PACER: Fine-grained Interactive Paper via Camera-touch Hybrid Gestures on a Cell Phone

Chunyuan Liao, Qiong Liu, Bee Liew, Lynn Wilcox
FX Palo Alto Laboratory
3400 Hillview Ave, Bldg 4, Palo Alto, CA 94304, U.S.A.
{liao, liu, bee, wilcox}@fxpal.com

ABSTRACT
PACER is a gesture-based interactive paper system that supports fine-grained paper document content manipulation through the touch screen of a cameraphone. Using the phone's camera, PACER links a paper document to its digital version based on visual features. It adopts camera-based phone motion detection for embodied gestures (e.g. marquees, underlines and lassos), with which users can flexibly select and interact with document details (e.g. individual words, symbols and pixels). The touch input is incorporated to facilitate target selection at fine granularity, and to address some limitations of the embodied interaction, such as hand jitter and low input sampling rate. This hybrid interaction is coupled with other techniques such as semi-real time document tracking and loose physical-digital document registration, offering a gesture-based command system. We demonstrate the use of PACER in various scenarios including work-related reading, maps and music score playing. A preliminary user study on the design has produced encouraging user feedback, and suggested future research for better understanding of embodied vs. touch interaction and one vs. two handed interaction.

Author Keywords
Cell phone, camera, touch, gesture, paper interface, fine-grained, embodied interface.

ACM Classification Keywords
H.5.2 User Interfaces (D.2.2, H.1.2, I.3.6): Interaction Styles

General Terms
Design, Human Factors

INTRODUCTION
Paper still plays an important role in many tasks even in this age of computers [25]. This phenomenon can be attributed to paper's advantages in display quality, spatial arrangement flexibility, instant accessibility and robustness. However, paper lacks computational capability and does not render dynamic information. In contrast, cell phones are becoming powerful in computation and communication, providing a convenient access to dynamic information and digital services [5, 8]. Nevertheless, cell phones are constrained by their limited screen size, relatively lower display quality and cumbersome input methods [2, 27]. Combining the merits of paper and cell phones for rich GUI-like interactions on paper has become an active research area [5, 8, 22].

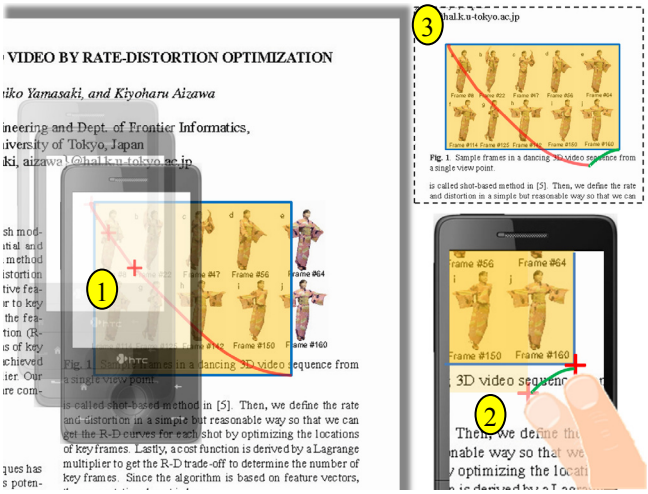

Figure 1. A hybrid marquee gesture for copying a rectangular region from a paper document. (1) Initial coarse selection (in orange) with a camera-detected embodied gesture (in red) through a Magic-Lens-like interface. (2) Fine-tuning with a touch gesture (in green) on the screen. (3) The whole hybrid gesture and the resulting region.

A typical approach is to use a cell phone to identify a specific paper document segment through its visual features [5], a barcode [22] or an electronic tag [21], and retrieve the associated digital information (e.g. a video clip), with which the user then interacts on the phone. Existing practices, although having made significant progress, still fall short of the goal of GUI experience on paper. First, the prior work does not offer sufficient fine-grained operations on user-specified document details, such as individual words, math symbols or an arbitrary image region. HotPaper [5] focuses on text-patch level multimedia annotation on paper, and Rohs [22] augments pre-defined regions in a printed map with dynamic weather information.

CHI 2010, April 10–15, 2010, Atlanta, Georgia, USA.

Second, current systems revolve around point-and-click interaction [8, 22], with a lack of gestures such as the marquee, bracket and lasso selectors, which offer more flexibility to manipulate document content and have been widely deployed in GUIs. Finally, document types handled by these systems are somewhat limited. [22] requires 2D bar codes on paper, and HotPaper [5] needs several sentences of western text for identification.

