

The Panlingual Mobile Camera: Tasks and Interface

Jonathan Pool, Nebiyeleoul Tadesse, Timothy Wong, Luke Woods

http://www.cs.washington.edu/education/courses/cse490f/06au/project_files/camera/need/

Problem and Solution Overview

Mobile individuals encounter signs, posters, labels, menus, captions, instructions, forms, memos, bills, and other texts they cannot read, because of (1) language barriers, (2) illiteracy, or (3) vision disability. Their ability to navigate, learn, communicate, and perform transactions is impaired. We proposed a camera phone with an enhanced functionality and interface to support such a customer by allowing the customer to efficiently photograph texts and get rapid explanations of them, produced locally and/or remotely. The device could also translate memos by the user for others and deliver identifications or explanations of non-text images. In the current stage, we focus on designing an efficient input interface for users seeking translations.

Contextual Inquiry Study

Design

To understand customer requirements, we studied three subjects moving amidst texts that they could not read. We approximated a translation need with different treatments, to avoid systematic substitution bias: (1) a subject wearing blurring goggles, (2) a subject amidst many non-English commercial signs and labels, and (3) a subject with low English fluency interviewed in the subject's native language, and (4) a subject with moderate English fluency interviewed in English. We varied the venues to include outdoor and indoor places, passive (e.g. sightseeing) and active (e.g., shopping) roles, and pedestrian and vehicular movement. To maintain motivation we filled the subjects' requests as they arose.

We sought to discover mainly (1) subjects' purposes, (2) kinds of text subjects wanted to understand, (3) the urgency of the understanding, (4) alternative sources of understanding, (5) means of coping without understanding, and (6) subjects' text-selection strategies.

An investigator, with or without an assistant, accompanied the subject, used questions and suggestions to elicit behavior and comments, and made written and photographic notes. We took advantage of natural in-situ props to elicit more behavior and less speculation.

Interviews

The investigators conducted a total of four interviews. The interviewees included students, professionals, and a homemaker. All had some experience with international travel.

Discoveries

Typically, subjects sought text understanding every 1-2 minutes, would be dissatisfied with a latency of more than about 30 seconds, asked about varied (stationary/moving, indoor/outdoor, large/small) texts, and sought general before specific understandings. They used non-textual cues (e.g., landmarks) to help infer text meanings or avoid the need to do so. Some facts subjects wanted to understand were not explained by any ambient text. Unanticipated discoveries:

- Subjects used font sizes, layouts, proximities to physical objects, and other format cues to decide which pieces of text to ask about. This suggests that effective text interfaces should offer physical cues of this kind.
- With large bodies of text, subjects asked not "Translate this", but "Where is the price?", "Which menu items contain no pork?", "What are the food's ingredients?", "Are any books on this shelf in Hattanese?", "Which movies are showing this week?", or "Is Jane Doe in this building directory?". This suggests that summarization or question answering would often help customers more than translation would.
- Seeing with many category labels (section signs in a bookstore), a subject preferred to get a translation of a category of interest into the local language and then either find (or ask for help in finding) an identical label.
- Seeing a segmented list with headers (a restaurant menu), a subject didn't always ask to understand the headers first, but rather a sample of items. This allowed him to see a sample of prices.
- Subjects encountered an inconsistent mix of text and graphic symbols, such as "WALK"/"DONT WALK" or glyphs. This suggests consistent representations could magnify the value of any acquired understandings, and customers might want explanations of non-obvious glyphs.
- Some texts might be difficult to interpret without local knowledge (e.g., "True Value" being a hardware-store chain). This suggests special criteria for translator competence.
- Subjects were ambivalent about some labels (e.g. on exotic foods) and proper names, since some were very informative, and yet many were untranslatable or, if translated, meaningless to the subjects.
- Some texts were abnormally difficult for subjects and likely abnormally complex for OCR, including logo typography (e.g., an initial "C" distorted to wrap around some subsequent letters), hand-written signs, and glass-mounted transparent signs seen from the back.
- Some texts in non-Latin scripts were accompanied by Latin transliterations, and

subjects sometimes found this helpful in guessing meanings. This suggests that there would be a demand for transliteration-cognizant OCR and language recognition.

