Understanding the Role of Image Recognition in Mobile Tour Guides

Nigel Davies, Keith Cheverst, Alan Dix
Computing Department
Lancaster University
Lancaster, UK
nigel,kc,@comp.lancs.ac.uk,
alan@hcibook.com

Andre Hesse
Technische Universität Ilmenau
Ilmenau
Germany
andre.hesse@stud.tu-ilmenau.de

ABSTRACT

Users of mobile tour guides often express a strong desire for the system to be able to provide information on arbitrary objects they encounter during their visit – akin to pointing to a building or attraction and saying "what's that ?" to a human tour guide. This paper reports on a field study in which we investigated user reaction to the use of digital image capture and recognition to support such functionality. Our results provide an insight into usage patterns and likely user reaction to mobile tour guides that use digital photography for real-time object recognition. These results include the counter-intuitive observation that a significant class of users appear happy to use image recognition even when this is a more complex, lengthy and error-prone process than traditional solutions. Careful analysis of user behavior during the field trails also provides evidence that it may be possible to classify tourists according to the methods by which they prefer to acquire information about tourist attractions in their vicinity. If shown to be generally true these results have important implications for designers of future mobile tour guide systems.

Categories and Subject Descriptors

H.5.2 [Information interfaces and presentation (e.g., HCI)]: User Interfaces; H.5.1 [Information interfaces and presentation (e.g., HCI)]: Multimedia

General Terms

Human Factors.

Keywords

Mobile tour guides, mobile camera phones, object recognition, user experience, user evaluation.

1. INTRODUCTION

The ability to identify and provide information on arbitrary objects in a city is an important requirement for any mobile tour guide system. In essence, the idea is to be able to reproduce the act of pointing at an object (e.g. a building or statue) and asking a human tour guide "What's that ?". In our experience of building and testing tour guide systems during the past six years this is an extremely common usage model and a feature that tourists often request – exceeding any desire users express for planned tours with heavily structured information.

In [5] we reported on a technique for providing this functionality within the Lancaster GUIDE system. GUIDE provides a complete mobile tour guide including features such as descriptions of local areas, structured tours, messaging between tourists and interactive services (e.g. booking theatre tickets) [4]. The GUIDE approach to providing users with information about objects of interest that they see relies on a dialogue with users to identify the object in question. More specifically, when users see an object that is of interest to them they can ask the system to "Tell me about something I can see" [5]. The GUIDE unit then determines their approximate position (to within about 100m) and asks the user whether they are "looking at something close by or far away" (figure 1). From these two pieces of coarse grained information – an approximate location and the distance to the object of interest the system builds a set of thumbnails of likely objects. This list is shown to users who can then select the object that they are interested in and obtain further information about it from the system.

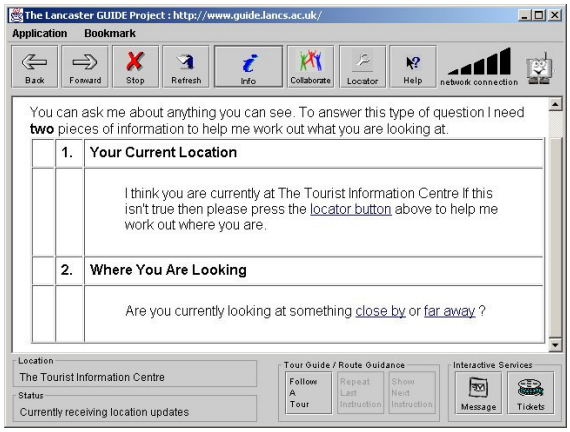

Figure 1. Interacting with the GUIDE system to identify objects of interest.

This system has a number of desirable properties – it obviates the need to attach physical tags to objects (e.g. RFID tags), does not require the construction of complex geometric models of the city and does not rely on highly accurate positional information. In

MobileHCI'05, September 19–22, 2005, Salzburg, Austria.

fact, as we point out in [5], no matter how accurate a positional fix one can obtain on a tourist it is usually impossible to identify uniquely the object that is the focus of their attention (consider, for example, a tourist standing at the top of a hill looking out over a city – one knows exactly where they are but has no idea what they are looking at).

Reports of extensive trials of the Lancaster GUIDE system (e.g. in [4]) have shown that users respond well to this form of interactive dialogue with the system and that the system is usually able to provide users with the information they require, i.e. the system is able to include the correct object in the set of thumbnails it displays to the user. Moreover, users appear to accrue benefits from seeing the thumbnails of other objects of interest in the neighborhood of their target.