The recent advances in content-based document recognition algorithms [16, 28], cell phone motion gestures [13, 29], touch interfaces [11, 27] and hardware production pave the way to new approaches. We present PACER (Paper And Cell phone for document Editing and Reading), an interactive paper system featuring a camera-touch hybrid interface. It recognizes documents based on natural document visual features instead of any special markers on paper or specific end user hardware. More importantly, it allows users to manipulate *fine-grained* document content with various *gestures* beyond point-and-click.

For instance, to copy and email an interesting region from a paper document, a user first aims a camera phone roughly at the region and captures a picture. PACER recognizes the picture, and presents on the screen the corresponding high quality digital version of the document, rather than the raw video frames (We call this design *loose registration*). The user then operates the phone as an embodied see-through "Magic Lens" [24], with its crosshair center treated like a mouse pointer. The user sweeps the phone over the paper to issue an embodied marquee gesture for selecting a rectangular region of the document (Figure 1-1). To fine-tune the ends of the gesture, the user can switch from the embodied interaction to the touch interaction by directly touching the screen and moving the pointer in a zoomed and scrollable view (Figure 1-2). Upon user confirmation, the selected region (Figure 1-3) of the high quality digital document is copied and sent via email.

With this camera-touch hybrid interaction, the user can also pick the title of a reference from a journal, and then search for its full text on Google Scholar (Figure 4); specify a word or math symbol for text search on paper (Figure 8-1); snap to a sightseeing drive on a paper map, and browse the street views on the phone while sweeping it along the route (Figure 8-2); or play a music score by moving the phone over the intended sections (Figure 8-3). These documents may be any mix of text (possibly in any language), images and graphics, and need no special markers. To our best knowledge, PACER is the first system with such features in the literature.

In the remainder of this paper, we first describe the related work, followed by the system architecture and the underlying document recognition and tracking techniques. We then discuss the PACER command system, focusing on the interaction techniques that enable the fine-grained gesture interaction, including semi-real time document tracking, loose registration and camera-touch combination. We then demonstrate several PACER application scenarios, and report the preliminary user study on the design with a mockup implementation.

RELATED WORK

Our work falls in the general category of interactive paper, which attempts to bridge the gap between paper and the digital world. The research can be traced back to pioneering systems like Digital Desk [31], which augments paper with digital video projected from overhead. A-Book [19] lays a PDA on a notebook, serving as a digital overlay for the notes beneath the PDA. In contrast, PaperLink [4] treats paper and a screen as two separate displays, and renders on the screen information associated with the paper.

The Anoto [3] digital pen enables pen-paper only interfaces. The digital pen traces handwriting on paper in real time by recognizing special dot patterns printed on the paper. Using this technology, Paper Augmented Digital Documents [7] and PapierCraft [15] enable users to annotate a digital document or issue pen gesture commands to manipulate its content with a printed copy. Recent Anoto-based systems like ButterflyNet [32], PaperProof [30] and Musink [26] share the general idea of pen gesture commands on paper, and focus on mobile data capturing, proofreading and music composing tasks respectively. PACER adopted this idea of gesture commands, applying them to a camera-touch phone based paper interface.

Interaction with Paper Documents Using Cell Phones

There are several recognition methods for linking paper to digital documents. Barcodes are one of the most popular approaches. Rohs [22] uses 2D barcodes to identify specific geographical regions on a paper map, through which users can retrieve the associated dynamic weather forecast with a camera phone. Marked-up Map [21] employs RFIDs to label map regions. These marker-based approaches are relatively reliable, but require alteration of the original documents and cannot support high spatial resolution. Also visual markers are visually obtrusive and take up valuable display space.

To avoid the limitations of the markers, other systems such as PBAR [10], HotPaper [5], Mobile Retriever [17] adopt a content-based recognition approach, identifying document patches through their text features, e.g. the spatial layout of words in the patches. PBAR allows users to open a printed URL on a camera phone through point-and-click. HotPaper facilitates creating and browsing multimedia annotations on paper. These approaches do not require markers, but cannot support patches with little or no text content (e.g. maps, photos and document figures).

Approaches based on general purpose image local features, such as SIFT [18], may recognize generic content including text, pictures and graphic elements. Using SIFT, MapSnapper [8] can locate the corresponding digital map for a picture of a small map region on paper. SIFT is also employed by Kim et al. [12] to track printed documents on

a desktop through an overhead camera. FIT [16] resembles SIFT, but is more efficient in feature calculation and storage. Wagner [28] proposed a SIFT-based algorithm optimized for cell phones, which has been deployed in MapLens [20] for augmented maps. Some commercial products like SnapTell (http://www.snaptell.com) and Google goggles (http://www.google.com/mobile/goggles) have been available for on-line visual search.

