novel smart camera app, we conducted a series of between-groups studies. We begin this chapter by describing how pattern recognition has been used in commercial applications, before reporting the individual studies and finally discussing the findings in relation to prior work.

[1]Ethics approval granted by the University of Southampton (ref: 27198)

6.1 Pattern Recognition in Apps

The keypoint marker feedback seen in many consumer applications is commonly derived from a feature matching algorithm, an intrinsic part of many smart camera apps, e.g. panorama stitching, object detection, gesture recognition and motion tracking. Most keypoint matching algorithms involve three stages of processing: (i) identify distinctive points of interest in an image (the keypoints), (ii) programmatically describe them, so that the description is resilient to geometric variations e.g. rotation, scale and perspective, and photometric variants e.g. contrast and brightness, and (iii) compare the descriptions with those of another image. How the results of this comparison process are used is application specific. In panorama stitching for example, the closest matching descriptions between images are assumed to represent the same point in the physical world. Using their relative changes in position the images can be transformed such that the keypoints overlap creating a new combined image with a wider field of view.

6.2 Study 1

We designed and conducted a between-groups study. This study began by examining two conditions, comparing keypoint markers with no feedback. Through a combination of quantitative and qualitative methods the results revealed that participants overwhelmingly misinterpreted the meaning of keypoint marker feedback. Participants interpreted them as indicating high level algorithmic explanations (e.g. about recognized objects), while in reality they refer to low-level features of the image (e.g. pixels). To better understand this finding, 20 new participants were exposed to two additional conditions designed around feedback that is actually related to higher level algorithmic explanations. More formally, this second iteration of Study 1 addresses a fourth question: does the processing stage (lower level vs higher level) from where the feedback is derived impact user understanding? For brevity we report these conditions together.

6.2.1 Study Design

Developing an ecologically valid and testable experimental task which incorporated a keypoint matching algorithm proved non-trivial. The task needed to provide sustained exposure to algorithmic feedback so that participants could observe and reason about the feedback. Further to this, participants must experience instances of failure and success. The task therefore should be controllable, but in a way that is not obvious to participants. In addition, it would be advantageous for the task to be enjoyable to motivate interaction, have a clear goal and provide discussion points. Through experimentation a task which best satisfied these criteria was developed, the creation of stop-motion animation.

To create a stop-motion animation, an animator must capture a series of still images (frames) of a given scene. By incrementally moving artifacts (characters) between frames the illusion of animation can be achieved i.e. when the frames are played back in order the characters appear to move autonomously in relation to the static elements of the scene (e.g. the background). Figure 6.1 demonstrates this process:

1. Set up the background scene with the character in its starting position and hold the tablet such that the camera's viewfinder encapsulates the scene and the character.
2. Capture a frame.
3. Place the tablet aside and manipulate the character in some way e.g. reposition or rotate.
4. Reposition the tablet and capture another frame.
5. Preview the captured frames / playback the animation. If the result is not acceptable then the frame can be deleted at this stage (or at any time later).
6. Repeat stages 3 to 5 until the animation is complete.

Traditionally stop motion animations are created using cameras where the position and angle are strictly controlled e.g. held in a tripod. To incorporate pattern recognition technologies in to our study design we replaced the controlled camera with a handheld tablet computer and bespoke app (Anim8[2]) which employs a keypoint matching algo-

[2]For more information about the Anim8 app visit: http://anim8.space/

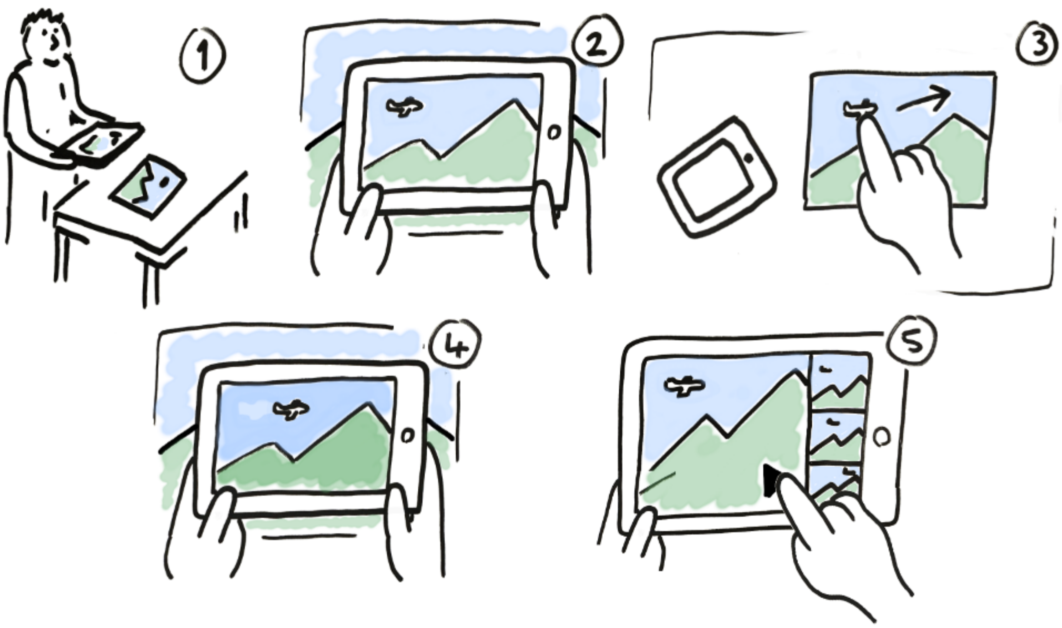


FIGURE 6.1: Creating an animation with Anim8

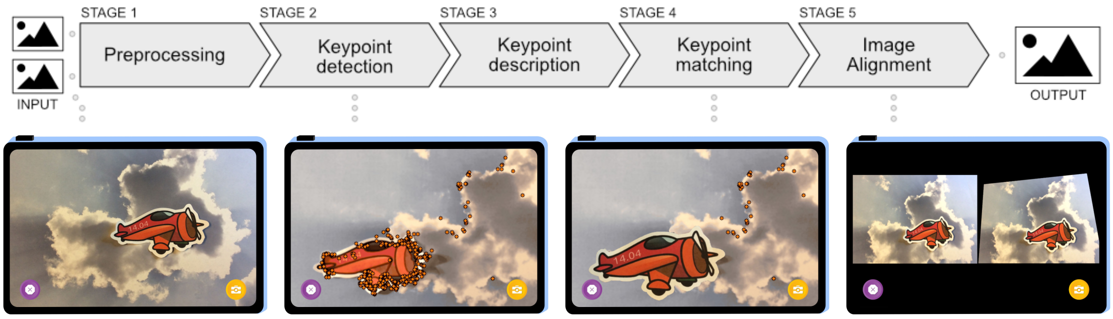
rithm[3] to align each frame to its predecessor - a process of stabilization. This process makes all frames appear to have been captured from the same physical location even though the camera's position and angle vary. The keypoints with the closest descriptions are matched and assumed to point to the same physical feature in both frames. The most recently captured image can then be transformed so that its keypoints overlap its predecessors. Characters which have been moved between frames will create erroneous mappings, however if enough matches are found for the elements of the scene which have remained static (e.g. the background) then the matches associated with the moving characters will be treated as outliers and ignored.

In order for the stabilization process to work effectively it is critical that the static elements of the scene are "feature rich", i.e. the algorithm can identify many keypoints. If there are too few then the transformation process may output an image where the background is distorted and the character remains stationary (Figure 6.2). Leveraging this limitation, the likelihood of whether the stabilization process will succeed or fail can be controlled - by providing "feature rich" and "feature poor" backgrounds participants of the study can be exposed to situations where the stabilization process succeeds and fails respectively. Factors such as lighting conditions, shadows and camera angle makes this form of manipulation not immediately obvious to study participants.

Through pilot studies we concluded that four animation tasks with 4 to 5 frames per task provides sufficient exposure. We designed the tasks to assess whether feedback derived from the stabilization process can help participants develop a better understanding of the systems' needs. To create discussion points and elicit user understanding we ask participants to choose one of three background options in the last two animation tasks (3 options per task). The feature richness of the three background options varied and thus the likelihood of the stabilization process succeeding varied (Figure 6.6).

[3] Through experimentation the ORB algorithm Rublee et al. (2011) proved to offer the best compromise of performance, speed and control for our study.

FIGURE 6.2: When too few matching keypoints are identified in the background, the stabilization process can result in an image transformed such that the character appears to remain stationary and the background becomes distorted.



(A) No-Feedback (B) Keypoints (C) Matching-Keypnts (D) Split-Screen

FIGURE 6.3: Examples of the feedback conditions presented by the Anim8 application and their relationship to the processing pipeline (a, b, c, d). Also the preview interface (e). Note: To see these images animate see supplemental materials.

6.2.2 Conditions

To explain the study conditions, we describe them in relation to the computer vision pipeline employed by Anim8 (Figure 6.3). It should be noted that we did not explain the feedback nor point out its presence to participants. This was done to mirror the experiences of current consumer smart camera app users.

No-Feedback (Figure 6.3a) This condition was included as a baseline. The input images to the pipeline were presented back to participants without any additional feedback.

Keypoints (Figure 6.3b) The camera's viewfinder was augmented with keypoint markers which indicate the locations at which keypoints had been detected in stage 2. It is important to note that not all the identified keypoints will be matched. Matches where the descriptions are considered too dissimilar are deemed outliers and are ignored by the stabilization process. Despite this, the location, distribution and volume of identified keypoints are good indicators for the potential success of the stabilization process.

Matching-Keypoints (Figure 6.3c) Again the viewfinder was augmented with keypoint markers, however in this case only those which have been successfully paired with keypoints in the previous frame were displayed (Stage 4).

Split-Screen (Figure 6.3d) This condition represents the final stage of processing. The viewfinder was divided into two equal halves. On the left: the input image updated in real-time (as per No-Feedback condition). On the right: the image outputted by the processing pipeline (update every ˜120ms).

The No-Feedback and Keypoints conditions were compared first, while the Matching-Keypoints and Split-Screen conditions were included at a later stage, as described in the Introduction.

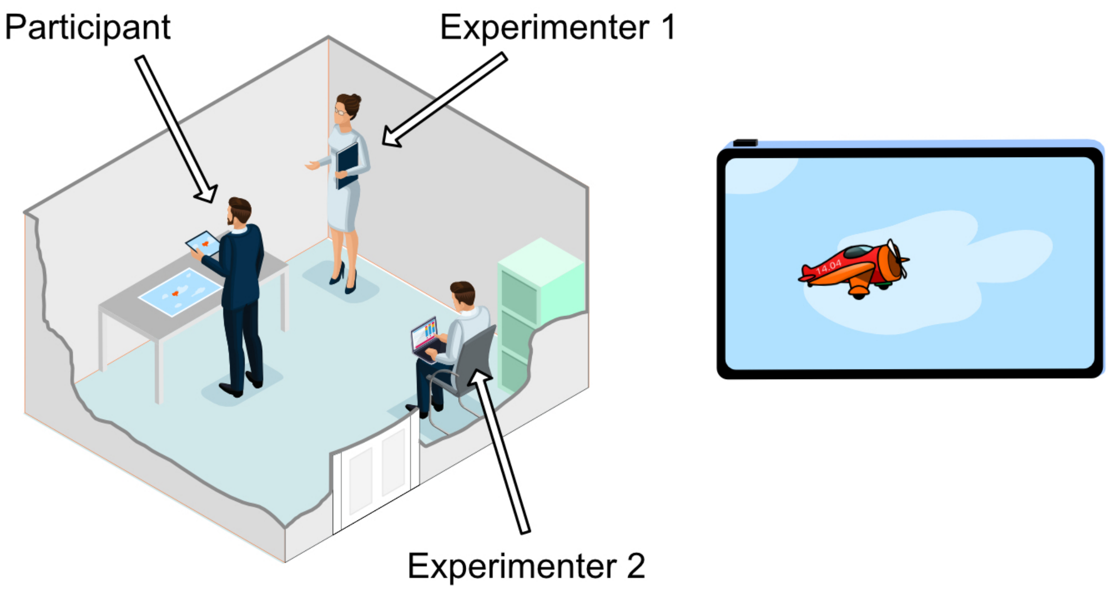


FIGURE 6.4: Diagram of experimental setup

6.2.3 Procedure

All studies were conducted in the same empty windowless meeting room (so lighting conditions could be controlled) on a university campus. Two experimenters were present at all times - one to conduct the experiment and the other to observe, take notes and make audio recordings.

At the start of the study participants received written instructions detailing: (i) the procedure necessary to create stop-motion animations, (ii) how Anim8 uses computer vision technologies to remove the need for a tripod, and (iii) a high level explanation of the image processing operations - that Anim8 tries to align images "by looking for things in each image which are not supposed to have moved, for example the background". After reading the instructions participants were asked to stand up while performing the animation tasks.