Motivated by the widespread deployment of digital cameras in devices such as Personal Digital Assistants (PDAs) and phones and more recent results from the GUIDE project which report that user preferences have shifted from large form-factor devices with text and images to small form factor devices that use audio as the principle means of information delivery [3] in this paper we explore user reactions to the use of digital image capture and recognition techniques. If users are receptive to such technology then we can envisage creating mobile tour guides that have cameras as one of their principal input mechanisms – enabling guide systems with new form factors and interface designs to be developed.

In particular, in this paper we investigate through field trials of a "Wizard of Oz" system users' reaction to the use of digital image capture and recognition to identify objects of interest – allowing the system to answer "what's that?" queries simply by the user taking photographs of the relevant objects. Whilst necessarily tentative in areas, the results of this work provide an insight into usage patterns and likely user reaction to mobile tour guides that use digital photography for real-time object recognition and highlight potentially critical issues for future design. These results include the counter-intuitive observation that a significant class of users appear happy to interact with the system by taking photographs of objects even when this is a more complex, lengthy and error-prone process than traditional dialogue based solutions. Careful analysis of user behavior during the field trails also provides evidence that it might be possible to classify tourists according to the methods by which they prefer to acquire information about tourist attractions in their vicinity. If shown to be generally true these results have important implications for designers of future mobile tour guide systems.

2. METHODOLOGY

We have chosen to study real tourists, but with an experimental system designed to investigate a particular issue: the usage patterns of camera-based "what's that?" interaction. We thus sit methodologically between controlled laboratory studies on the one hand and more open ethnographic approaches on the other. There is always a tension between ecological validity and analytic power and there is currently some debate over the merits of lab-based versus field-based studies for evaluating the usability of mobile systems. It has been argued by other researchers in the mobile HCI community that "the added value of conducting usability evaluations in the field is very little" [8]. However, we would argue strongly that had our study not been 'situated in the field' we would not have observed the patterns of voluntarily use that were outside our expectations.

There are of course limitations with such an approach: the practicalities of setting up an experiment in a 'wild' environment, obtaining sufficient feedback from busy users etc. The last of these is particularly problematic since participating in our experiment requires real tourists to sacrifice precious vacation time. Whilst they may have volunteered to use the device for a while, it was clear that once they handed back the unit they considered the experience over and wanted to get away. With only a few hours to 'visit' Lancaster long post-use interviews were impossible, hence our analysis dwells strongly on the logged data as well as more qualitative findings.

For similar reasons the average usage time was short (see section 4.2), reflecting the tourists' real 'guerilla' behaviour: short bursts at one activity before moving elsewhere. However, as tourist guide systems are 'walk up and use', it is the first few minutes of use that are crucial; preferences and patterns of use are established quickly and as quickly forgotten when the device is returned.

One of the problems therefore is how to get useful measurements at all. In a laboratory we could create an artificial task, give subjects a variety of techniques to use and measure some form of performance for each. However, for tourists the real issues involve a combination of performance, enjoyment etc. which are harder to evaluate. Although we were administering a post-test questionnaire, we also wanted some more objective measures. In part we obtained this by producing a system that could be used in both a location-based "what's near?" mode similar to the conventional GUIDE dialogue and also a camera-based "what's that ?" mode. The tourists' moment-to-moment *choices* between location and camera modes was thus a *dependent variable* in our experiment rather than an independent variable as it would likely be in a performance comparison. As this choice of mode is important to our experiment it is crucial that the design of the interaction does not unduly influence the tourists' choices. To achieve this, the two interaction techniques were designed to be as similar as possible except in the crucial issue of camera vs. location based dialogue. This meant that we deliberately eschewed 'tweaks' such as saving pictures and in general pared down the system as much as possible given the constraint of making it sufficiently useful and usable for real use.

In a real system there would be many extra features that would alter use. For example, if object recognition were a side-effect of taking a photograph, then we may see use of this as a form of incidental interaction [6]. For a more formative evaluation to improve tour guide systems in their entirety we would undoubtedly have looked at such features. However, in this study our desire for a more controlled exploration of the use of image recognition led us towards greater control in the ecological spectrum. In summary, this is a study using *real* tourists but designed to explore a particular issue (the use of image recognition). As such it has inherent limitations but we believe represents a methodologically necessary stage.

3. THE EXPERIMENTAL PLATFORM

In order to examine the use of image recognition in future tour guides we developed a simple mobile tour guide application that could be used by city visitors (see Figure 2). The system offers

very basic tour guide functionality: users can find out about objects of interest in one of two ways:

1. engaging in a dialogue with the system using the GUIDE interaction model described in [5].
2. taking a photograph of the object of interest.