Mobile Interface

PACER draws upon mobile interface research to handle target acquisition, document navigation and multi-modal input. In particular, to address the finger occlusion issue for touch screen interaction, we borrowed ideas from Shift [27], which shows a callout for the occluded screen area. Other techniques like Cross-Keys and Precision-Handle [2] interactively amplify screen regions to facilitate small target selection. AppLens and LaunchTile [11] focus on one-handed thumb operation for quickly accessing applications.

With the popular and pervasive camera phones, TinyMotion [29] and REXplorer [13] exploit the built-in camera to detect the phone user's hand movement, which is used for tasks such as selecting menus [29] and issuing gesture commands [13]. Adams et al. [1] proposed an algorithm to align successive phone viewfinder frames in real time.

DOCUMENT RECOGNITION AND TRACKING

The PACER system consists of three bottom-up layers, namely *paper document recognition and tracking*, *command system* and *applications*. We employ image local feature based algorithms [8, 16, 18, 28] for recognizing camera phone-captured images of paper document patches (called *camera image* hereafter). This approach does not require any special paper, markers or dedicated devices, and is applicable to generic document types including text, graphics and pictures. It is also robust to image scaling, rotation and occlusion, which is very desirable for the flexibility of cell phone based interfaces. We used FIT [16] for our current implementation. SIFT [18] is also possible, but less efficient than FIT in feature calculation and search [16], and subject to commercial product licenses. Wagner's algorithm [28] is promising, but the detailed implementation is not available, and its recognition accuracy with larger databases (400+pages) is uncertain.

Figure 2 illustrates the system architecture. The printed document is sent to a PACER server, which identifies feature points in every page and calculates a 40-dimension FIT feature vector for each point [16]. The vectors are clustered into a tree for ANN (Approximate Nearest Neighbor) correspondence search [16]. Other metadata such as text, figures and hot spots in each document page are also extracted and indexed at the server. The same feature calculation is applied to a subsequent query camera image, and the resulting features are matched against those in the tree. The page with the most matches (if above a threshold) is taken as the original digital page for the camera image.

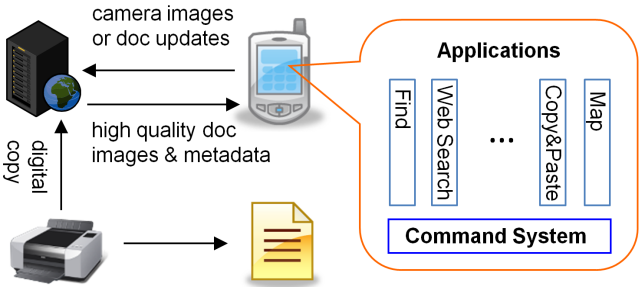

Figure 2. The architecture of the PACER system

To realize the mouse pointer metaphor with the phone crosshair, we derive a homographic transform H between the camera image and the matched page based on their corresponding feature points, in a way similar to MapSnapper [8]. With H, the whole camera image is mapped to a quadrangle patch in the original document image, and the crosshair center is mapped to a point P. The metadata associated with the patch and P can then be retrieved from the server, including text, hotspot definitions and high resolution page images of the document.

Due to the high computational complexity, our current implementation performs the feature calculation, matching and metadata retrieval on a separate PC server. The user interface runs on a cell phone, communicating with the server wirelessly. In the future, more powerful cell phones may take over the tasks and run as a self-contained system.

Semi-Real Time Document Tracking

The embodied interaction, such as drawing the marquee gesture in Figure 1, relies on real time tracking of the paper document relative to the phone. FIT requires about 1 second to accomplish a query session through the cell phone's wireless link. In contrast, camera-based motion detection is much faster (~15fps on an HTC Touch Pro cell phone), but merely detects relative movement and is subject to accumulated errors [1, 29]. We opted to combine the two methods. The relatively accurate recognition is performed upon user request or automatically at fixed intervals of time (e.g. 1~2 seconds). Based on the result, the crosshair position relative to the document is estimated with the motion detection between two consecutive frames. Every recognition session resets the motion detection to reduce the accumulated error.