Participants were tasked with creating 4 stop-motion animations. Animating a two dimensional cardboard character (approximately 8cm by 5cm in size) moving across an A3 printed background (see Figure 6.5 for examples). To ensure that all participants had a good understanding of how to use the Anim8 application, the experimenter demonstrated the capture, playback and delete operations prior to the first task commencing. Whilst demonstrating the capturing of a frame, the participants were advised to ensure the printed background scene was fully encapsulated in the camera's viewfinder and that the desk should not be visible. This was done to prevent features other than those in the scene impacting the outcome of the experiment (this was not explained to the participant). The participants were also advised that if they needed any assistance regarding

(A) Task 1 (B) Task 2 (C) Task 3 (D) Task 4

FIGURE 6.5: Example frame for each of the animation tasks.

the operation of the application during the study, then they could ask at any time.

Prior to each animation task, the experimenter provided each participant with the necessary materials (i.e. a character to animate and static background scene / scenes) and an instruction sheet detailing an example path for the character to follow, along with the number of frames expected (4 to 5). On completion of the task, the participant was asked to play back the animation they had created to the experimenter. The tasks were conducted in the same order for all participants to ensure that they experienced both successful and unsuccessful attempts. The tasks were structured as follows:

6.2.3.1 Task 1

Task 1 was designed to allow participants to familiarise themselves with the UI and reassure them that the app works as described. To this end, a feature rich background (Figure 6.5a) which proved in testing to work with almost no failures was provided, making the task easy to succeed. On completion, the experimenter asked how the participants found using the app and if they had any queries.

6.2.3.2 Task 2

Task 2 was designed to highlight the limitations of the system. The background in this task (Figure 6.5b) proved in testing to always fail. As it was impossible to complete this task, the experimenter would intervene after a time limit of 2 minutes, if the participant had not already raised concerns. The experimenter would ask the participants to explain what was happening and if they knew why it did not work, before suggesting that they proceed to the next task for brevity.

6.2.3.3 Task 3

Task 3 was designed to assess users' understanding and create a point of discussion in the interview. Participants were asked to choose the background they felt would work best for the app from a selection of 3 backgrounds (see Figure 6.6). Participants were advised that they could preview them through the application's viewfinder if they wished.

FIGURE 6.6: Background options presented to participants in Tasks 3 (Top row) and Task 4 (Bottom row). Left: Likely to fail, Center and Right: Likely to succeed.

The backgrounds offered had previously been assessed and ranked according to the algorithm's ability to effectively identify features within them. One of the backgrounds consistently failed in testing and the remaining two consistently worked well, although one was more visibly "feature rich" than the other. The motivation for presenting users with this range of background options was to make the different levels of detail between the backgrounds less obvious. Once the participant completed this animation task, they were asked why they had selected that specific background.

6.2.3.4 Task 4

Task 4 followed the same structure as Task 3, with a new character and set of 3 backgrounds (see Figure 6.6). This last task was designed to sustain participant interaction with the application, collect an additional data point and further assess user understanding i.e. what, if anything, had been learned in Task 3.

At the end of the study a semi-structured interview was conducted. The interview began by asking participants if their experience in Task 3 and Task 4 had given them a better understanding of why the animation in Task 2 resulted in failure. Using this as a starting point, the experimenter asked further questions to assess the participants' understanding of the algorithm and their motivations for selecting the backgrounds in Task 3 and Task 4. For the participants of conditions where feedback was presented in the viewfinder, the experimenter also asked what they thought it represented and if they used it in their decision making.

6.2.4 Participants

We recruited 40 participants (15F, 25M) from the university participant pool which includes university staff, students and the general public. Anyone who expressed interest was allowed to participate in the study, so long as they did not identify as having technical hobbies or interests (e.g. computer programming), were not in technical employment (e.g. lab assistant) and were not technically educated (e.g. no degree in computing or engineering related subjects). Participants were also required to have normal or corrected to normal vision. Each participant received a £10 payment for their participation. Of the 40 participants 29 reported to be in education and 11 in full time employment. Participants' backgrounds were diverse with the most common being Business & Economics (13) followed by Social Sciences (9) Law (5), Languages (5), Art (4), Accountancy (2), Medicine (1) and Geography (1). One participant was aged between 40 and 49 years, 6 between 30-39 and 33 between 20-29. For more detailed information please see the supplemental materials.

Ten participants were randomly assigned to each condition. For conciseness, we will refer to participants by condition and subject number, for example, K7 was subject number 7 of the Keypoints condition. Prefixes "N", "M" and "S" refer to the No-Feedback, Matching-Keypoint and Split Screen conditions respectively.

6.2.5 Data Collection

Detailed notes were recorded on paper, using forms which were designed during pilot studies. These forms provided a framework to ensure that all the necessary data was collected in an efficient manner. The forms also provided space for unexpected observations to be logged. In addition, audio recordings we recorded for transcriptions and analysis.

6.2.6 Quantitative Findings

To quantitatively assess the effect of feedback across the conditions, three researchers independently coded participants' responses to questions pertaining to their background selections (taken from researcher notes and transcripts of audio recordings). This coding process was specifically focussed on the participants' understanding of how the system works (in contrast, in the next section we report a further analysis of the data through broader, more general coding). In particular, a participant's response was coded as "correct understanding" if they described how the presence of distinctive shapes and features in the background positively impacted the app's ability to align frames. For example, the following statements were coded as demonstrating a correct understanding: "I think it picks up the shapes on the picture and it [...] then compares the position of the dots on the other one [...] the next picture? So it can tilt the frame accordingly" (K9) or "because the background is distinct enough" (N6). If a participant reported motives not connected to the requirements of the app or their understanding of what is significant was incorrect they were coded as "incorrect understanding". For example, the

Table 6.1: No. Participants who selected a "correct background".

Condition	Task 3	Task 4
No-Feedback	10	10
Keypoints	7	10
Matching-Keypoints	10	10
Split-Screen	9	10

following statements were coded as demonstrating an incorrect understanding: "Because it's nice and colourful" (N8) or "[...]it looked more homogenous than the other ones. So I thought [...] it would be easier to take the photos like this" (K2).

Table 6.1 summarizes the background selections made by participants in Task 3 and Task 4 and Figure 6.7 shows whether their selection was based on a correct understanding of the stabilization processes.

To compare participants' understanding between the conditions we consider the total number of answers which demonstrated a correct understanding in Task 3 and Task 4 (Figure 6.7). For example, 7 of the 10 participants in the Split-Screen condition demonstrated a correct understanding in Task 3 and 9 participants in Task 4, giving a summed value of 16. A chi-square test of the summed values revealed a statistically significant difference (chi-square=8.33, p=.040, df=3, Cramer's V=0.323). To better understand the differences between the conditions, we analysed the chi-squared standardized residuals (presented in Table 6.2). It can be noticed that the standardized residuals are larger (in absolute value) for the Keypoints and Split-Screen conditions, suggesting that these two conditions explain the significance of the chi-square test. A chi-square test also shows no statistically significant differences for correct background selections (chi-square=6.316,p=.097,df=3), nor when testing the tasks individually[4]. It should be noted that participants sometimes selected a 'feature-rich' background for aesthetic reasons rather than because it would make the app work better (as instructed), failing to demonstrate correct understanding. In the next section we discuss our qualitative findings and the role of background selection further.

[4]understanding on Task 3: chi-square=3.509, p=.320, df=3, Cramer's V=0.296; understanding on Task 4: chi-square=5.812, p=.121, df=3, Cramer's V=0.381; correct selections on Task 3: chi-square=6.667,p=.083,df=3; all selections were correct in Task 4, so no statistical test needed

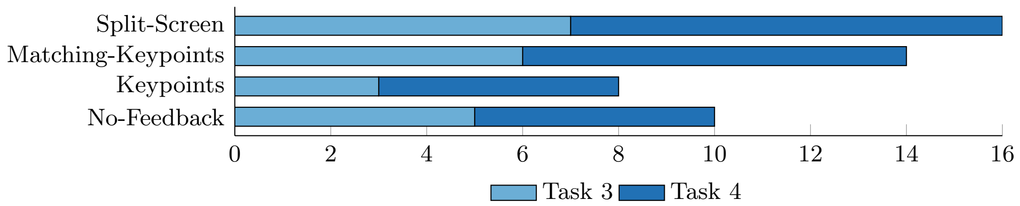


Figure 6.7: No. Participant responses coded as "correct understanding" when reporting their motivation for background selection in task 3 and task 4.

Table 6.2: Standard residual results of the No. participants who demonstrated a "correct understanding".

	Count	Expected	Std Residual
No-Feedback	10	12	-0.6
Keypoints	8	12	-1.2
Matching-Keypoints	14	12	0.6
Split-Screen	16	12	1.2

6.2.7 Qualitative Findings

Transcripts of all audio recordings and researchers' notes collected during the studies were independently coded by three researchers in a second round of analysis. Codes were initially drawn from research questions and then supplemented with those that emerged from the interviews before being grouped by consensus. In the subsequent subsections we detail these groups and give example quotations. First however, we would like to note that overwhelmingly participants reported the task to be interesting and entertaining. This suggests that the experimental task was sufficiently engaging and participants were invested in creating animations successfully.

6.2.7.1 Participants drew from their existing knowledge

When asked about previous experience with computer vision applications, participants mentioned QR Code scanning, Facebook and Instagram (none of which provide visual feedback). No participants reported using Amazon or Bixby's search by image, or any other application which provides keypoint feedback. In the No-Feedback condition, half of the participants demonstrated a correct understanding. These participants explained that having elements in the background which were "more detailed" (N1), "most defined" (N7), "distinct" (N6) or "prominent" (N2) would help the app because they were good reference points for alignment. The remaining five participants had an incorrect understanding and in the main focussed on the aesthetics, e.g "I thought the clouds would go really well with [...] the hot air balloon" (N9). Interestingly, participants in the No-Feedback condition selected a correct background more often than participants in the Keypoints condition (Table 6.1). Participants K2, K4 and K8 of the Keypoints condition made associations between the keypoint markers and their experience of other applications, suggesting that the keypoint markers functioned in much the same way as the autofocus on digital cameras, in that they highlight regions on which the camera is focussing. Whether these analogies are helpful is not clear. One of the participants who drew such parallels made good choices when selecting backgrounds, while the remaining two were misled by their assumptions - K2 for example, selected a feature poor background for Task 3, expecting that a plain background would make it easier for the app to identify the character.

6.2.7.2 Early stage keypoint marker feedback is not easy to understand

Participants of the Keypoints condition broadly failed to understand the meaning of keypoint markers and how it related to low-level features of interest to the algorithm (30% demonstrated a correct understanding in Task 3 and 50% Task 4). Participants K1, K2 and K3 incorrectly thought that the keypoint markers were highlighting regions where the algorithm had identified a moving object, something the user intended to animate. These participants theorised that if the algorithm succeeds in finding the objects which are meant to move, then the algorithm will be able to successfully transform the captured images to create animations e.g. K2 said "these dots might help show that the focus of the photo is the [character] [. . .] if I have these dots around the [character] then the image will be clearer". K2 and K3 both selected the worst background option for Task 3. They justified their choice by saying that among the three options the plainest background would work best because it would make the identification of the character easier for the algorithm e.g. K3, when asked why they chose a plain background in Task 3, said it was "because [the app] could be confused about the subject of the picture". Both K2 and K3 expressed confusion when keypoint markers appeared in locations which did not fit their understanding of how the system works i.e. on the background instead of the character. K2 remarking: "[keypoint markers] try to capture the [character] in the photo, a balloon, [. . .], but it's not on the balloon" and K3, "[if keypoint markers] mean the [character] is moving, [. . .] I don't understand why [keypoint] markers are showing up on the cloud, not the [character]". Despite witnessing evidence to the contrary both participants failed to correct their misunderstanding, a behaviour pattern previously reported in work on intelligent system (Tullio et al., 2007).

6.2.7.3 When keypoint marker feedback was helpful

The quantity of the keypoint markers was the most commonly reported explanation of how participants took into account Keypoint feedback. For example, K1 explained that if "[. . .] in background, [I] see a lot of dots. I can tell that background is definite. When I did the [animation of the] plane [for which the app failed], there were only 1 or 2 dots". K6 stated that "if there is nothing [in the background], it's not going to work. [If] something is there it's going to work". However, only four participants demonstrated a better understanding which was consistent with the workings of the stabilization process. These participants noticed how and where the keypoint markers appeared and were able to develop more specific theories of how the algorithm identifies keypoint markers within an image. For example, K10 correctly speculated that the algorithm "pick[s] up the shape" and "areas of heavy contrast".

In the Matching-Keypoint condition, six of the ten participants reported the feedback to be helpful. Of these participants, three described the keypoint markers as indicators, reporting what the algorithm was doing: "I can see what the dots are surrounding. [...] I know what it's doing" (M10), "when I saw [keypoints markers], it was more reassuring [...] saying you're doing it right" (M7), and "the app is trying to match between images [...] things which the app sees in this image which it also saw in the previous image" (M1). The other three participants explained that they saw the keypoint markers as guides, that the keypoint markers were designed to help them test if the background image would work or not: "the dots showed if the picture would work out" (M6), "I can tell what's the problem of the image" (M8) and "[the keypoints] might help you pick a background" (M5).