- One subject had more oral than written fluency in the idiographically written language of some texts and indicated that hearing the texts read aloud might suffice. This suggests foreign-language customers sometimes have needs like illiterate customers.
- A subject seeing a map suggested that getting the map back with a "You are here" mark would be useful.
- A store ordered us to stop photographing in it. This suggests limits on our device's usability or marketing its benefits to premise proprietors as well as end-users (cf. service dogs).
- Subjects sometimes made decisions (e.g., giving money to a beggar holding a sign, or guessing which route to choose) without waiting or even asking for explanations of highly relevant text. Assuming that these texts were not superfluous, we surmise that the cost of merely asking and waiting for an explanation is subjectively significant.

Experience Sampling Study

Design

While the contextual-inquiry study revealed when and why users might want translations (or other interpretations) of texts, we also wanted to learn how, and how well, users could photograph various kinds of text, in case they wanted translations. For this purpose we designed an experience-sampling experiment.

We gave each of two subjects a camera phone for 48 hours. From 8:00 a.m. to 9:30 p.m., the phones alerted the subjects randomly, at a mean rate of once per hour, and invited them to make one or more photographs of text. Subjects were encouraged to photograph diverse texts so we could evaluate the amenability of such photographs for text translation. Subjects were invited to photograph additional texts at their discretion and to record any audio-visual comments they thought might be useful.

The sensing equipment available to us did not permit the collection of data on the locations where photographs were taken or where subjects were when they received alerts. We did not ask subjects to make records of their locations, because this would have added enough to the burden on subjects to put their willingness to participate at risk.

Experiments

The subjects were a professional librarian and a student. They spent their study periods in various indoor, outdoor, and travel contexts. One made 38 and the other made 49 photographs.

Discoveries

Our main discovery of interest was that a substantial fraction of all photographs taken by subjects were too blurred, dark, small-scale, mis-aimed, or low-contrast to permit all the text in them to be read by a human or to make any text visible. In our judgment, this was true of 42% of the photographs in experiment 1 and 92% of those in experiment 2. This suggests that help to users in assuring and verifying text clarity (e.g., zoom lenses, sharpness feedback, automatic sharpening by burst capture) could be beneficial. Most photographs containing illegible text also contained legible text. This suggests that an opportunity to mark text of interest (e.g., when users want newspaper headlines but not stories) could help prevent expensive attempts to translate marginally legible text that customers didn't need anyway. Photographed texts appeared in various orientations and included handwriting, suggesting additional complexity for OCR.

Post-experiment annotations from one subject suggested that text illegibility was especially likely with photographs of three kinds: (1) those made out the windows of moving vehicles, (2) those made at a range of 12" or less, and (3) those of texts on video monitors.

Another discovery was that not all photographed texts were unilingual. This suggests that a system might be inadequate if it presupposes that each text is in a single language.

Further study would be more informative if we asked subjects to mark the texts that they intended to photograph immediately afterward.

Task Analysis

- **Who is going to use the system?** The initial target customers are people not knowing the local language(s). People with limited literacy or vision and people in unfamiliar places are other potential customers.
- **What tasks do they now perform?** Customers read and try to understand texts in signs, labels, menus, periodicals, books, etc. Their goals encompass all activities of visiting or living in a place, including personal care, learning, entertainment, transportation, and commerce.
- **What tasks are desired?** Simulated customers showed an interest in delegating some reading to agents and querying the agents for answers to questions, including which text passages warrant translation.
- **How are the tasks learned?** Customers learn text understanding with literacy education and practice. Where they don't know a text's language, practice has helped them infer text relevance and classification (e.g., whether a text is information, a warning, an advertisement, etc.).
- **Where are the tasks performed?** Interesting texts are almost ubiquitous. Moreover, people appear to be encountering text (as opposed to oral communication) more pervasively with time, because of the growth of email, text messaging, and lecture presentations.
- **What's the relationship between user and data?** Customers mainly consume data

(texts). But each translation for a customer could become a sharable datum. Moreover, customers could write notes, or enter responses in text-based interfaces, and have them translated into local languages.