With this approach, the precision of physical-digital document registration depends on recognition accuracy, motion detection granularity and recognition interval. The motion detection in the current prototype is based on TinyMotion [29] with 16x16 pixel macro blocks in 240x320 video frames, therefore the granularity is roughly 8 pixels. More accurate motion detection could be achieved with new algorithms like [1]. Accelerometer-based detection, although consuming fewer CPU cycles, needs double integral options on the raw acceleration readings to estimate the phone position, so is less reliable than the camera-based detection [29]. Moreover, the shorter the recognition interval, the smaller the accumulated error. When the interval shrinks to 1 frame (i.e. every frame is

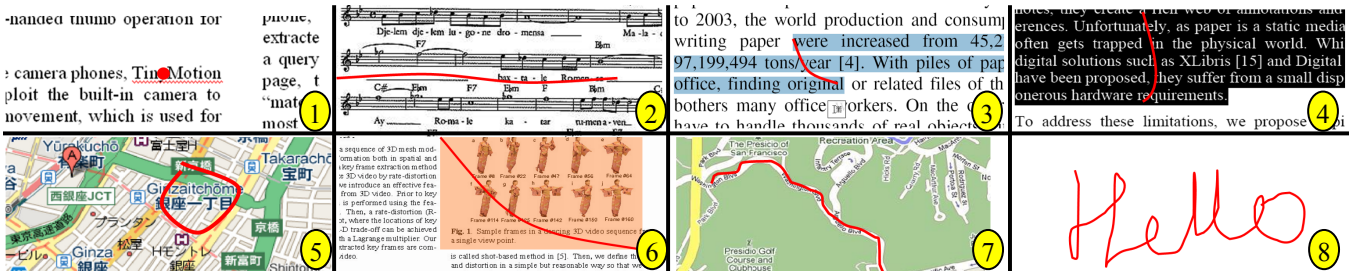

Figure 3. Gestures (in red): 1) Pointer, 2) Underline, 3) Bracket, 4) Vertical Bar, 5) Lasso, 6) Marquee, 7) Path, 8) Freeform

recognized), the approach becomes fully real-time. We expect decreasing recognition intervals in the future due to increased processing speed. Note that even if the physical-digital document registration is imperfect, the fine-grained interaction is still possible with the loose registration strategy that is described later in this paper.

With the semi-real time tracking, the pointing and moving of the phone over a paper document are mapped to those of a pointer in the corresponding digital page. This mapping enables users to interact with document content through an embodied "see-through" interface similar to the Augmented Reality interfaces [5, 20]. Different from the prior work though, the users do not interact with the raw camera images, but the retrieved high quality document images.

CAMERA-TOUCH HYBRID COMMAND SYSTEM

Based on the document recognition and tracking, we propose a command system for users to select arbitrary paper document content (e.g. individual words, characters, math symbols, and image regions) via a set of gestures like an underline, marquee and lasso. The users can also choose an action via a menu system. We integrate the camera and touch input to address challenges such as low camera image quality, inaccurate registration and hand jitter.

PACER Gestures

Motivated by the computer GUIs like MS Windows and Tablet PC interfaces like Scriboli [9], PACER offers much richer interaction beyond point-and-click, facilitating command target selection at a wide range of granularities. Figure 3 illustrates the eight PACER gestures currently implemented: *Pointer* is suitable for the point-and-click interaction with pre-defined objects (e.g. words, East Asian characters, math symbols and icons); *Underline* is used to select a line of text or music notes; *Bracket* and *Vertical bar* ease quoting text in a sentence and multiple lines respectively; *Lasso* and *Marquee* support selecting an arbitrary document region; *Path* can be employed to set a route on a map, and *Freeform* is interpreted in an application-specific way. PACER currently neither supports multi-stroke gestures nor performs gesture recognition. Users need to choose a gesture type manually before issuing a gesture.

Pure Embodied Interaction for Issuing Gestures

As shown by Rohs [23], in map browsing tasks, an embodied Magic Lens interface combined with a paper map is significantly more efficient than a button-based cell phone interface with only digital maps, since the paper provides more contextual information and the direct point-and-read interaction is intuitive. Inspired by this finding, we began with a pure embodied interface for PACER.