Participants in the Keypoints condition tended to overestimate the meaning of the Keypoint feedback and relate the meaning to higher level concepts, such as the separation of background and foreground objects. In this regard Matching-Keypoints appeared to be more intuitive as its meaning is more inline with user expectation. M1 for example, reported that when the app didn't work in Task 2 he did not know why. During Task 3, he speculated that the colour might have an effect (lighter or darker colour), but found through experimentation that this was not the case. He then correctly theorized that the app needed distinct features. He explained, "The dots meant like it's picking distinct points throughout the image. [...] I think [the app is] re-mapping the points that [it had] taken in an image before. I think that's what it's trying to do".

6.2.7.4 Split Screen feedback was helpful, but not in the way we expected

Seven participants in the Split-Screen condition also reported the feedback to be helpful. Four participants suggested that it acted as a cue, indicating when best to capture a frame e.g. "The preview helped me decide when to take a picture" (S7) or "I [wait] for the preview to stabilize before taking the picture" (S3). An artifact of the stabilization processes implementation is a "flickering effect" which occurs when the system is rapidly toggling between a successful transform and a failure. This strictly speaking is a usability "bug" which participants reappropriated, using it as a means of gauging the likelihood of a successful transform e.g. "If it was flickering I wouldn't take the picture" (S7), and "I waited for a clear picture [. . .] then hit capture" (S4).

Another unexpected way of using Split-Screen feedback was described by two participants (S7 and S2). They used the feedback to position the camera in the same place as the previous image, S7 commenting "the preview tells me what angle to take the picture from". Both participants would keep moving the camera until the left and right images matched in the preview i.e. the alignment transformation was minimal. This approach does in fact help make better quality animations, however it is not how the app was intended to be used and this process of positioning was very time consuming for the participants.

6.2.7.5 When feedback was not helpful

Five participants in the Split-Screen condition and three in the Matching-Keypoints condition reported the feedback to be distracting or unhelpful. For example, “I found the split screen very distracting and would rather not see it” (S4), “I found the dots distracting because it ruined the focus at times” (M4), “They were a bit annoying, they get in the way” (M1) and “they could be obstructive” (M6). Interestingly, S6 described the feedback as unhelpful because they prefered to frame the photo from memory, using the viewfinder to align the camera with features they had identified in the background. To this end the preview was unhelpful because the split screen design reduced the size of the viewfinder. These comments illustrate the risk that feedback visualisations can be distracting.

6.2.7.6 Background selection motivation

Although all participants selected a correct background in Task 4, not all provided a correct explanation. Participants responses when asked why they chose the background image they selected in Task 3 and Task 4 were coded into one of two categories: aesthetic - they were motivated by how the image looked, and detail - where they stated in some way that the level of detail was important (including incorrect understandings). Aesthetics was the primary motivation for 27 selections out of 80 (10 No-Feedback, 9 Keypoints, 5 Matching-Keypoints and 3 Split-Screen), with detail accounting for the remaining 53 selections (10 No Feedback, 11 Keypoints, 15 Matching-Keypoints and 17 Split-Screen). It should be noted that it is by chance that some of our participants considered the correct background to be more aesthetically pleasing.

6.2.8 Discussion

In the introduction we set out a series of questions. In this section we discuss the outcomes of our study using these questions as a scaffold.

6.2.8.1 Does the processing stage from which feedback is derived impact user understanding?

Our results indicate that feedback derived from the later stages of the processing pipeline (Matching-Keypoints and Split-Screen) are more effective at informing users’ understanding. The chi-square test of “user understanding” reveals a significant difference between conditions, with the standard residuals indicating the Keypoints and Split-Screen are responsible. More participants of the Split-Screen condition demonstrated a correct understanding of how the system works than participants of any other condition (Figure 6.7), with Matching-Keypoints second. In contrast, participants in the Keypoints condition performed worse than participants who received no feedback at all.

Despite users understanding varying between conditions, most participants across all conditions were successful in selecting a correct background (see Figure 6.1). As mentioned above, participants sometimes selected the correct background for aesthetic reasons, rather than to make the algorithm work (as requested by the study instructions). As a consequence, instead of using selection as a measure of understanding, we rely only on the participants' explanations of *why* they selected a specific background.

6.2.8.2 Is keypoint marker feedback intelligible to users?

More participants in the Matching-Keypoints condition were able to correctly describe the input requirements of the system in comparison with those who received no additional information in the form of feedback (No-Feedback). Interview responses indicate that users have a tendency to interpret feedback as an outcome rather than a progress notification of an intermediary stage. In this regard Matching-Keypoints appeared to be more intuitive, as their meaning is more inline with user expectation. We tentatively propose that keypoint markers can be used to inform user understanding, so long as the meaning being conveyed is inline with user expectations.

6.2.8.3 Can keypoint markers mislead if misunderstood?

Given that the Keypoints and Matching-Keypoints conditions utilise exactly the same feedback visualisation (keypoint markers), the result showing that Keypoints condition participants were least able to understand the needs of the algorithm (Figure 6.7) suggests that they may have been detrimental to user understanding. While the keypoint markers are a good indicator of the future stabilization processes success, participants commonly understood them to represent the final output, that they represented regions where the stabilization process had identified matches. It is feasible that this misconception could result in users using the markers in ways which inhibit their interactions. Indeed, Keypoints condition participants' interview responses indicate a disconnect between their interpretation of feedback and the actual information conveyed e.g. K3, "[if keypoints] mean the [character] is moving, [...] I don't understand why keypoints are showing up on the cloud, not the [character]".

6.2.8.4 Can keypoint markers improve usability and aid users' interaction?

The inherently visual nature of computer vision processes, both in their input and also the intermediate stages, makes visual feedback the logical medium through which to deliver feedback (Kato et al., 2012). However, participants in our studies, at times reported the feedback to be distracting or obtrusive (e.g. M1 "They were a bit annoying, they get in the way"). This highlights a design tension between attracting attention and causing distraction, and between being informative and not overwhelming. These tensions are well understood in graphic design, particularly around the design of interactive visualizations. However, the situation here is more complex. Some aspects of algorithm design are conceptually simple and naturally map to visual representations. Keypoints for example, are a concept that lend themselves to being represented pictorially e.g. by marking their physical location with geometric points. It could at first be tempting to see this as an example of "form follows function" (Sullivan, 1896), however when dealing with the design of feedback for systems which employ pattern matching algorithms, we argue that the "form follows function" principle requires careful interpretation. What is "function" in this case? At first, it may seem to be the "technical" function of the algorithm, but this is not the case. We need to remind ourselves that the "function" is instead the function to help users understand what the system does. One implication then, is that to design feedback, it may be beneficial to distance oneself from the question of how algorithmic steps and internal states map to form, and instead think about the end result of the system and how it will be used. Moreover, in some cases, it may be challenging, or even impossible, to map the function of the algorithm to form.

6.2.9 Summary

In this study, we examine the role of visual feedback in smart camera apps. Leveraging a novel experimental design centered on the creation of stop-motion animation, 40 participants were exposed to four different levels of feedback. Through a combination of quantitative and qualitative methods, our findings indicate a disconnect between user expectations and the information actually represented by the feedback. Participants exposed to keypoint marker feedback derived from early stages of processing showed a tendency to misunderstand it and overall they performed worse than participants who received no feedback at all. Conversely, participants who received keypoint marker feedback derived from later stages of processing demonstrated an improved understanding of the system operation. We conclude that the stage of processing from which feedback is derived plays an important role in users' ability to develop coherent and correct understandings of a system's operation.

6.3 Study 2

In Study 1 (Section 6.2), to mirror the experiences of users of existing commercial applications, participants were exposed to feedback without an introduction, and without an explanation of what information the feedback represented. In this second study, we investigate whether priming users' understanding with written instructions can align their expectations with the intended meaning being conveyed by feedback. To this end, we replicate the study design and procedures of Study 1, testing three new conditions where participants are provided with information relating to the feedback and underlying processed.

6.3.1 Study Design

The study design and procedure replicate that of the Study 1 in every way, bar one alteration: the instructions presented to participants at the beginning of the study were extended with the following:

- A high level overview of how the processing works: "The application tries to stabilise the animation by reshaping all the images so they look as if they have been taken from the same position. It does this by looking for things in each picture which are not supposed to have moved, for example the background."
- An introduction of the feedback (when appropriate) e.g. "In the app you are testing today the camera preview will show some orange dots called keypoints".
- A description of what the feedback represents (when appropriate) e.g. "Keypoints represent pixels which the app finds to be distinctive - please note that what is distinctive to the app may be different from what is distinctive to the human eye".
- An explanation of how best to use the feedback in order to meet the needs of the computer vision processing: "For the app to work well the distinctive features should appear on the background, and be spread evenly, rather than concentrated on a small area. Distinctive features on a moving object (e.g. the character) don't help".

6.3.2 Conditions

The three conditions examined in this second study, are duplicated from the first study and only the instructions modified. Note: the Matching-Keypoints condition was not tested as it already proved effective without instructions.

No-Feedback-With-Instructions As per the first study, this condition was included as a baseline. The input images to the pipeline were presented back to participants without any additional feedback.

Keypoints-With-Instructions The camera's viewfinder was augmented with keypoint markers which indicate the locations at which keypoints had been detected in stage 2.

Split-Screen-With-Instructions This condition represents the final stage of processing. The viewfinder was divided into two equal halves. On the left: the input image updated in real-time (as per No-Feedback condition). On the right: the image outputted by the processing pipeline.

6.3.3 Participants

A further 30 participants (16F, 14M) were recruited from the university participants pool. As before, each participant received a £10 payment for their participation and were required to meet the same criteria i.e older than 18, with no technical background and normal / corrected to normal vision. Participant age ranges were as follows: 3 aged between 50 and 59 years, 3 between 40-49, 4 between 20-29 and 20 between 18-19. Ten participants were randomly assigned to each condition. For conciseness, we refer to participants by condition and subject number. Prefixes "NI", "KI" and "SI" refer to the No-Feedback-With-Instructions, Keypoint-With-Instructions and Split-Screen-With-Instructions respectively.

6.3.4 Data Collection & Analysis

In line with Study 1, audio recordings and detailed notes of observations were collected by experimenters during the study and independently coded by three researchers in two

Table 6.3: No. Participants who correctly selected a "correct background" (i.e. one suited to the needs of the stabilisation process).

	Task 3	Task 4
No-Feedback (N)	10	10
No-Feedback-With-Instructions (NI)	10	10
Keypoints (K)	7	10
Keypoints-With-Instructions (KI)	8	10
Matching-Keypoints (M)	10	10
Split-Screen (S)	9	10
Split-Screen-With-Instructions (SI)	10	10

rounds of analysis. The first quantitative analysis focussed specifically on participants' responses to questions pertaining background selections and were coded focussed on their understanding of how the system works. The second qualitatively analysis was not so constrained and all recorded material was coded i.e. codes were drawn from research questions, the prior study and any which emerged from the interviews.

6.3.5 Quantitative Findings

In this section, we report the results of the quantitative analysis. For comparison, the results are reported alongside those of the previous study (i.e. Study 1).

Table 6.3 summarizes the background selections made by participants for Tasks 3 and 4. Figure 6.8 shows whether their selection was based on a correct understanding of the stabilization processes. A chi-square test of the summed answers which demonstrated a correct understanding in Task 3 and Task 4 (as per Study 1) revealed a statistically significant difference across all conditions, i.e. those of Study 1 and Study 2 (chi-square=18.27, p=.006, df=6, Cramer's V=0.361).

Figure 6.9 presents how many participants were able to correctly describe the reason for Task 2's failure: (i) immediately after the event, and (ii) when asked in the interview at the end of the study. A chi-square test of the number of participants able to demonstrate a correct understanding of why Task 2 was problematic, revealed a statistically significant difference (chi-square=18.27, p=.006, df=6, Cramer's V=0.551) across all instruction conditions and their counterparts (i.e. all conditions excluding Matching-Keypoints), with the Cramer's V score greater than 0.5 indicating a large effect. The most sizeable differences can be seen between Keypoints (where less than half of participants demonstrated a correct understanding) and Keypoints-With-Instructions (where almost all did). To a lesser extent, but still notable is the difference between No-Feedback (50%) and No-Feedback-With-Instructions (70%).