Users were free to swap between the two modes of use at any time. In both cases the user is returned a set of images of possible matching objects. Tapping on one of these images provides more information (where available) in the form of textual and audio descriptions of the object (Figure 3). We reused elements of content we had developed for trials described in [3] and [4] and known to meet the needs of tourists.

Figure 2. Prototype in use

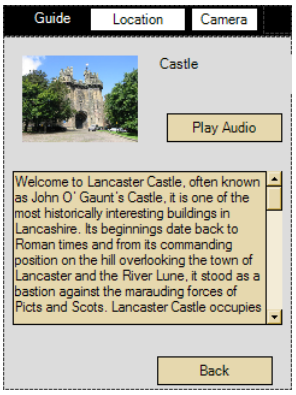

Figure 3. Information screen

In this study we were specifically interested in examining user interaction issues for a system based on digital image capture and recognition rather than attempting to explore the use of specific image recognition technologies. As a result we adopted a "Wizard of Oz" approach to providing the required functionality. Specifically, photographs taken by the user were transmitted to a tablet PC where a researcher used a custom developed application called GECS (GUIDE Experimenter's Control System) to dynamically select an appropriate candidate *image set* to return to the user. In this way we were able to adjust the perceived accuracy and timeliness of the recognition process in order to explore how these factors impacted user reaction. The same approach was used for returning candidate image sets in response to GUIDE style dialogue interactions – we did not use a GPS compass or wireless network for positioning but instead relied on observations to return an appropriate image set.

We did not support additional functionality such as general information on an area, route-guidance, ticket booking or messaging; all of which have been deployed and evaluated by researchers in systems such as GUIDE [4].

Hardware. The Client used a PDA (HP iPAQ 5550) with built-in 802.11 networking, an attached mobile camera (HP Photosmart Mobile Camera) with 320*240 resolution and an expansion-pack with a 4.66 GB PC card hard-drive to store the application content. The resulting tourist system (iPAQ and accessories) weighs approx. 400g with a battery life of approx. 2 hours during field trail conditions. We anticipate that the application could be ported to a camera phone without significant difficulty.

The experimenter followed the tourist and carried a tablet-PC (Fujitsu-Siemens Stylistic ST5011), also with built-in 802.11 networking. The PDA and the tablet PC communicated in ad-hoc-mode, removing the need to have wireless network infrastructure in the area where the field trials were conducted. The tablet PC ran the GECS application that enabled the researcher to control many aspects of the experience of the user interacting with the mobile unit.

Client Interface. When users are first handed the unit they see the project logo and a welcome message that asks them to select from one of two modes of use for the system: *location mode* or *camera mode* (in the experiments described later in this paper each of these modes is explained to users before they are given the unit). Users can swap between the two modes at any point during the experiment.

Location Mode. When this mode is selected it displays a welcome message explaining to the user that the unit "can tell you about objects that are around you" and asks the user to tell the system if they are looking at something nearby or faraway.

Once the user makes an appropriate selection the welcome screen disappears and a new message appears telling the user that the system needs some time to work out their physical location and a wait-cursor is displayed. During this time the system sends a request for an image set to the tablet PC running the GECS application. Once an appropriate image set has been selected by the experimenter it is returned to the client. At this point the wait-cursor disappears and the system explains to the user that there is more than one object that could be the subject of their enquiry and that they will need to identify the object themselves from a set of pictures. The user can then move through a list of pictures (one per screen) using Previous and Next buttons (see Figure 4). Once the user sees an image of the object that is of interest to them they can click on the image and will be taken to a screen that provides more information about that object in both textual and audio format (see Figure 3). Users can then either continue to browse the image set or issue another query about an object that they can see.

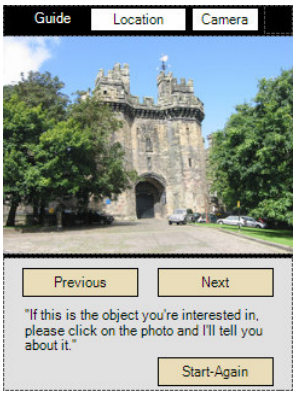

Figure 4. Location mode

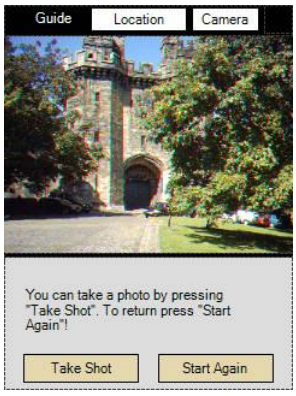

Figure 5. Viewfinder screen

Camera Mode. When this mode is selected it displays a welcome message, this time explaining to the user that they should "Press the 'Take Photo' button to take a picture of an object of interest". When the user presses the *Take Photo* button a new screen is displayed with a 240*180 viewfinder window showing the current view of the camera and a new button labeled *Take Shot* that

enables the user to take the photograph once the camera is pointing at the appropriate object (Figure 5).