- **What other tools does the user have?** Local persons may translate texts into a customer's language or use gestures or graphics to explain texts. Graphic symbols and physical objects accompanying texts may permit educated guesses of meanings.
- **How do users communicate with each other?** Typically, customers are alone or are in groups that communicate orally *in situ*. A customer might also communicate with remote colleagues.
- **How often are the tasks performed?** Customers in our studies sought to perform tasks about once every 1-2 minutes on average.
- **What are the time constraints on the tasks?** Customers typically appear to suffer significant impairment of benefits if latency exceeds about 30 seconds.
- **What happens when things go wrong?** Customers who know they don't understand a text usually seem to (conservatively) avoid acting on it, but unnoticed misunderstandings of critical texts could cause serious harm.

Tasks Supported

We propose a service that would deliver text translation. A sample of existing and new customer tasks that the service would support, at three levels of anticipated difficulty, is:

1. The customer sees a sign on a road and tries to infer whether it states a speed limit. Easy, because numerals and international symbols are likely to give clues, and traffic signage is likely to be consistent, so once the customer has identified one speed-limit sign it will be straightforward to identify others.
2. The customer is traveling by bus as a sightseer, not caring where any bus is going. The customer wants to know which buses are for the general public's use, and how to pay the fare. Easy, because buses other than public ones will likely not stop for the customer, and, although the fare-paying instructions on the buses may be difficult to read, the situation makes it obvious what the customer needs to do, and the driver, any fare collector, and other passengers can help if the customer shows some money and looks inquisitive. In addition, other passengers may serve as examples for which bus to take and how to pay.
3. The customer is looking for the philosophy section in a bookstore, whose sections are labeled with overhead signs. Moderate, because it would not be obvious to others what the customer wanted, but, even if the section labels are not understood, some books may be in an understood language, and the store's staff or customers may know a language that the customer knows. The staff has pecuniary as well as humanitarian motives to help.
4. The customer is visiting an ancestral homeland and has found a cemetery where a great-grandparent's tomb is believed to be. By comparing the ancestor's written name with inscriptions, the customer has found what appears to be the ancestor's tomb. Now

the customer wants to know what the inscription says about the person buried there. Moderate, because inscriptions tend to wear with age, but the customer's curiosity is likely to be obvious to others in such a location, who are likely to be sympathetic and helpful if they can find a common language with the customer. If no solution on site is found, the customer can make a rubbing or a photograph and take it to others for analysis.

5. The customer wants to dine and is looking for a restaurant that offers at least some food that is not extremely spiced. Difficult, because menus are likely to offer few or no clues to the optionality of spicing, it is probably not obvious to others that this might be the question on a customer's mind, and universal gestures to express this question do not exist.
6. The customer, on a street devoid of other people, approaches an apartment building and wants to know whether a person whose name is known by ear is listed in the building's directory, but the names are written in a script that the customer doesn't know. Difficult, because the customer can't convert from sound to script and nobody is there to help.

Interface Design

Functionality

The envisioned system includes a client and a server. The client obtains requests for translation, delivers results, and manages prior requests and results. The server produces translations (using some combination of OCR, language identification, and translation, by local and/or remote human and/or artificial agents).

We imagine that related functionalities aimed at text understanding could cooperate with these. Two examples are support for knowledge sharing among customers and support for customer research. We withhold judgment on whether to treat them as additional functionalities of the same system. We are not considering such additional functionalities at this stage.