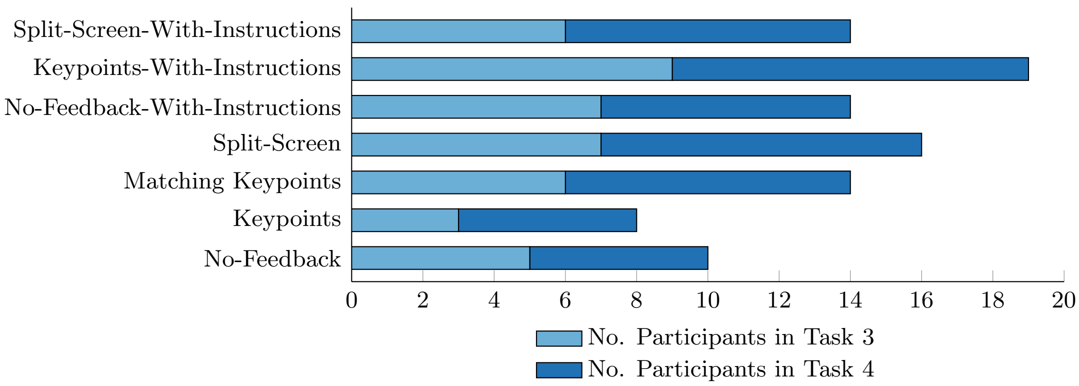


FIGURE 6.8: Participant Understanding

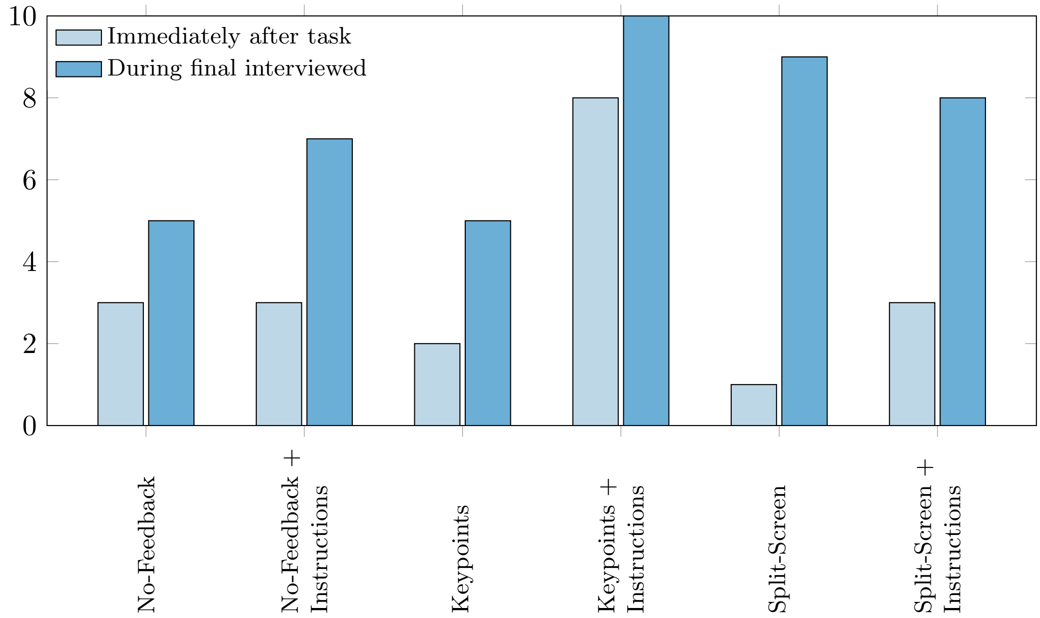


FIGURE 6.9: Participants who correctly explained why Task 2 failed - After task and at the end of study

6.3.6 Qualitative Findings

Participants, once again, overwhelmingly reported the task to be interesting and entertaining. Further supporting our assertions that the experimental task was sufficiently engaging and participants were invested in creating animations successfully.

6.3.6.1 Participants existing knowledge

No participants reported using existing application which utilised keypoint marker feedback. As in Study 1, participants mentioned Facebook and Instagram, neither of which provide visual feedback for their computer vision interactions.

6.3.6.2 Instructions and No-Feedback

In the No-Feedback-With-Instructions condition 7 out of the 10 participants demonstrated a correct understanding e.g. NI10 said "the more details the better" when describing how they made background selections (Figure 6.8). They went on to correctly explain that "Task 2 could be fixed [by] adding some things to the background". Similarly N9 said: "I could fix task 2 by adding some features to the background. Maybe some objects in the sky" as did NI5 who suggested the "addition of birds".

6.3.6.3 Instructions and Keypoints

In stark contrast to the Keypoints condition of Study 1, Keypoints-With-Instructions participants demonstrated a correct understanding most often in Study 2 (Figure 6.8). For example, KI2 said the "plane didn't work because [there were] no background dots. More colours and lines would help make more dots." and KI5: "The dots are needed so you know if the background is going to work or not". While the instructions were reported to be clear by all participants, KI4 did remark "It took me a minute to understand how the dots worked, where they would appear". All bar one of the participants who demonstrated a correct understanding, reported the keypoint markers to be helpful when choosing a background. For example, KI3 who said that "the dots helped me choose a good background. [I] had sense if it would work or not. If they [the keypoints] weren't there then how could you know why it fails". KI9 however said that "the keypoints are helpful when picking a background, but after that [I would] turn them off", they felt that once the background selection was complete that they could rationalise how to succeed. This is interesting because in the controlled lighting conditions of our user study this is to some degree true. However, if lighting and other environmental factors changed it is possible that the features on which the pattern matching relies would be less detectable.

6.3.6.4 Instructions and Split-Screen

Split-Screen-With-Instructions was the only condition in Study 2 which performed worse than its Study 1 counterpart (i.e. the Split-Screen condition). However, the majority of participants were able to develop a correct understanding of the systems input requirements and participants performance was on a par with Matching-Keypoints (Study 1) and No-Feedback-With-Instructions (Study 2). SI8 for example was able to explain that " of the lack of detail and contrast in the background" caused task 2's failure and SI4: "I chose the rocks in task 3 because there was more detail, but it wasn't perfect. I thought the bubbles were too light. I chose the clouds in task 4 because there was good shape and distinctive features". As per Split-Screen in Study 1, three participants of Split-Screen-With-Instructions reported using split screen as a cue, indicating when best to capture a frame . SI4 described that: "when it was stable I used it. It told me when to take a photo". SI5 described how "when it wasn't flickering [they] could take a picture" and SI2 reported: "I used it to know if I could take a picture. If the right was all crazy then I would adjust the app angle. If it was still I used the left to get the best framing".

While the instructions appear to have made an impact on user understanding, some participants in the Split-Screen-With-Instructions condition still found the feedback at times to be unhelpful. SI7 reported the split-screen feedback to be disorientating: "I found it confusing, it made me muddled up", prefering to take a photo and see if it worked, while SI8 said they "the split screen was a bit off putting" and SI1: "it lets you know if it fails. It is good in failure cases, but not so helpful when success". SI9 failed to understand the feedback at all, saying: "I feel like it's user error, [...], I think it's more about me, I was a bit unsteady".

6.3.6.5 Background selection motivation

Once again the number of participants who selected a good background varies little between conditions (Table 6.3). Further demonstrating that this is not a good measure of understanding. However, the qualitative data pertaining to the motives behind selection reveals something much more interesting. In the same manner as Study 1, participants responses when asked why they chose the background image they selected in Task 3 and Task 4 were coded into one of two categories: aesthetic - they were motivated by how the image looked, and detail - where they stated in some way that the level of detail was important (including incorrect understandings). In a change from Study 1, detail accounted for 47 selections (13 No-Feedback-With-Instructions, 20 Keypoints-With-Instructions and 14 Split-Screen-With-Instructions). With aesthetics in second, 13 selections out of 60 (7 No-Feedback-With-Instructions, 0 Keypoints-With-Instructions and 6 Split-Screen-With-Instructions).

6.3.7 Discussion

The composition of the guidance materials was a key consideration during the design of this study iteration. It was imperative that the instructions provided sufficient information to allow participants to better understand the systems input requirements, whilst ensuring that user understanding could be differentiated from simple repetition of the provided materials. In this regard we assert that our quantitative findings indicate that we were successful in this goal, whilst there was broadly an improvement in user understanding, this was not distributed evenly across conditions and a significant difference is present.

6.3.7.1 Do instructions improve user understanding?

The quantitative findings show that instructions had a significant impact on participants ability to understand system operation. The motivation for background selection demonstrates this. In Study 1 the primary motive for background selection was its aesthetics, while in this Study the level of detail became the primary motivation.

Comparing the Kepoints condition of Study 1 with the Kepoints-With-Instructions condition further demonstrates the impact of instructions. In the first study, Kepoint condition participants were the least able to demonstrate a correct understanding of the system needs, performing worse that participants shown no keypoints. In this second study, the same keypoint marker visualisation coupled with instructions was the most effective of all feedback conditions (Figure 6.8). Our hypothesis that a misalignment between user expectation and meaning being conveyed by feedback was the route course appear to have been validated by this result, with the instructions bringing participants expectations in to line the information actually being conveyed by the keypoint markers.

It is also evident that instructions aided participants who received no feedback at all. While the improvement is not as substantial as other conditions, No-Feedback-With-Instruction participant understanding is equivalent to Matching-Keypoints condition participants of the first study. This highlights an important **design implication**: if later stage feedback is not feasible (e.g. computational power prohibits it), providing earlier stage feedback coupled paired with instructions is a viable alternative. However, written instructions are often avoided by designers who want intuitive interfaces, creating a tension between the efficacy of feedback and the design ideals.

6.3.7.2 Instruction and Experimentation

Participants of the Keypoints-With-Instruction condition were significantly more able to correctly describe the reason for Task 2's failure immediately after the task than any other condition. This difference however, appears to diminish by the end of the study (i.e. once participants have completed two further tasks). While we acknowledge the limitations of this measure, this finding coupled with those of Chapter 5 suggests that while instructions expedite the formation of understanding, it can also be achieved through experimentation. The **design implication** then, is that short infrequent interactions (e.g. Amazon and Samsung's apps) will benefit from instructions, while applications which expose users to sustained periods of feedback will allow users to learn through experimentation.

6.3.7.3 When feedback was helpful

Participants of the feedback conditions broadly reported considering the feedback when taking a picture (9 participants in the Keypoints-With-Instruction condition and 8 in Split-Screen-With-Instruction). The quantitative data demonstrates that participants of the Keypoints-With-Instruction conditions developed a correct understanding more often and sooner than participants of other conditions. However, this was not the case for Split-Screen-With-Instruction condition participants, who performed marginally worse than their counterparts in study 1.

6.3.7.4 When feedback is not helpful

Of the conditions examined in this follow-up study, participants of all conditions, except Split-Screen-With-Instructions, were better able to correctly describe the input requirements of the system in comparison to their study 1 counterparts. Qualitative analysis did not reveal any specific misconceptions nor reasons for the variation in the split-screen conditions and given the small quantitative differences, it is not possible to draw any firm conclusions. However, we note a recurrence of a tension between attracting attention and causing distraction, first identified in the study 1. Split-Screen-With-Instructions condition participants reported the feedback to be "off putting" and "annoying", while only one participant of the Keypoints-With-Instructions condition said that they would prefer not to see feedback.

6.3.8 Summary

In this study, we extend the work reported in Study 1 (Section 6.2), exploring the role of written instructions. Employing the same novel experimental design as Study 1, 30 new participants were exposed to three additional conditions. The No-Feedback, Keypoints and Split-Screen conditions of Study 1 were copied and supplemented with written instructions which detailed the system and when necessary the meaning of feedback. Further to reporting findings which support those identified in Study 1, we also find that instructions are an effective way of bringing user expectations inline with the meaning being conveyed by feedback. We also find that instructions can expedite the formation of coherent user understanding.

6.4 Limitations

Designing a controllable study requires a degree of compromise between ecological validity and the gathering of comparable data. To further validate our findings, examination of feedback in less controlled conditions would be advantageous. Moreover, the application we developed to conduct the user studies reported in this chapter, centers around a task specifically designed to accelerated interactions with feedback. Commercial applications such as Amazon and Samsung's Bixby have shorter periods of interactions and while our findings hold, examination of short users interaction (e.g. app telemetry) may yield further insight into the efficacy of our findings.

Furthermore, in study 2, a participant commented that the "the keypoints are helpful when picking a background, but after that [I would] turn them off", they felt that once the background selection was complete that they could succeed without it. While under the controlled lighting conditions of our user study this is to some extent true, in the real world this unlikely that the environment will remain unchanged. This highlights an opportunity for future work, while the consistency of environmental conditions was necessary for comparison in our studies, addressing how changes in the environment affect users ability to develop coherent understanding os systems is worthy of investigation.

6.5 Implications

In the first lab study we saw a disconnect between user expectations and the information actually represented by the feedback. This led to misconceptions and ineffective interactions. Participants exposed to keypoint marker feedback derived from early stages of processing overall performed worse than participants who received no feedback at all. Conversely, participants who received keypoint marker feedback derived from later stages of processing demonstrated an improved understanding of the system operation. The implication then is that the stage of processing from which feedback is derived plays an important role in users' ability to develop coherent and correct understandings of a system's operation.

In the second iteration of our user study (Study 2), we found that issuing explicit guidance relating to feedback and what it represents has a significant impact on the speed at which users understand the cause of failures. Moreover, participants of the Keypoints-With-Instructions condition were better able to demonstrate a correct understanding of the systems needs than any condition of either Lab Study. This is in stark contrast to its counterpart condition (Keypoints) in the first study, in which participants performed worse than even those participants who receive no feedback at all. We conclude that, feedback derived from the later stages of processing are better suited to instruction-free user experience, however, guidance can be used to bring user expectations inline with the meaning of early stage feedback and improve user interactions overall.

Finally, we hope that our novel experimental design and study method can be used by HCI researchers in future work exploring the design space of feedback and cues.