Once the user has taken a photograph it is displayed on the screen and the user is provided with an option to either take another photograph or to ask the system about the object in the photograph (by pressing a button labeled *What's this?*) (Figure 6).

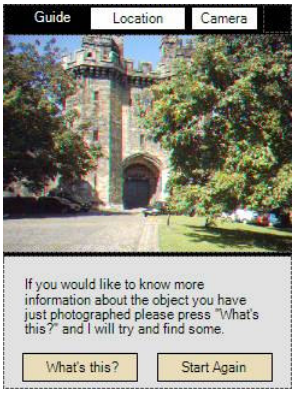

Figure 6. "What's this?"

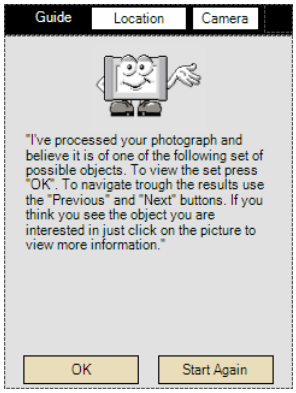

Figure 7. Result collection

At this point the system sends the photograph to GECS and requests an image set to display to the user. Once an appropriate image set has been selected by the experimenter it is returned to the client and, after a configurable delay, a new screen appears telling the user that the system could not uniquely identify the object they photographed and that they should select the correct object from a set of images they will be shown (Figure 7). The user interaction for browsing through the image set, obtaining further information and returning to earlier screens in this mode is similar to that described in the previous section on location mode.

The GECS Application. This provides facilities for the researcher controlling the experiment to return appropriate responses to the mobile unit based on user requests. The experimenter can select appropriate image sets to return to the client as a result of a query in either location or camera mode. These image sets are chosen from a list of possible image sets created prior to the experiment using a separate tool to construct image sets from collections of photographs, text and audio data. Each image set is a small collection of entries (typically 5) with each entry consisting of an image and links to further information about that image. The perceived accuracy of the system's responses to requests can be adjusted by returning more or less appropriate image sets. In addition, GECS allows responses to be delayed by a specified number of seconds – simulating processing delays.

4. FIELD TRIAL

4.1 Objective and Method

Our aim was to gain early insight into user reaction to the use of digital image recognition to identify objects in an electronic tour guide. To help provide a reference point for assessing user reactions to image recognition we designed our experiment to enable us to capture user preferences for the two modes described above.

We followed broadly the same experimental methodology successfully used for GUIDE trials [3], [4] in which a combination of observation, experimental log data and semi-structured interviews were used to capture data.

The experiments were conducted by researchers operating in pairs. One researcher was the primary interaction point for visitors: making introductions, explaining the system to the tourist, helping them with problems, observing their behavior and conducting the post experiment interview. The second researcher acted as the "wizard", controlling the experiment through GECS. In the remainder of this paper we will refer to the researchers as the *observer* and the *GECS operator* respectively.

Tourists were selected at random in an area that includes a number of tourist attractions (e.g. a Castle, a Priory and a Museum) as well as the Tourist Information Centre which we used as a base for our work. Of those tourists we stopped and asked to participate in a trial of the new system roughly 50% agreed and hence the subjects are, inevitably, self-selecting to include only those with an interest in trying a new electronic guide. This is likely to be reflected in generally more positive responses than might be the case with a truly random selection of

Table 1. Frequency analysis for final preference against start mode, end mode and time spent in each mode, (a), (b) & (d) n.s., (c) sig. p<0.005 (Chi sq. = 12.98, 3 df.)

	Preference →	Location	Camera	Not Sure	Only Used One
(a) Start Mode	Location	7	4	2	3
	Camera	3	6	0	2
(b) End Mode	Location	5	4	0	3
	Camera	5	6	2	2
(c) >50% time in mode	Location	9	1	1	3
	Camera	1	9	1	2
(d) Error condition	No errors	4	6	0	4
	Errors	6	4	2	1

tourists. We did not select on the basis of age, gender or whether the visitor was alone or in a group.