The current design stage addresses one of the client functionalities: obtaining from the customer requests for translation. We make the simplifying assumption that the customer desires a translation of a single text that can be captured in a single photograph. The functionality supports the customer's submission of the last-made photograph. If the customer wishes to review, compare, or submit an earlier photograph, this activity requires support from the management functionality, which the current design stage does not address.

Description

Our strategy is to design each interface in an idealized version and subsequently to modify it as required for the constraints of particular devices. In this stage we are

designing the idealized version of the interface for the solicitation of translation requests. It makes use of some features not widely available yet on camera phones, particularly touch-sensitive displays with softbuttons, and thus as many buttons in any state as the facts of the state warrant.

The interface has five states:

- standby
- camera
- inspection
- modification
- request

The standby state is the normal state of a camera phone. In the camera state, the display displays the image that can currently be captured, gives the user control over the capture settings, and permits the user to capture the current image. In the inspection state, the user can inspect the captured image. In the modification state, the user can modify the captured image and add metadata to it. In the request state, the user can request a translation of text in the captured image.

The transition possibilities between pairs of the states are shown in Figure 1, where the normal (non-aborted) sequence, beginning and ending with the standby state, is shown with teal arrows.

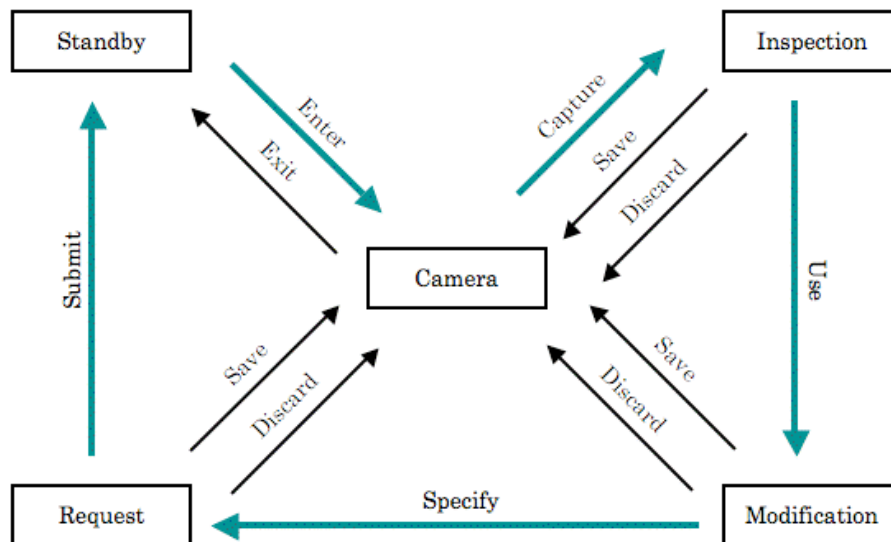


Figure 1. Translation Solicitation Interface State Transitions

The user powers provided by the interface are intended to support commonly wanted functionality elements and avert damage from errors.

The main interface elements are these:

- Each command that produces an inter-state transition is executed with single (hard or soft) button.
- Controls in the camera state include zoom, brightness, contrast, color balance, and, of course, capture. All but capture are treated as continuous one-dimensional variables, which have separate buttons that make them current. They share a bipolar control that

increases or decreases the current variable while pressed. A variable flash depends on the brightness setting and the ambient light, so is not directly controlled by the user.

- In the inspection state, the user zooms in by touching the image where the enlarged image is to be centered and zooms out by pressing and holding a button (the view, not the image, is changed).
- In the modification state, the user can rotate, crop, and voice-annotate the image and marquee and delimit the text. While in this state, the user can also zoom the view. Voice annotation can include point references; the user makes them by touching the display while speaking. Rotation works like the camera-state controls. To crop the image, the user touches the upper-left and lower-right corners of the new image. Modifications are combined into a single unlimited undo stack.
- In the request state, the user chooses between two speeds and two qualities, and requests a written and/or aural translation. The submission of any translation request automatically implies the saving of the image.