6.6 Summary

In this chapter we report on two iterations of a lab study examining the role of visual feedback for smart camera apps. Through a comparative between-groups study, 70 participants were exposed to four different levels of feedback without instructions and three levels with instructions. The study leveraged a novel experimental design enabled by a bespoke smart camera app and set of experimental tasks around the creation of stop-motion animation. Through a combination of quantitative and qualitative methods, our findings indicate that the experimental design and tasks were successful in engaging participants and generating controlled failures. Our findings suggest that non-technical users tend to relate feedback to the later stages of processing and feedback not inline with this expectation (i.e. feedback from early and intermediate stages) can mislead and confuse users. However user expectations can be managed with instructions and earlier stage feedback can be made effective with guidance.

Chapter 7

General Discussion & Conclusions

Norman (2013) describes feedback as one of the key tools which designers can employ to bridge the gulf of evaluation. The work presented in this thesis examines how feedback derived from system processes can be used to encourage the development of greater user understanding and improve user interaction. To this end, a series of 6 user studies were designed, developed and conducted (excluding iterations), centered around sensor based smart systems. These studies were developed to evaluate the role of feedback and its effect on "user understanding" - a term used in this thesis to express the knowledge held by a user about how a system works, their past experiences of it and how they reason about future interactions. While the definition of user understanding draws heavily on prior work relating to Mental Models (e.g. Yarosh and Zave (2017); Zhu et al. (2017); Huang and Cakmak (2015)), it was selected to address valid criticism of the overuse and misrepresentation of Mental Models (Carroll, 2003; Rogers, 2012).

7.1 Summary & Key Findings

This thesis begins (Chapter 1) by outlining the research problem to which this work contributes, i.e. that while there exists a wealth of literature examining user interactions with traditional systems, the body of work relating specifically to smart systems is less well developed. Further to this, an opportunity for investigation was identified, namely; how feedback could be used to inform user understanding such that they could interact more effectively. In particular, the following research question was defined:

> How does the design of smart system feedback impact users' with no specialist knowledge of the underpinning technologies ability to develop useful understandings of system operations such that they can interact more effectively? And what are the implications of bad design?

Chapter 2 provided a backdrop to this thesis, reviewing literature from a range of disciplines. In particular, literature exploring the following questions was reviewed; how smart systems have been defined in prior work, how user understanding can be modelled and how existing research has tried to inform user understanding of both traditional systems and more recently smart technologies. In addition, literature specific to the sensor based prototype systems developed for experimental purposes were considered.

Chapter 3 reports an investigation of how user understanding of information derived from sensor based systems can be fostered through natural language feedback, in a controlled lab study. The findings suggest that the level of detail contained within textual messages will impact users' impetus to take action. Information-poor messages (i.e. short messages with limited detail) are sufficient to invoke action, however they can result in unnecessary action as they do not inform understanding sufficiently for users to reason about their urgency. In contrast, information-rich messages can reduce unnecessary action, but their verbosity is undesirable.

Building on this user study, Chapter 4 examines more dynamic interactions with sensor data. Through a series of user studies (including one conducted in the field) the work reported in this chapter explored how real-time sensor and connectivity feedback can be used to inform users with no experience of the technological aspects of wireless sensor deployment, so that they can effectively identify *suitable installation locations*. The findings suggest that offering a suggested course of action significantly improves users' ability to find *suitable installation locations* (in comparison to no feedback and real time visualisations of sensor data). Expanding on the notion of suggested course of action, further studies explore how it can be represented visually. The findings of further studies showed that a spatial representation which encodes a suggested direction of travel, significantly expediates users' search time. The studies reported in this chapter were intentionally designed to constrain participants by location alone. In the next chapter, the concept of a *suitable installation location* is refined to take into consideration the requirements of the processes which consume the data yielded by sensors.

Chapter 5 reports an investigation of how visual feedback common to many smartphone applications (keypoint markers) can be used to inform users with no specialist knowledge of pattern matching technologies, such that they can reason more effectively about expected outcomes. In a between-groups study, 18 participants were tasked with examining a series of image pairs and estimating whether a computer vision process would be able to match artifacts within the image pairs. The findings suggested that keypoint marker feedback had no significant effect on these users' ability to predict outcomes. Interviews and quantitative data however revealed a potential limitation in the study design, that the use of still images inhibited participants' ability to experiment and theorise about actions which might affect outcomes.

To address these concerns, Chapter 6 reports two follow up studies, with the aim of developing a better understanding of how non-explicit feedback can help users with no specialist knowledge of pattern matching technologies better understand which actions are most likely to result in a satisfactory outcome. These studies utilised a novel experimental task, in which participants created stop motion animations with a handheld camera whilst subject to one of eight feedback conditions. The findings suggest that users tend to relate feedback to the later stages of processing and that feedback not in line with this expectation (i.e. feedback from early and intermediate stages) can mislead and confuse users. However, user expectations can be managed with instructions, while earlier stage feedback can be made effective with guidance.

7.2 Design Implications

Throughout this thesis a number of **design implications** have been highlighted. This section summarizes them.

Chapter 3 Design Implications

1. Information-poor textual messages are sufficient to invoke user action, but contain too little information for unnecessary actions to be avoided.
2. Long and detailed messages improve users' ability to avoid unnecessary action, by allowing them to better understand when to act. However, the verbosity of such messages is undesirable and could potentially impede efficient interactions.

Chapter 4 Design Implications

1. Real-time connectivity and sensor fidelity feedback is helpful to users and should be made available whenever possible.
2. Textual prompts suggesting a course of action significantly improve users ability to understand how to act in response to feedback.
3. Novel mechanisms of guiding users warrant further exploration. In particular, given the ubiquity of AR technologies in modern smartphones, the opportunity to foster user understanding in this way is considerable.
4. Users tend to fixate on small fluctuations in feedback, which can be detrimental to their efficiency and mechanisms to "smooth out" such variations should be found.
5. When developing hardware it is important to ensure they are robust both in the making and also their design, as users exhibited considerable concern over the protection of devices.

6. While using smartphones and tablets is a desirable alternative to the incorporation of screens in smart devices, they can introduce considerable confusion for users, who can misunderstand the relationship between what is being monitored and what is doing the monitoring.

Chapters 5 and 6 Design Implications

1. Interactivity and the freedom to experiment plays an important role in users ability to develop an understanding of the meaning of feedback.

2. The stage of processing from which feedback is derived is an important consideration when designing interactions as users expect feedback to represent the outcomes of system processes.

3. While not necessarily desirable, instructions are an effective means of bringing user expectations into line with the meaning being conveyed by feedback.

4. Without instruction later stage feedback is more likely to be in line with user expectations and thus leads to more effective interactions.

5. Misunderstandings of early processing stage feedback can be detrimental to user understanding and potentially lead to actions which are contrary to the needs of dependent processes.

7.3 Discussion

Having recapped the work reported in this thesis, in this section the common threads are identified and discussed in relation to prior work. To avoid repetition, below are three key terms (adapted from Norman (2013)) which are frequently used in the subsequent sections.

1. *Conceptual model* - The system designers model of how the system should work i.e. how the designers envisage the system to work and how it should be perceived.

2. *System image* - How the system is portrayed through the user interface and supporting materials (e.g. instruction manuals).

3. *Mental model* - How a user understands the system to work based on (i) how they perceive it to work i.e. from the system image and (ii) what they discover through interactions with the system.

7.3.1 When Feedback *Is* Helpful

The prevailing wisdom in HCI is that greater understanding leads to better interaction and that it is the system designer's job to provide users with sufficient information so that a useful understanding of how a system operates can form (Norman, 2010), i.e. users develop a useful *mental model*. While bridging the gulf of execution (i.e. educating users on what interactions are possible and how they can be achieved) can be addressed through the careful design of signifiers, mappings and constraints (Norman, 2013), these approaches lend themselves to the specifics of a smart systems design. This thesis instead focuses on the more generalisable challenge of bridging the gulf of evaluation, i.e. helping users reflect on the consequences of their actions.

7.3.1.1 Explicit Guidance

Guidance materials are common to almost all man-made systems (Norman, 2010) and while the nature of these materials varies from quick tooltips to multi-page user manuals, they are broadly accepted to benefit to user understanding (or at least when they are read (Novick and Ward, 2006)). This thesis supports this proposition, demonstrating how instructional materials can aid user interactions with smart systems: (i) in Chapter 3, investigations of how textual messages can be used to report the outcomes of fault sensing, demonstrate that information-rich messages are significantly more likely to avoid unnecessary action, (ii) in Chapter 4, studies of wireless sensor placement, reveal that explicit messages suggesting a course of action, significantly expedited participants search for connectivity, and (iii), in Chapters 5 and 6 the findings indicate how written instructions helped bring users' expectations of feedback into line with the meaning truly being conveyed.

Early work in this space investigated how users of programmatic calculators develop *mental models*. Bayman and Mayer (1984) demonstrates the effectiveness of written instructions and strongly advocates their use in end user systems e.g. "written instructions help users of calculators develop deeper mental models so that they [can] interact more effectively". While users familiarity with such technologies has undoubtedly increased since this work was first published, with computing technologies now pervading almost every aspect of everyday life, the potential for complicated and confusing interactions has also risen thanks to the increasing complexity of interconnected smart systems. The findings of this thesis support Bayman and Mayer (1984)'s recommendation that designers should provide users with explicit explanations of the system's *conceptual model* so that they can better relate their "hands-on experience" with the *system image*.

More recently in the domain of cognitive science, (Harold et al., 2015) demonstrates how simple "linguistic warnings" are an effective way to improve user interaction with time-series visualizations. Similarly, Kong et al. (2018) shows the titles of time-series visualizations have the potential to influence interpretation. The studies reported in Chapters 3 and 4 support this work, showing how natural language textual notifications can improve users' understanding of when to act and in what way.

This thesis and the prior work in this space also highlight the design tension between the ideals of ubiquitous computing (i.e. interactions are imperceptible) and the effectiveness of explicit instruction. While Novick and Ward (2006) highlight the reluctance of users to read lengthy written instructions prior to interaction, the capacity for instructions to positively impact user understanding is apparent.

7.3.1.2 Exposing Systems

Common to the user studies reported in this thesis is the exposure of underlying system processes to users. As discussed in the introduction, Norman (2010) champions visibility and describes how with careful design the complications borne out of complexity can be tamed. This thesis complements this notion, demonstrating the fine line between designs which promote understanding through the exposure of underlying system processes, and designs which add confusion making interactions more complicated for users. In the subsequent paragraphs the positive effects are discussed, and in the following section (7.3.2) the potential negative implications are considered.

The findings of Chapter 3 show how the degree of information encapsulated in feedback messages affects users' capacity to reason about when to take action. In Chapter 4 observations of users exposed to the hidden landscape of radio propagation when searching for locations to install wireless smart sensors, revealed how exposure to this information significantly improved their ability to find suitable installation locations. Chapters 5 and 6 demonstrate how users' understanding of pattern matching processes can be fostered through feedback derived from intermediate processing stages. This work is in line with Lim et al. (2009)'s work, which shows how the information content of messages plays an important role in how users reason about system actions and Kulesza et al. (2015) who demonstrates how explanations of machine learning predictions can help users debug system operation. The findings of this thesis suggest that feedback can be used to effectively build useful *mental models* which are not only helpful for normal interactions, but can also aid users in overcoming unexpected outcomes.

7.3.2 When Feedback *Is Not* Helpful

While the findings of this thesis are in line with Norman (2013)'s assertions that feedback can foster more useful *mental models* and help users overcome the gulf of evaluation, they also support his emphasis on good design. The studies reported in this thesis highlight the pitfalls of poor feedback design. In this section the potentially negative impact of feedback are discussed.

7.3.2.1 Feedback Can Be Confusing

Across this thesis, analysis of the quantitative and qualitative data collected through user studies has revealed instances where confusion has arisen from interactions with feedback. This is most apparent in Chapter 6, where participants exposed to early-stage keypoint marker feedback (when creating animations with the Anim8 app) were less able to explain what constituted a suitable input than participants who were shown no feedback at all. Analysis of interview responses revealed a range of misinterpretations which highlighted a misalignment between user expectations and the meaning of the feedback, i.e. users expected feedback to represent processing outcomes, when in reality it was an indicator of the potential success or failure of the process. To a lesser extent, Chapter 3 shows how information-poor messages invoke unnecessary action, failing to provide sufficient information for users to reason about the urgency with which action should be taken. Chapter 4 reports how participants searching for *suitable installation locations*, fixated on short lived anomalies in radio signals (e.g. caused by electronic interference). These anomalies resulted in users abruptly interrupting their search and spending a considerable amount of time performing a detailed search of the small area, trying to find the signal they had once seen. These findings highlight the challenges of designing effective feedback and the tension that exists between providing sufficient information to overcome confusion and presenting too little information increases the likelihood of misinterpretation. Alan et al. (2016b) reports observations of how such misunderstandings can result in confusion and be detrimental to user interaction.

Kizilcec (2016)'s work with algorithmic peer-assessment interfaces reports similar findings. They show that when user expectations are violated (i.e. when they receive a grade less than they expected), explanations of system operation limited the erosion of their trust in the system. Interestingly, they also show that if the system is too transparent the erosion of trust increases, highlighting the careful balance which must be struck. Similarly, Lim et al. (2009) shows that explanations which express *why* a context-aware system acted as it did are more effective than the logical equivalent *why not* explanations, further highlighting the nuanced nature of design.