Once the subject had agreed to participate they were given a brief introduction to the system by the observer. The different modes were explained to the subject and they were told they could use it for as long as they would like to explore the local area. The aim of the experiment was stated only as being to gain user feedback on the new system. The GECS operator was introduced as a colleague who would be helping with the experiment. During the briefing the subjects were always told that "the system" would perform the image recognition, etc. – the role of the GECS operator in the actual performance of the system was not mentioned. This is because we were concerned that subjects would respond differently (and have different expectations) if they felt a human rather than a machine was doing the recognition – though we did not attempt to test this hypothesis.

Subjects were then handed the unit and were free to explore the area. The observer remained with the subject to help with any difficulties that they may experience and to observe the subject's behavior. The GECS operator retreated to a discrete distance (typically about 15m) which they maintained throughout the experiment, following the subjects when they explored the area. We conducted a range of experiments involving returning image sets designed to simulate image recognition algorithms of varying degrees of accuracy and latency but in all cases the parameters were fixed throughout the duration of an individual experiment with a subject, i.e. subjects experienced consistent levels of performance from the system.

At the conclusion of the experiment subjects were asked a series of questions designed to provide data on their overall impression of the system and specifically on their reactions to each mode of operation. Demographic information (age, gender and solo or group tourist) was also captured. At this point, with the experiment and data capture complete the subjects were debriefed and the role of the GECS operator explained. Finally, once the subjects had departed the observer completed a separate questionnaire relating to their observations of the subject and the environment (weather, unusual events, level of activity in the area etc.) during the experiment.

4.2 Results

Participation. We recruited a total of 27 participants for our study: 6 female, 21 male with a wide range of ages including one septuagenarian. Some of the participants (16) were alone while others (11) were in groups. Note, we did not try to balance gender, age, cultural background, etc., in the experiment. This is because the range of factors is too large in a group of real tourists to effectively balance and because we rely on volunteers. In practice we did not notice a significant difference in behavior between females and males when using the system. All the subjects were visitors to the city and unfamiliar with the tour guide system. Subjects interacted with the system for an average of 6 mins. 37 secs. but there was a significant deviation up to 14 minutes in total.

In all but one case the subjects remained totally unaware of the role of the GECS operator in the experiment. When debriefed most subjects expressed genuine surprise that the image recognition was not being carried out by the system. The only subject that enquired about the role of the GECS operator during the experiment was a Psychology student who guessed that the operator was playing an active role in the experiment but assumed it was that of an observer.

Overall Mode Usage. 22 out of 27 subjects used both the location mode and the camera mode. 16 started with location mode and 11 with camera mode. Overall 74% of subjects expressed a preference for either the location mode or the camera mode with the remainder being either unsure of which mode they preferred (7%) or having used only one mode during their trial (19%). Those who expressed a preference were divided exactly equally between the two modes. We detected no relationship between the mode that the subject started with and the mode they said they preferred when interviewed at the end of the experiment (Table 1.a). However, with this and other results it should be noted that with the numbers of participants only large differences would be statistically detectable.

Similarly, there was no statistically significant relationship between the mode the user finished on and the mode they said they preferred (see Table 1.b). This is interesting as it would have been reasonable to find that subjects having made a preference choice would only use that mode. In fact even when users expressed a preference in words and use they continued to use both modes.

A total of 5 subjects used only one of the modes, with 3 only using the location mode and 2 only using the camera mode. Most subjects switched mode only once during the course of the experiment but 22% swapped more than once – the maximum number of swaps between modes was 4 by a female subject aged between 18 and 30, and a male subject aged between 30 and 40. This is important as one potential problem would have been if users simply became comfortable with the mode they first used and therefore did not make any real choices. We did find a significant statistical relationship between the mode that the user spent the most time in and the mode they said they preferred (Table 1c).

Camera Mode. In general subjects were enthusiastic about the use of the camera mode with 37% stating that this was their preferred mode of operation. Subjects did not appear to have any difficulty in using the system to take photographs even though the system was not familiar to them. However, we did note that the unusually bright sunlight present during the field trial caused some subjects to have difficulty seeing the screen on the PDA. This was a particular difficulty when taking photographs since the screen acted as the viewfinder for the camera and hence users had to see the screen to determine what they were photographing – furthermore, the necessity to point the unit at the object to be photographed meant that subjects could not freely adjust the viewing angle as they would when using the PDA normally.

The HP camera supports manual focusing but we set this to infinity and users did not need to adjust this. While the camera could be rotated to enable, for example, the PDA to be held horizontally we did not explicitly inform users of this fact and hence in general they held the unit vertically in an outstretched arm as shown in Figure 2.