Task Scenarios

Scenario 1. The customer sees a bus and takes a photograph of its identifying sign, adjusting the brightness to make the text clear. The customer then inspects the photo and enlarges it with the sign at the center, but then decides to reverse the enlargement. The customer decides to request a translation of the sign and identifies the text with a rectangular frame. Then the customer submits the request, specifying that the translation should be written rather than spoken, should be fast, and does not need to be high-quality.

This customer proceeds through the five states from standby and back to standby. Figure 2 shows a possible interface at one point of each state in this scenario.

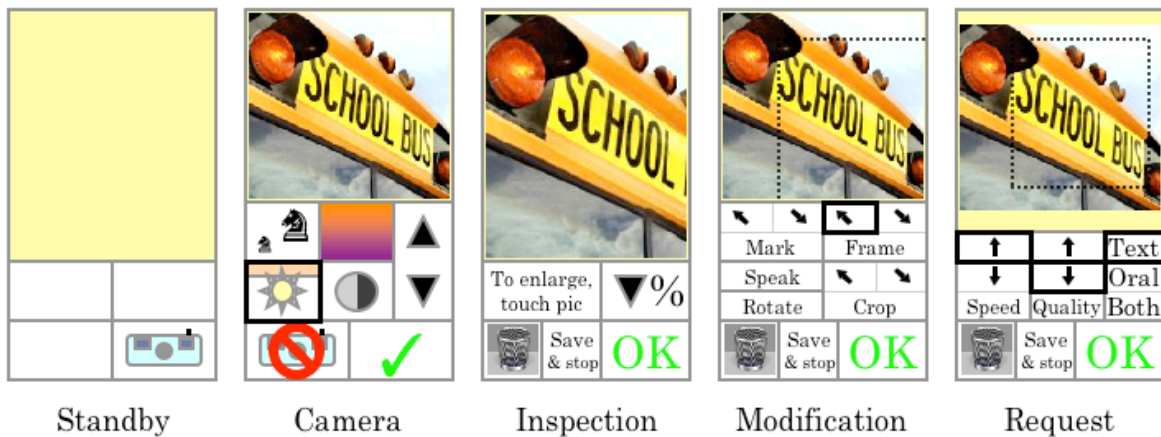


Figure 2. Sample Views of States in Scenario 1

Scenario 2. The customer sees a tombstone inscription of interest and decides to photograph it for translation, but the inscription has worn with age and shows little color contrast. So, the customer experiments with the controls offered by the interface for image manipulation in the camera state: brightness, contrast, color balance, and

magnification. As illustrated in Figure 2 in the case of the brightness control, the interface gives feedback to the user about the current value of the variable that the user is currently manipulating. If a photo is obviously unusable when viewed in the inspection state, the customer can discard it and return automatically to the camera state. If the photo seems questionable, the customer can save it instead and return automatically to the camera state. If the questionable photo turns out to be the best one, the customer can retrieve and submit it (retrieval is a functionality that is excluded from the current design phase).

Scenario 3. The customer has found an attractive restaurant in a small Togolese town, but has no hope of finding somebody there who knows Mongolian. So, the customer photographs the menu posted on the wall. By using the "Speak" button in the modification state's interface, the customer records a voice annotation to explain that a translation of the entire menu is not wanted at this time; instead, the customer wants to know whether the menu indicates that it is possible to request moderation in the usually extreme spicing of the food. While speaking, the customer may touch the image to point to any items being mentioned, such as a picture of a dish that looks interesting.

Addendum



Figure A1. Barely Legible Photo from Experience Sampling Study



Figure A2. Human-Readable but Non-Machine-Readable Photo from Experience Sampling Study