7.3.2.2 Feedback Is Not Always Enough

Participants of the studies reported in this work also demonstrated how some *mental models* are robust and difficult to change. In Chapter 6, three participants developed flawed *mental models* from the feedback conveyed by the Anim8 application, incorrectly believing the keypoint marker feedback to represent regions where the algorithm had identified a moving object (rather than static elements). Despite expressing confusion when witnessing evidence to the contrary, they maintained their flawed *mental models*. Similarly, in Chapter 5 participants who incorrectly postulated that keypoint marker feedback was indicating areas where the object detection algorithm was "having difficulties" remained convinced despite each instance being followed by an outcome which suggested otherwise. In Chapter 4 three participants incorrectly conflated the sensing device's connectivity with the sensor's fidelity, i.e. they reasoned that poor connection would reduce the quality of data captured. Despite their own experimentation demonstrating this was not the case, they remained resolute in their beliefs. Tullio et al. (2007) reports similar observations. During a six week long field study they note how some office workers had high-level beliefs which were so robust that even when contradictory aspects of the systems true implementation were revealed they failed to change their understanding. They go on to suggest that low level feedback may not be enough and that users might need higher level information to develop better *mental models*.

7.3.3 Designing Experiments

This thesis documents two novel experimental designs, making a methodological contribution in addition to the experiments they permitted. In Chapter 4 an experimental task was developed which exposed participants to the complex landscape of radio propagation (i.e. wireless connectivity), challenging participants to find a balance between connectivity and sensor fidelity in an increasingly difficult conditions. In Chapter 6 an experimental task was developed focusing on complexity within the system (i.e. internal processes of pattern matching) as opposed to complexity in the environment. A bespoke app was created, permitting interactions with computer vision processes where the outcomes (i.e. failure and success) can be manipulated in an inconspicuous way.

Through numerous pilot studies and design iterations, these experimental designs were refined to discreetly elicit user understanding while users experienced interactions at the boundaries of satisfactory system outcomes, whilst ensuring meaningful conclusions could be drawn i.e. they have ecological validity, repeatability and controllability. In addition to reporting the study designs, the software and hardware has been released as open source so that future work in this space may benefit.

7.4 Limitations & Future Work

One of the strengths of this thesis is the design of controlled lab studies to elicit and assess user understanding while conducting realistic tasks. However, as with all experiments, a compromise must be reached between ecological validity and structure which permits meaningful measurement. The work presented in this thesis focuses largely on meaningful measurement, as such controlled studies were most commonly employed (one field study was conducted). While this does not detract from the findings, additional field studies would complement this thesis and further validate the findings.

In the late 1980s, as HCI emerged as a separate discipline, Nievergelt and Weydert (1987) proposed that a user understanding was comprised of three components: knowledge of the past (how the system reached its current state), knowledge of the present (the current state of the system) and knowledge of the future (how an action will affect the state of the system). The majority of feedback tested in this thesis relates to knowledge of the present. However, there are a few exceptions. For example, investigations of the role of feedback in sensor placement (Chapter 4), test examples of knowledge of the past and the future. The Bar+Arrow condition (which used augmented reality) presented historic data, guiding the participants to where good connectivity had previously been detected, and the Bar+Msg condition presented a suggested course of action, which highlighted the potential of a future action. While these two examples are representative of exploiting past and future information to build knowledge, more work is needed to explore the considerable design space. For example, how can historic data supplement real time feedback (something Renaud and Cooper (2000) is confident has the potential to benefit user interactions), or how can predictive machine learning be used to suggest future actions.

This thesis focuses on feedback and how it can help users develop better understandings of system operation so that they can interact more effectively. While some consideration is paid to how this applies to overcoming unexpected outcomes, this is only explored from the perspective of the system being the cause of erroneous interaction. However, human error is also a potential factor and warrants further study. Jewell et al. (2015)'s investigations of users configuring wireless devices, draws on the work of (Norman, 2013) who defines two types of human error: Slips (when a user intends to carry out one action, but ultimately carries out another) and Mistakes (when an action is carried out deliberately because of a floored plan or lapse of memory) to catalogue a series of user errors and demonstrate how good interaction design can mitigate them. Understanding how feedback can help users identify their slips and mistakes is another interesting avenue for future work and has the potential to improve user understanding.

7.5 Conclusion

This thesis reports a series of user studies which examined how feedback yielded from smart sensor based systems can foster the formation of greater user understanding. In so doing, this thesis makes a methodological contribution in the form of experiments designed to examine user interactions at the boundaries of successful and failed interactions.

Building on prior work, this thesis demonstrates that feedback plays an important role in the formation of greater user understanding of smart systems (i.e. more useful mental models). In (Chapter 3) how natural language output can be utilised to develop user understanding is explored so that informed choices can be made about when to take action. Chapter 4 investigates how feedback can help users better understand system constraints when physically placing wireless smart sensors, aiding their search for suitable installation locations. Chapters 5 and 6 identify the challenges faced by users of pattern matching technologies common to many smart system applications and explore how feedback can be used to reveal the inner workings of such systems.

In addition, from the findings of this work a set of **design implications** are derived, which both highlight the pitfalls of bad design and emphasise the key considerations when designing feedback for smart systems.

Appendix A

Feedback Derived from Different Stages of Processing - Follow Up

A.1 Introduction

In Chapter 6 we report two user studies designed to address whether keypoint maker feedback: (i) is intelligible to lay users (ii) can improve usability and aid users' interaction around failures, and conversely (iii) mislead users if misunderstood. In so doing we identify a design tension between attracting attention and causing distraction, and between being informative and not overwhelming (Section 6.2). To investigate the design space of feedback and how the visualisations design can address the challenge of attracting attention without causing distraction, while being informative without being overwhelming, we developed a new visualisation based on keypoint marker feedback. This new visualisation (we refer to as "Reveal") was designed as a direct replacement for the keypoint marker feedback tested in the Keypoints and Matching-Keypoints condition. Here, we report our findings from a user study (n=10) designed to evaluate the effectiveness of this visualisation.

A.2 Reveal Visualisation

The design of Reveal visualisation draws on the metaphor that the Anim8 system only "sees" the regions of the images where keypoints are detected. The camera preview is fully covered with an opaque black layer. For each keypoint a circular portion of this layer is removed, revealing the region of the camera preview where the keypoint was detected. The radius of the circle is derived from the same keypoint marker data and reflects the scale spaces at which the keypoint was identified. Figure A.1 shows an example of how this visualisation appears in the Anim8 interface.

FIGURE A.1: Reveal Visualisation Example

A.3 Study Design

The study's design and procedure replicates the study reported in Section 6.2. The only alteration is the change of feedback conditions tested. In this study we test only the Reveal feedback so that it can be compared with the feedback conditions of the prior studies i.e. those reported in Chapter 6.

A.4 Participants

A further 10 participants (4F, 6M) were recruited from the university participants pool. As before, each participant received a £10 payment for their participation and were required to meet the same criteria. Participant age ranges were as follows: 5 between 20-29 and 5 between 18-19. We refer to participants using the condition prefix "R", and subject number.

A.5 Quantitative Findings

Table A.1 summarizes the background selections made by participants for Tasks 3 and 4. Figure A.2 shows whether their selection was based on a correct understanding of the stabilization processes.

A.6 Qualitative Findings

The majority (6) of participants were not able to interpret the Reveal feedback's meaning. Most participants did not understand what it meant. One participant (R31) noted

TABLE A.1: No. Participants who correctly selected a "correct background" (i.e. one suited to the needs of the stabilisation process).

	Task 3	Task 4
No-Feedback (N)	10	10
Keypoints (K)	7	10
Matching-Keypoints (M)	10	10
Split-Screen (S)	9	10
Reveal (R)	9	9

that it was something technical, that the app was trying to detect something, but she did not know what, and said, "On the preview, there were little bubbles. I wondered why they were there. [...] it's something to do with a program". Since she was not sure what it meant, she "didn't think much about it". In the most extreme case a participant (R26) thought the app was "faulty". The lack of understanding expressed by participants in the Reval condition explains why most participants in this condition did not report considering the feedback. Three participants stated that they considered the Reveal feedback when choosing a background image. Of these, the quantity or scale of the feedback was the most commonly reported explanation of how the feedback was taken into account. For example, R23 explained, "More noise in the background, so it is more easier. [...] It picks up more background".

One participant (R27) in the Reveal condition thought that the feedback indicated a boundary, an area beyond which an object should not be moved. He remarked, "I understood how it works. The main thing is, I should have the object inside the visible area". His assumption was that the black area concealed the region of the image the algorithm was considering to be the background and therefore moving the character into this space would lead to distortion of the transformed image. At the end of the experiment, he still was not able to explain why he had failed to succeed in animation 2. R23 thought that the key point markers were in some way similar to their previous experiences of the "Lasso" highlighting tool in Photoshop.

The majority (7) of participants in the Reveal condition found the feedback distracting and in some cases annoying. Unlike key point markers which largely went without question, three participants during the animation tasks asked an experimenter what

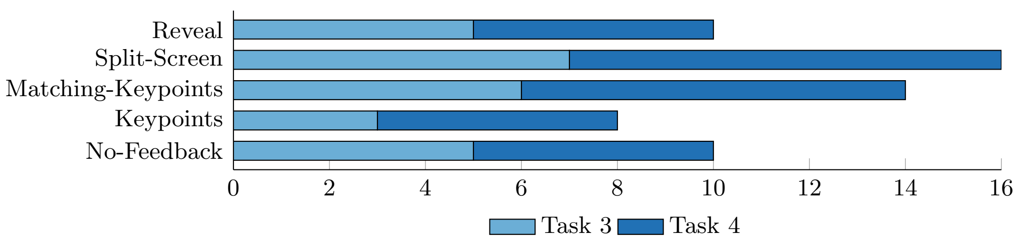


FIGURE A.2: No. Participant responses coded as "correct understanding" when reporting their motivation for background selection in task 3 and task 4.

the Reveal feedback was. R32 on seeing the Reveal feedback remarked, “What’s the bubbly things? That’s too much. [...] That’s crazy!” Other participants shared similar sentiments, going so far as to suggest that something might be wrong with the app: “Why is this camera like this?” (R26) “The focus of the camera is not accurate” (R25), “Why isn’t the background showing up?” (R27), “What is it with [this camera preview]?” (R28).

One participant (R31) described how she felt the feedback increased her workload, describing it as an additional task for which she had to “calculate” the meaning. She was “doing not just one task, but several tasks at the same time for each shot”. When asked, she said that she would have preferred not to have seen the feedback because it was “more work” for her.

A.7 Discussion

We found that the Reveal feedback was helpful to the minority of participants (4) who demonstrated a better understanding of how the underlying algorithms of the app worked. When participants already had a basic idea of what the algorithm was looking for, feedback helped them to deepen their understanding. However, the majority of participants found the feedback to be distracting (7 out of 10 participants) which affected their ability to perform the animation tasks. A common trait observed in the Reveal condition, was participants’ tendency to position the camera closer to the character. This was most common when a less feature rich background was used. In these instances the Reveal visualisation concealed the majority of the users view and only revealed the character. As a result, participants moved the camera closer to frame the visible part of the preview more centrally and to occupy more of the frame. This behaviour made it even more difficult for the app to match pictures.

We conclude that while the metaphor on which the Reveal feedback is based (i.e. the conceptual model) is logical, the majority of participants failed to understand it, nor did it aid their interaction, in fact it was detrimental. These findings highlight the challenges faced by designers to convey a system’s conceptual model through the system image.

Bibliography

Alper T. Alan, Enrico Costanza, Sarvapali D. Ramchurn, Joel Fischer, Tom Rodden, and Nicholas R. Jennings. Tariff agent: Interacting with a future smart energy system at home. *ACM Trans. Comput.-Hum. Interact.*, 23(4):25:1–25:28, August 2016a. ISSN 1073-0516.

Alper T. Alan, Mike Shann, Enrico Costanza, Sarvapali D. Ramchurn, and Sven Seuken. It is too hot: An in-situ study of three designs for heating. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5262–5273, New York, NY, USA, 2016b. ACM. ISBN 978-1-4503-3362-7.

Saleema Amershi, Dan Weld, Mihaela Vorvoreanu, Adam Fourney, Besmira Nushi, Penny Collisson, Jina Suh, Shamsi Iqbal, Paul N. Bennett, Kori Inkpen, Jaime Teevan, Ruth Kikin-Gil, and Eric Horvitz. Guidelines for human-ai interaction. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 3:1–3:13, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2.

Jonas Auda, Dominik Weber, Alexandra Voit, and Stefan Schneegass. Understanding user preferences towards rule-based notification deferral. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI EA '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356213.

Sebastiano Bagnara and Gillian Crampton Smith. *Theories and practice in interaction design*. CRC press, 2006.

Sumit Basu and Alex Pentland. Smart headphones. In *CHI '01 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '01, page 267–268, New York, NY, USA, 2001. Association for Computing Machinery. ISBN 1581133405.

Piraye Bayman and Richard E. Mayer. Instructional manipulation of users' mental models for electronic calculators. *International Journal of Man-Machine Studies*, 20 (2):189 – 199, 1984. ISSN 0020-7373.