We recorded all photos taken by users (examples shown in Figure 8) and noticed a strong degree of similarity between the photos

taken by different users. We started the experiments in two distinct locations – outside the Tourist Information Centre and adjacent to the Shire Hall entrance to the Castle. In the former case users almost always took an initial photograph of the main entrance to the Castle (the most striking object likely to be of interest to a tourist in the area) while in the latter case there was more variety in the ordering but many attractions such as the Priory appeared in the sets of photographs taken by different subjects. This high degree of similarity between the photographs taken by tourists gives us confidence that using image recognition techniques to identify objects automatically is a tractable problem. Again, we observed no significant differences in behavior on the basis of age or gender.

Figure 8. Examples of photographs of the same object taken by different subjects

Effect of Errors in Image Recognition. We wished to investigate whether introducing simulated errors into the recognition process would affect subject's reaction to the camera mode. For a total of 13 subjects we simulated errors in the recognition process and returned image sets in which the image of the target object appeared 3rd in the list of returned images. In such cases users had to click through a series of incorrect responses to find their target object in the image set (recall that we only display a single candidate image on each screen and provide the user with *Previous* and *Next* buttons for navigation). We were unable to detect any significant statistical relationship between error rates and final preferences as expressed by the user (Table 1.d). Whilst the power of Chi squared test is low for this number of subjects, the balance in usage and preference noted previously makes it surprising that differences in usability due to higher error rates does not lead to substantial and thus measurable differences. This statistical result is backed up by our own observations of user behaviour – even when errors were introduced this did not seem to affect a subject's perception of the system.

Location Mode. In common with subjects that used the camera mode there was a generally positive reaction to the location mode – 37% of subjects preferred this mode of operation (exactly the same percentage as favored the camera mode). Subjects requested information on objects close by in 71% of the total number of requests. This is in keeping with experiences reported for the original GUIDE system in which subjects tended to use this feature to obtain information on objects close by.

During our observations of subjects using the location mode it became clear that they were not using the system in the way that we had expected. In particular, our expectation had been that subjects would use the system to find out information about a specific object that they could see – providing them with an answer to the question "tell me more about something I can see?". However, our observation is that subjects did not use the mode in this way but instead used it to provide a "comprehensive" list of objects of interest in their location (either near-by or far-away). We are basing this conclusion on responses to interviews, observations and log data. Firstly, we note that during the interviews subjects frequently made statements to the effect that the location mode was good because it told you about things that were near-by. One subject even claimed that it would be good for "lazy people" since they wouldn't have to issue a specific query to the system. When we observed users interacting with the system it was clear they did not identify an object in the real world and then try and find out about it using the location mode but rather used the location mode to find a list of possibly interesting objects that they then endeavored to locate in the real world. Finally, our log data reveals that in 69% of cases the image sets returned as a result of a query in location mode are viewed in their entirety whereas this is only true for 35% of queries issued in camera mode. This suggests that in location mode the users are not searching for a specific object but rather are browsing through the list of returned images looking for objects of interest.

Other Feedback. General feedback from participants was extremely positive - several people enquired where they could buy the system or whether it was available for hire from the Tourist Information Centre. One older subject – a local woman who observed us conducting experiments and wanted to try the system - commented that she would feel vulnerable carrying this equipment in a busy city but once it was explained that such a system could be built into her mobile phone her fears seemed to be allayed (note that we did not use this subject in our data sets since she was not a tourist and was herself working on a project using GPS traces as a form of art).

Finally, we observed the same reluctance to use headphones reported in [3]. During our experiments none of the subjects wanted to use headphones despite the fact that the audio was extremely quiet and subjects often had to hold the PDA to their ear in order to hear the audio properly.

4.3 Discussion

The results we have presented clearly need to be treated with some caution – while the sample size (27 subjects) is certainly comparable with many previous trials of context-aware tour guides, the size is rather small to base any significant conclusions on. The novelty of using new technology will also clearly have played a part in the user response to the system as will the desire to be polite when responding to questions. Furthermore, we cannot claim that the subjects actually used the system as part of their normal tourist activities: the duration of the trials was determined by the subjects themselves but it is clear that the subjects were not really using the tour guide units to explore the area. Rather they were experimenting with the technology and providing an initial impression of the system. Finally, we observe

that the trials were conducted during an unusually warm and sunny spell which may account for the good nature of many visitors.