Chris Beckmann, Sunny Consolvo, and Anthony LaMarca. Some assembly required: Supporting end-user sensor installation in domestic ubiquitous computing environments. In Nigel Davies, Elizabeth D. Mynatt, and Itiro Siio, editors, *UbiComp 2004:*

Ubiquitous Computing, pages 107–124, Berlin, Heidelberg, 2004. Springer Berlin Heidelberg. ISBN 978-3-540-30119-6.

Victoria Bellotti and Keith Edwards. Intelligibility and accountability: Human considerations in context-aware systems. *Human–Computer Interaction*, 16(2-4):193–212, 2001.

L. Bourikas, E. Costanza, S. Gauthier, P. A. B. James, J. Kittley-Davies, C. Ornaghi, A. Rogers, E. Saadatian, and Y. Huang. Camera-based window-opening estimation in a naturally ventilated office. *Building Research & Information*, 46(2):148–163, 2018.

A.J. Bernheim Brush, Bongshin Lee, Ratul Mahajan, Sharad Agarwal, Stefan Saroiu, and Colin Dixon. Home automation in the wild: Challenges and opportunities. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '11, pages 2115–2124, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0228-9.

Bruce G. Buchanan and Edward H. Shortliffe. *Rule Based Expert Systems: The Mycin Experiments of the Stanford Heuristic Programming Project (The Addison-Wesley Series in Artificial Intelligence)*. Addison-Wesley Longman Publishing Co., Inc., Boston, MA, USA, 1984. ISBN 0201101726.

John M Carroll. *HCI Models, Theories, and Frameworks: Toward a Multidisciplinary Science*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 2003. ISBN 9780080491417.

Nico Castelli, Corinna Ogonowski, Timo Jakobi, Martin Stein, Gunnar Stevens, and Volker Wulf. What happened in my home?: An end-user development approach for smart home data visualization. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 853–866, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9.

Meghan Clark, Mark W. Newman, and Prabal Dutta. Devices and data and agents, oh my: How smart home abstractions prime end-user mental models. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 1(3):44:1–44:26, September 2017. ISSN 2474-9567.

Enrico Costanza, Ben Bedwell, Michael O. Jewell, James Colley, and Tom Rodden. 'a bit like british weather, i suppose': Design and evaluation of the temperature calendar. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4061–4072, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7.

Enrico Costanza, Jacques Panchard, Guillaume Zufferey, Julien Nembrini, Julien Freudiger, Jeffrey Huang, and Jean-Pierre Hubaux. Sensortune: A mobile auditory interface for diy wireless sensor networks. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2317–2326, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-929-9.

Kenneth James Williams Craik. *The nature of explanation*. Cambridge University Press, Cambridge, 1943.

Henriette Cramer, Vanessa Evers, Satyan Ramlal, Maarten van Someren, Lloyd Rutledge, Natalia Stash, Lora Aroyo, and Bob Wielinga. The effects of transparency on trust in and acceptance of a content-based art recommender. *User Modeling and User-Adapted Interaction*, 18(5):455, Aug 2008. ISSN 1573-1391.

Rodrigo de Oliveira, Mauro Cherubini, and Nuria Oliver. Movipill: Improving medication compliance for elders using a mobile persuasive social game. In *Proceedings of the 12th ACM International Conference on Ubiquitous Computing*, UbiComp '10, pages 251–260, New York, NY, USA, 2010. ACM. ISBN 978-1-60558-843-8.

Deloitte. Global mobile consumer trends. Technical report, Deloitte, 2017.

Graham Dove, Kim Halskov, Jodi Forlizzi, and John Zimmerman. Ux design innovation: Challenges for working with machine learning as a design material. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 278–288, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9.

Benedict du Boulay, Tim O'Shea, and John Monk. The black box inside the glass box: presenting computing concepts to novices. *International Journal of Man-Machine Studies*, 14(3):237 – 249, 1981. ISSN 0020-7373.

Mary T. Dzindolet, Scott A. Peterson, Regina A. Pomranky, Linda G. Pierce, and Hall P. Beck. The role of trust in automation reliance. *International Journal of Human-Computer Studies*, 58(6):697 – 718, 2003. ISSN 1071-5819. Trust and Technology.

Herman D'Hooge, Linda Dalton, Helen Shwe, Debra Lieberman, and Claire O'Malley. Smart toys: Brave new world? In *CHI '00 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '00, page 247–248, New York, NY, USA, 2000. Association for Computing Machinery. ISBN 1581132484.

W. Keith Edwards and Rebecca E. Grinter. At home with ubiquitous computing: Seven challenges. In Gregory D. Abowd, Barry Brumitt, and Steven Shafer, editors, *Ubicomp 2001: Ubiquitous Computing*, pages 256–272, Berlin, Heidelberg, 2001. Springer Berlin Heidelberg. ISBN 978-3-540-45427-4.

Yrjö Engeström. *Learning by expanding: An activity-theoretical approach to developmental research*. Cambridge University Press, 1987.

Yrjö Engeström. Expansive learning at work: Toward an activity theoretical reconceptualization. In *Journal of Education and Work*, volume 14, pages 133–156. Routledge, 2001.

Jerry Fails and Dan Olsen. A design tool for camera-based interaction. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '03, pages 449–456, New York, NY, USA, 2003. ACM. ISBN 1-58113-630-7.

Joel E. Fischer, Andy Crabtree, James A. Colley, Tom Rodden, and Enrico Costanza. Data work: How energy advisors and clients make iot data accountable. *Comput. Supported Coop. Work*, 26(4-6):597–626, December 2017. ISSN 0925-9724.

James Fogarty, Carolyn Au, and Scott E. Hudson. Sensing from the basement: A feasibility study of unobtrusive and low-cost home activity recognition. In *Proceedings of the 19th Annual ACM Symposium on User Interface Software and Technology*, UIST '06, pages 91–100, New York, NY, USA, 2006. ACM. ISBN 1-59593-313-1.

Pedro Garcia Garcia, Enrico Costanza, Sarvapali D. Ramchurn, and Jhim Kiel M. Verame. The potential of physical motion cues: Changing people's perception of robots' performance. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, pages 510–518, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4461-6.

Dedre Gentner and Albert L Stevens. *Mental models*. Psychology Press, 1983.

Georg Gerstweiler, Emanuel Vonach, and Hannes Kaufmann. Hymotrack: A mobile ar navigation system for complex indoor environments. *Sensors*, 16(1):17, Dec 2015. ISSN 1424-8220.

Jordan Harold, Kenny R Coventry, Irene Lorenzoni, and Thomas F Shipley. Making sense of time-series data: How language can help identify long-term trends. In *CogSci*, 2015.

Tove Helldin, Ulrika Ohlander, Göran Falkman, and Maria Riveiro. Transparency of automated combat classification. In *Engineering Psychology and Cognitive Ergonomics*, pages 22–33. Springer, 2014. ISBN 978-3-319-07515-0.

Michael E. PorterJames E. Heppelmann, Marco Iansiti, Karim R. Lakhani, and Michael E. Porter. How smart, connected products are transforming competition, Mar 2017.

Jonathan L. Herlocker, Joseph A. Konstan, and John Riedl. Explaining collaborative filtering recommendations. In *CSCW '00: Proceedings of the 2000 ACM Conference on Computer Supported Cooperative Work*, pages 241–250, New York, NY, USA, 2000. ACM. ISBN 1-58113-222-0.

Marc Hesenius, Ingo Börsting, Ole Meyer, and Volker Gruhn. Don't panic!: Guiding pedestrians in autonomous traffic with augmented reality. In *Proceedings of the 20th International Conference on Human-Computer Interaction with Mobile Devices and Services Adjunct*, MobileHCI '18, pages 261–268, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5941-2.

Yang Hu, Dominique Tilke, Taylor Adams, Aaron S. Crandall, Diane J. Cook, and Maureen Schmitter-Edgecombe. Smart home in a box: usability study for a large scale self-installation of smart home technologies. *Journal of Reliable Intelligent Environments*, 2(2):93–106, Jul 2016.

Justin Huang and Maya Cakmak. Supporting mental model accuracy in trigger-action programming. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 215–225, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3574-4.

Edwin L. Hutchins, James D. Hollan, and Donald A. Norman. Direct manipulation interfaces. *Hum.-Comput. Interact.*, 1(4):311–338, December 1985. ISSN 0737-0024.

Timo Jakobi, Gunnar Stevens, Nico Castelli, Corinna Ogonowski, Florian Schaub, Nils Vindice, Dave Randall, Peter Tolmie, and Volker Wulf. Evolving needs in iot control and accountability: A longitudinal study on smart home intelligibility. *Proc. ACM Interact. Mob. Wearable Ubiquitous Technol.*, 2(4):171:1–171:28, December 2018. ISSN 2474-9567.

N. R. Jennings, L. Moreau, D. Nicholson, S. Ramchurn, S. Roberts, T. Rodden, and A. Rogers. Human-agent collectives. *Commun. ACM*, 57(12):80–88, November 2014. ISSN 0001-0782.

Rikke Hagensby Jensen, Yolande Strengers, Jesper Kjeldskov, Larissa Nicholls, and Mikael B. Skov. Designing the desirable smart home: A study of household experiences and energy consumption impacts. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, New York, NY, USA, 2018. Association for Computing Machinery. ISBN 9781450356206.

Michael O. Jewell, Enrico Costanza, and Jacob Kittley-Davies. Connecting the things to the internet: An evaluation of four configuration strategies for wi-fi devices with minimal user interfaces. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 767–778, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3574-4.

Jeff Johnson and Austin Henderson. Conceptual models: Begin by designing what to design. *interactions*, 9(1):25–32, January 2002. ISSN 1072-5520.

Philip N Johnson-Laird. *Mental models: Towards a cognitive science of language, inference, and consciousness*. Harvard University Press, 1983. ISBN 0674568818.

Natalie Jones, Helen Ross, Timothy Lynam, Pascal Perez, and Anne Leitch. Mental models: an interdisciplinary synthesis of theory and methods. *Ecology and Society*, 16(1), 2011.

Victor Kaptelinin. Activity theory, 2020.

Victor Kaptelinin and Bonnie Nardi. Activity theory as a framework for human-technology interaction research. *Mind, Culture, and Activity*, 25(1):3–5, 2018.

Jun Kato, Sean McDirmid, and Xiang Cao. Dejavu: Integrated support for developing interactive camera-based programs. In *Proceedings of the 25th Annual ACM Symposium on User Interface Software and Technology*, UIST '12, pages 189–196, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1580-7.

Aqeel H. Kazmi, Michael J. O'grady, Declan T. Delaney, Antonio G. Ruzzelli, and Gregory M. P. O'hare. A review of wireless-sensor-network-enabled building energy management systems. *ACM Trans. Sen. Netw.*, 10(4):66:1–66:43, June 2014. ISSN 1550-4859.

David E. Kieras and Susan Bovair. The role of a mental model in learning to operate a device. *Cognitive Science*, 8(3):255 – 273, 1984. ISSN 0364-0213.

Jacob Kittley-Davies, Ahmed Alqaraawi, Rayoung Yang, Enrico Costanza, Alex Rogers, and Sebastian Stein. Evaluating the effect of feedback from different computer vision processing stages: A comparative lab study. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, pages 43:1–43:12, New York, NY, USA, 2019. ACM. ISBN 978-1-4503-5970-2.

René F. Kizilcec. How much information?: Effects of transparency on trust in an algorithmic interface. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 2390–2395, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7.

A. Komninos, J. Besharat, V. Stefanis, and J. Garofalakis. Perceptibility of mobile notification modalities during multitasking in smart environments. In *2018 14th International Conference on Intelligent Environments (IE)*, pages 17–24, June 2018.

Ha-Kyung Kong, Zhicheng Liu, and Karrie Karahalios. Frames and slants in titles of visualizations on controversial topics. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 438:1–438:12, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6.

Jeamin Koo, Jungsuk Kwac, Wendy Ju, Martin Steinert, Larry Leifer, and Clifford Nass. Why did my car just do that? explaining semi-autonomous driving actions to improve driver understanding, trust, and performance. *International Journal on Interactive Design and Manufacturing (IJIDeM)*, 9(4):269–275, Nov 2015. ISSN 1955-2505.

Gerd Kortuem, Fahim Kawsar, Vasughi Sundramoorthy, and Daniel Fitton. Smart objects as building blocks for the internet of things. *IEEE Internet Computing*, 14(1): 44–51, 2010.

Josua Krause, Adam Perer, and Kenney Ng. Interacting with predictions: Visual inspection of black-box machine learning models. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 5686–5697, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-3362-7.

T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, and W. K. Wong. Too much, too little, or just right? ways explanations impact end users' mental models. In *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, pages 3–10, Sept 2013.

Todd Kulesza, Margaret Burnett, Weng-Keen Wong, and Simone Stumpf. Principles of explanatory debugging to personalize interactive machine learning. In *Proceedings of the 20th International Conference on Intelligent User Interfaces*, IUI '15, pages 126–137, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3306-1.

Todd Kulesza, Simone Stumpf, Margaret Burnett, and Irwin Kwan. Tell me more?: The effects of mental model soundness on personalizing an intelligent agent. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '12, pages 1–10, New York, NY, USA, 2012. ACM. ISBN 978-1-4503-1015-4.