Despite the lack of experience of long term use of the system it is interesting to note that the vast majority of tourists quickly formed a strong opinion as to which mode they preferred. Indeed, only 26% were either unsure of which mode they preferred or used only a single mode. Of those that did express a preference the subjects were divided equally between the two modes we offered. User's total usage times varied considerably and we did not observe any significant relationship between total usage time and mode preference.

One of our original aims had been to provide two functionally equivalent methods of identifying an object (i.e. using a camera and engaging in a dialogue with the system) in order that we could determine which of these methods users preferred. However, our observations and data have led us to believe that the functionality of these modes was not perceived by subjects as being the same. Specifically, subjects saw the camera mode as performing object identification while the location mode simply provided a means of browsing information on the local area. Thus the modes were used in different ways and the question of preference in reality appeared to include a component that tested whether subjects preferred to explore an area on their own or to be presented with a list of available information – with half of the subjects expressing a preference for each.

The fact that tourists can be divided into two distinct classes that expressed these traits should not have come as a surprise to us – in earlier GUIDE systems it is reported that while some tourists wished to navigate the city using a map a significant number preferred to be given detailed route guidance instead.

More interestingly, we note that those users who expressed a preference for the camera mode did so despite the fact that the camera mode was clearly more difficult to use and more time consuming than location mode. Moreover, the camera mode often returned a larger resulting image set and this set contained images that were similar in appearance to the target image rather than necessarily relating to the same location. Thus, with the exception of the image of the target, the image set was less generally useful to subjects. Indeed, even when we introduced errors into the perceived recognition process – returning an image of the target object much later in the image set – and added simulated processing delays of up to 30 seconds this did not seem to affect the distribution of users opting for each of the different modes.

We believe that we can draw a number of important lessons from the work described in this paper:

1. A significant number of tourists (37%) embraced the use of digital images for object identification despite the fact that it required extra effort and had significant disadvantages compared to a dialogue-based system.
2. Tourists enquire about many of the same objects in a given area and take similar photographs from which they expect the system to be able to identify the object.
3. Testing subject preference for one method of identifying objects over another is difficult in isolation – the absence of other tour guide features such as tours and information overviews impacted the way in which users interacted with the system. For example, in the original GUIDE system an alternative mechanism was used to provide tourists with an overview of a location and its attractions without necessitating the user to engage in the dialogue described in this paper [4].
4. Tourists appear to be classifiable by their underlying approach to information discovery: those who wish to discover new information about objects they find in the physical world and those that wish to browse information and then find corresponding objects in the physical world.

We also note that our initial concerns that subjects would be unable to use the technology or take good quality photographs using a PDA and attached camera appear to have been unfounded.

5. RELATED WORK

Research in mobile tour guides has continued at a rapid pace since the early work of Georgia Tech on the CyberGuide project [9]. The most comprehensive system developed to date is the GUIDE system at Lancaster [4] and our work builds on many of the ideas from GUIDE. Other tour-guide systems have targeted specific domains such as museums [7]. A summary of work in the area of mobile guides can be found in [2].

Work on understanding user interfaces for tour guide systems has often been an integral part of projects in this area but has received particular attention from Aoki and Woodruff who suggested adopting a task-oriented design approach for dealing with object selection [1]. The three sub-tasks identified by Aoki and Woodruff are determining the user's location (location), identifying which objects users are expressing a tentative interest in (intimation) and identifying which objects users have selected (selection). For the location mode there is a clear mapping to these tasks (this is perhaps unsurprising since the GUIDE system is cited in [1]): location is performed by the "Wizard of Oz" system pretending to provide GPS functionality; intimation is through pressing the *Nearby* or *Far away* buttons and selection is through clicking on an appropriate image from the returned image set. There is no direct mapping to the camera mode in the current system, though there is of course an implicit assumption that the location is Lancaster. In future implementations a location system could be used to provide fine-grained location information that would assist in providing context for the image recognition system. However, we would not expect to have clearly identifiable intimation and selection sub-tasks. We do not consider this to be a significant problem with our system despite some of the issues raised in [1].

Davis at Berkeley is using camera phones, contextual information such as location and image similarity algorithms to assist users in automatically creating meta-data for their digital photographs [10,11]. Clearly there are similarities in the underlying system requirements – Davis needs to identify objects in order to know which meta-data to return to the user while we need to identify the object in order to provide tourist information. We anticipate that many mobile applications will have these requirements and that image processing for similarity or recognition purposes will become common network services that can be used by a wide range of mobile applications.

6. CONCLUSIONS

This paper reported on a field trial to investigate user response to the use of digital image capture and recognition for object identification in a mobile tour guide system. Our results show a

good level of acceptance of such technology with half of those subjects that expressed a valid preference choosing this form of interaction and no subjects reporting any problems with using the technology. Our results also show that the problem of image recognition for mobile tour guides is likely to be tractable – images of the same object have a high degree of similarity when taken by different tourists and, crucially, tourists appear to be forgiving of errors and delays when using image recognition technology. However, our trials also highlighted that a significant number of tourists preferred to browse their location for objects of interest and are likely to be unsatisfied with a system that provides only object identification based on image recognition. This has led us to speculate that it may be possible to classify tourists based on their preferred approach to information discovery while exploring a new city. As an item of future work we wish to explore this in more detail, for example, is one class driven by exploration in the physical world and the other by exploration in the virtual?

In conclusion, our results show that developers should not be concerned about user acceptance of digital image recognition techniques for object identification. However, developers should ensure that future mobile tour guide systems support both exploratory and browsing modes of information discovery. In essence, systems need to support answers to two distinct questions: "What's that?" and "What's here?".

7. ACKNOWLEDGEMENTS

This work was partially funded by the European Union under project Simplicity (IST-2004-507558). Andre Hesse was a visiting researcher at Lancaster University and we are grateful to Prof. Jochen Seitz from TU Ilmenau for his continuing support of GUIDE related activities. Chu Wang and Elinor Ollen Pink provided much needed assistance with the field trials.

8. REFERENCES

[1] Aoki, P.M. and Woodruff, A.: Improving Electronic Guidebook Interfaces Using a Task-Oriented Design Approach. *Proc 3rd ACM Conf. DIS*, 2000, pp. 319-325.

[2] Baus, J., Cheverst, K. and Kray, C.: A Survey of Mobile Guides. A Survey of Map-based Mobile Guides in: *Map-based mobile services – Theories, Methods and Implementations*. Chapter 13. Springer-Verlag., Zipf, A., Meng, L. and Reichenbacher, T. (Eds).2005.

[3] Bornträger, C., Cheverst, K., Davies, N., Dix, A., Friday, A. and Seitz, J.: Experiments with multi-modal interfaces in a context-aware city guide. *Proc. Fifth Int. Symp. on Human Computer Interaction with Mobile Devices and Services*, Udine (Italy), September 2003.

[4] Cheverst, K., Davies, N., Mitchell, K., Friday, A. and Efstratiou, C.: Developing a Context-aware Electronic Tourist Guide: Some Issues and Experiences. *Proc of CHI'00*, Netherlands. 17-24 March 2000.

[5] Davies, N., Cheverst, K., Mitchell, K. and Efrat, A.: Using and Determining Location in a Context-Sensitive Tour Guide: The GUIDE Experience. *Special issue of IEEE Computer on Location-based Computing*. August 2001.

[6] Dix, A.: Beyond intention - pushing boundaries with incidental interaction. *Proc. Building Bridges: Interdisciplinary Context-Sensitive Computing*, Glasgow University, 9 Sept 2002

[7] Fleck, M., Frid, M., Kindberg, T., O'Brien-Strain, E., Rajani, R. and Spasojevic, M.: From Informing to Remembering: Ubiquitous Systems in Interactive Museums. *IEEE Pervasive Computing*, Vol. 1, No. 2. 2002.

[8] Kjeldskov J., Skov M. B., Als B. S. and Høegh R. T.: Is it Worth the Hassle? Exploring the Added Value of Evaluating the Usability of Context-Aware Mobile Systems in the Field . *Proc. 6th International Mobile HCI 2004 conference*, Glasgow, Scotland. Lecture Notes in Computer Science, Berlin, Springer-Verlag, pp. 61-73

[9] Long, S., Kooper, R., Abowd, G.D. and Atkeson, C.G.: Rapid Prototyping of Mobile Context-Aware Applications: The CyberGuide Case Study. *Proc. of 2nd ACM International Conference on Mobile Computing (Mobicom)*. 1996.

[10] Sarvas, R., Herrarte, E., Wilhelm, A. and Davis, M.: Metadata Creation System for Mobile Images. *Proc of Second International Conference on Mobile Systems, Applications, and Services (MobiSys2004)*, Boston, Massachusetts. ACM Press, 2004.

[11] Wilhelm, A., Takhteyev, Y., Sarvas, R., Van House, N. and Davis. M.: Photo Annotation on a Camera Phone. Extended *Abstracts of CHI 2004*. Vienna, Austria. ACM Press, 1403-1406, 2004.