Gierad Laput, Yang Zhang, and Chris Harrison. Synthetic sensors: Towards general-purpose sensing. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 3986–3999, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9.

Amanda Lazar, Christian Koehler, Joshua Tanenbaum, and David H. Nguyen. Why we use and abandon smart devices. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, pages 635–646, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3574-4.

Michael Lewis. Designing for human-agent interaction. *AI Magazine*, 19(2):67, Jun. 1998.

Frank Li, Zakir Durumeric, Jakub Czyz, Mohammad Karami, Michael Bailey, Damon McCoy, Stefan Savage, and Vern Paxson. You've got vulnerability: Exploring effective vulnerability notifications. In *25th USENIX Security Symposium (USENIX Security 16)*, pages 1033–1050, Austin, TX, 2016. USENIX Association. ISBN 978-1-931971-32-4.

Brian Y. Lim and Anind K. Dey. Investigating intelligibility for uncertain context-aware applications. In *Proceedings of the 13th International Conference on Ubiquitous Computing*, UbiComp '11, pages 415–424, New York, NY, USA, 2011. ACM. ISBN 978-1-4503-0630-0.

Brian Y. Lim, Anind K. Dey, and Daniel Avrahami. Why and why not explanations improve the intelligibility of context-aware intelligent systems. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '09, pages 2119–2128, New York, NY, USA, 2009. ACM. ISBN 978-1-60558-246-7.

Javier Lopez, Ruben Rios, Feng Bao, and Guilin Wang. Evolving privacy: From sensors to the internet of things. *Future Generation Computer Systems*, 75:46 – 57, 2017. ISSN 0167-739X.

Dan Maynes-Aminzade, Terry Winograd, and Takeo Igarashi. Eyepatch: Prototyping camera-based interaction through examples. In *Proceedings of the 20th Annual ACM Symposium on User Interface Software and Technology*, UIST '07, pages 33–42, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-679-0.

Abhinav Mehrotra, Robert Hendley, and Mirco Musolesi. Prefminer: Mining user's preferences for intelligent mobile notification management. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, page 1223–1234, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344616.

Jussi Mikkonen and Riikka Townsend. Frequency-based design of smart textiles. In *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems*, CHI '19, New York, NY, USA, 2019. Association for Computing Machinery. ISBN 9781450359702.

Neville Moray. Mental models in theory and practice. *Attention and performance XVII: Cognitive regulation of performance: Interaction of theory and application*, pages 223–258, 1999.

Chandrakana Nandi and Michael D. Ernst. Automatic trigger generation for rule-based smart homes. In *Proceedings of the 2016 ACM Workshop on Programming Languages and Analysis for Security*, PLAS '16, pages 97–102, New York, NY, USA, 2016. ACM. ISBN 978-1-4503-4574-3.

Jakob Nielsen. *Usability Engineering*. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1994. ISBN 9780080520292.

J. Nievergelt and J. Weydert. Sites, modes, and trails: Telling the user of an interactive system where he is, what he can do, and how to get to places (excerpt). In R. M. Baecker and W. A. S. Buxton, editors, *Human-computer Interaction*, pages 438–441. Morgan Kaufmann Publishers Inc., San Francisco, CA, USA, 1987. ISBN 0-934613-24-9.

Marius Noreikis, Yu Xiao, and Antti Ylä-Jääski. Seenav: Seamless and energy-efficient indoor navigation using augmented reality. In *Proceedings of the on Thematic Workshops of ACM Multimedia 2017*, Thematic Workshops '17, pages 186–193, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-5416-5.

Don Norman. *The design of everyday things: Revised and expanded edition*. Basic books, 2013.

Donald Norman. *On the relationship between conceptual and mental models. In D. Gentner and A. L. Stevens Eds 'Mental Models'*. Psychology Press, 1983.

Donald A Norman. *Emotional design: Why we love (or hate) everyday things*. Basic Civitas Books, 2004.

Donald A Norman. *Living with complexity.* MIT press, 2010. ISBN 9780262528948.

Donald A. Norman and Stephen W. Draper. *User Centered System Design; New Perspectives on Human-Computer Interaction.* L. Erlbaum Associates Inc., USA, 1986. ISBN 0898597811.

David G. Novick and Karen Ward. Why don't people read the manual? In *Proceedings of the 24th Annual ACM International Conference on Design of Communication*, SIGDOC '06, pages 11–18, New York, NY, USA, 2006. ACM. ISBN 1-59593-523-1.

Carmine Ornaghi, Enrico Costanza, Jacob Kittley-Davies, Leonidas Bourikas, Victoria Aragon, and Patrick A.B. James. The effect of behavioural interventions on energy conservation in naturally ventilated offices. *Energy Economics*, 74:582 – 591, 2018. ISSN 0140-9883.

Donggun Park, Yu Shin Lee, Sejin Song, Ilsun Rhiu, Sanghyun Kwon, Yongdae An, and Myung Hwan Yun. User centered gesture development for smart lighting. In *Proceedings of HCI Korea*, HCIK '16, page 146–150, Seoul, KOR, 2016. Hanbit Media, Inc. ISBN 9788968487910.

Kayur Patel, Naomi Bancroft, Steven M. Drucker, James Fogarty, Andrew J. Ko, and James Landay. Gestalt: Integrated support for implementation and analysis in machine learning. In *Proceedings of the 23Nd Annual ACM Symposium on User Interface Software and Technology*, UIST '10, pages 37–46, New York, NY, USA, 2010a. ACM. ISBN 978-1-4503-0271-5.

Kayur Patel, James Fogarty, James A. Landay, and Beverly Harrison. Investigating statistical machine learning as a tool for software development. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '08, pages 667–676, New York, NY, USA, 2008. ACM. ISBN 978-1-60558-011-1.

Shwetak N. Patel, Sidhant Gupta, and Matthew S. Reynolds. The design and evaluation of an end-user-deployable, whole house, contactless power consumption sensor. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '10, pages 2471–2480, New York, NY, USA, 2010b. ACM. ISBN 978-1-60558-929-9.

Stephen J Payne. Mental models in human-computer interaction. *The Human-Computer Interaction Handbook*, pages 63–75, 2007.

Ioannis Politis, Stephen Brewster, and Frank Pollick. To beep or not to beep?: Comparing abstract versus language-based multimodal driver displays. In *Proceedings of the 33rd Annual ACM Conference on Human Factors in Computing Systems*, CHI '15, pages 3971–3980, New York, NY, USA, 2015. ACM. ISBN 978-1-4503-3145-6.

S. Pradhan, L. Qiu, A. Parate, and K. Kim. Understanding and managing notifications. In *IEEE INFOCOM 2017 - IEEE Conference on Computer Communications*, pages 1–9, May 2017.

K. Renaud and R. Cooper. Feedback in human-computer interaction - characteristics and recommendations. *South African Computer Journal*, 2000(26):105–114, 2000. ISSN 1015-7999.

Yvonne Rogers. Hci theory: Classical. *Modern, and Contemporary, Morgan & Claypool Publishers*, 2012.

Yvonne Rogers, Helen Sharp, and Jenny Preece. *Interaction Design: Beyond Human - Computer Interaction*. Wiley Publishing, 3rd edition, 2011. ISBN 0470665769, 9780470665763.

William B Rouse and Nancy M Morris. On looking into the black box: Prospects and limits in the search for mental models. *Psychological bulletin*, 100(3):349, 1986.

Ethan Rublee, Vincent Rabaud, Kurt Konolige, and Gary Bradski. Orb: An efficient alternative to sift or surf. In *Proceedings of the 2011 International Conference on Computer Vision*, ICCV '11, pages 2564–2571, Washington, DC, USA, 2011. IEEE Computer Society. ISBN 978-1-4577-1101-5.

Dan Saffer and Larry Tesler. Larry tesler interview: The laws of interaction design, 2007.

Martina Angela Sasse. *Eliciting and describing users' models of computer systems*. PhD thesis, University of Birmingham, 1997.

E. Sezer, D. Romero, F. Guedea, M. Macchi, and C. Emmanouilidis. An industry 4.0-enabled low cost predictive maintenance approach for smes. In *2018 IEEE International Conference on Engineering, Technology and Innovation (ICE/ITMC)*, pages 1–8, June 2018.

Frank Siegemund. *A Context-Aware Communication Platform for Smart Objects*, pages 69–86. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

Manuel Silverio-Fernández, Suresh Renukappa, and Subashini Suresh. What is a smart device? - a conceptualisation within the paradigm of the internet of things. *Visualization in Engineering*, 6(1):1–10, 5 2018.

N. A. Streitz, C. Rocker, T. Prante, D. van Alphen, R. Stenzel, and C. Magerkurth. Designing smart artifacts for smart environments. *Computer*, 38(3):41–49, March 2005. ISSN 0018-9162.

Eulim Sull and Youn-kyung Lim. Designing health-promoting technologies with iot at home. In *Extended Abstracts of the 2018 CHI Conference on Human Factors in*

Computing Systems, CHI EA '18, pages LBW083:1–LBW083:6, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5621-3.

Louis H Sullivan. The tall office building artistically considered. *Lippincott's Magazine*, 57(3):406, 1896.

Craig W Thompson. Smart devices and soft controllers. *IEEE Internet Computing*, 9 (1):82–85, 2005.

Joe Tullio, Anind K. Dey, Jason Chalecki, and James Fogarty. How it works: A field study of non-technical users interacting with an intelligent system. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '07, pages 31–40, New York, NY, USA, 2007. ACM. ISBN 978-1-59593-593-9.

Tom R Tyler. The psychology of procedural justice: a test of the group-value model. *Journal of personality and social psychology*, 57(5):830, 1989.

Jhim Kiel M. Verame, Enrico Costanza, and Sarvapali D. Ramchurn. The effect of displaying system confidence information on the usage of autonomous systems for non-specialist applications: A lab study. In *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, CHI '16, pages 4908–4920, New York, NY, USA, 2016a. ACM. ISBN 978-1-4503-3362-7.

Jhim Kiel M. Verame, Jacob Kittley-Davies, Enrico Costanza, and Kirk Martinez. Designing natural language output for the iot. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing: Adjunct*, UbiComp '16, pages 1584–1589, New York, NY, USA, 2016b. ACM. ISBN 978-1-4503-4462-3.

Dhaval Vyas, Wim Poelman, Anton Nijholt, and Arnout De Bruijn. Smart material interfaces: A new form of physical interaction. In *CHI '12 Extended Abstracts on Human Factors in Computing Systems*, CHI EA '12, page 1721–1726, New York, NY, USA, 2012. Association for Computing Machinery. ISBN 9781450310161.

Mark Weiser. The computer for the 21 st century. *Scientific American*, 265(3):94–105, 1991. ISSN 00368733, 19467087.

Muchen Wu, Parth H. Pathak, and Prasant Mohapatra. Monitoring building door events using barometer sensor in smartphones. In *Proceedings of the 2015 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '15, page 319–323, New York, NY, USA, 2015. Association for Computing Machinery. ISBN 9781450335744.

Qian Yang, Nikola Banovic, and John Zimmerman. Mapping machine learning advances from hci research to reveal starting places for design innovation. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, CHI '18, pages 130:1–130:11, New York, NY, USA, 2018. ACM. ISBN 978-1-4503-5620-6.

Rayoung Yang and Mark W. Newman. Learning from a learning thermostat: Lessons for intelligent systems for the home. In *Proceedings of the 2013 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '13, pages 93–102, New York, NY, USA, 2013. ACM. ISBN 978-1-4503-1770-2.

Rayoung Yang, Mark W. Newman, and Jodi Forlizzi. Making sustainability sustainable: Challenges in the design of eco-interaction technologies. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, CHI '14, pages 823–832, New York, NY, USA, 2014. ACM. ISBN 978-1-4503-2473-1.

Svetlana Yarosh and Pamela Zave. Locked or not?: Mental models of iot feature interaction. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems*, CHI '17, pages 2993–2997, New York, NY, USA, 2017. ACM. ISBN 978-1-4503-4655-9.

Yuhang Zhao, Sarit Szpiro, Jonathan Knighten, and Shiri Azenkot. Cuesee: Exploring visual cues for people with low vision to facilitate a visual search task. In *Proceedings of the 2016 ACM International Joint Conference on Pervasive and Ubiquitous Computing*, UbiComp '16, page 73–84, New York, NY, USA, 2016. Association for Computing Machinery. ISBN 9781450344616.

Kening Zhu, Xiaojuan Ma, Haoyuan Chen, and Miaoyin Liang. Tripartite effects: Exploring users' mental model of mobile gestures under the influence of operation, handheld posture, and interaction space. *International Journal of Human–Computer Interaction*, 33(6):443–459, 2017.

Davide Zilli, Oliver Parson, Geoff V Merrett, and Alex Rogers. A hidden markov model-based acoustic cicada detector for crowdsourced smartphone biodiversity monitoring. In *Proceedings of the Twenty-Third International Joint Conference on Artificial Intelligence*, IJCAI '13, page 2945–2951. AAAI Press, 2013. ISBN 9781577356332.