As an example, Figure 4 illustrates the basic steps of issuing a bracket gesture to select a paper title for full document search with Google Scholar. The phone is initially in the *snapshot* state while being pointed to the first title word. The user takes a snapshot to retrieve the corresponding high fidelity document image and metadata from the server. When the cell phone receives all the metadata, the UI turns to the *embodied navigation* state, indicated by a blue crosshair (Figure 4-2). The user adjusts the crosshair location by panning the phone. The word closest to the crosshair is highlighted for feedback (Figure 4-3). The user can also click the up and down buttons (Figure 4-1) to zoom in and out, with different control-to-display ratios for easy target selection. Clicking the ENTER button confirms the gesture' starting point. The UI then starts *embodied gesture* state, signified by changing the crosshair color from blue to red (Figure 4-4). The user sweeps the phone toward the last title word, with all the currently selected words being highlighted (Figure 4-5). Again, clicking the ENTER button finalizes the ending point. A menu of available actions pops up (Figure 4-6).

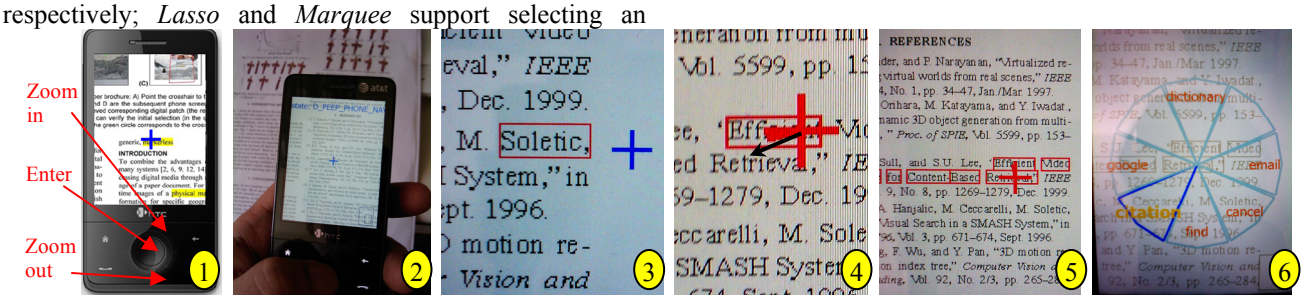

Figure 4. Select a paper title with a bracket gesture. 1) Button layout. 2) Embodied navigation. 3) Snap-to-word. 4) Embodied gesture starts, moving toward the last title word "Retrieval". 5) Feedback for the embodied gesture. 6) Pop-up menu for actions.

The user can either select "citation" to submit the selected paper title words to Google Scholar, or "cancel" to return back to the *embodied navigation* state for a different target. The state transition is denoted by the red paths in Figure 5.

Loose Registration

PACER substitutes the retrieved high fidelity document images for the camera images, and does not require persistent and precise alignment between the background paper document and the screen content (i.e. the physical-digital document registration) except for the initial pointing for recognition. We call this strategy *loose registration*, in contrast to the s*trict registration* in conventional Augmented Reality (AR) systems [5, 10, 20], which demand continuous and accurate physical-digital document registration for users to directly interact with the phone-captured video frames of a paper document. A-Book [19] too allows for interaction with the digital version of the augmented paper notebook. PACER extends the idea to camera-based handheld interfaces. HotPaper [5] shows a thumbnail of the recognized document for an overview, but not for content manipulation.

This design has several advantages. First, loose registration works around the continuous and precise registration of digital overlays with camera images. Therefore, even if the camera-to-digital-document coordinate transform is imperfect (e.g. due to inaccurate feature point matching or accumulated error in the semi-real time document tracking), the overlays such as the highlighting boxes (Figure 4-3,4,5) are perfectly aligned with the retrieved document contents. Although the screen content might not exactly match the phone-occluded portion of the paper document, we believe that it would not be a big problem, since the paper only serves as a coarse context, and the user's eyes focus on the phone screen, where the fine-grained interactions are performed. Second, loose registration avoids the low quality cell phone video, which is caused by, for example, out-of-focus, low resolution and undesirable lighting conditions [5, 20]. The retrieved high fidelity document images, together with the perfectly matched overlays, effectively facilitate fine-grained interaction.

Loose registration also enables more flexible manipulation on the document content. The users can navigate to the document segments that are not covered by camera images (Figure 6-2), and can adjust the zoom factors and control-to-display ratio regardless of the actual phone-paper pose, which is nearly impossible in conventional AR systems [5, 10, 20]. As Figure 4-(3,4) shows, a user can zoom into a small document segment, move the crosshair pixel by pixel, and therefore manipulate the document at its pixel level.

Finally, loose registration can relieve the users of repeatedly coordinating the paper and the cell phone in space, because that is only necessary for the initial snapshot recognition and some tasks that require the paper as a continuous context reference (e.g. map navigation). This distinguishes PACER from other systems. We reflect more on loose registration later in this paper.

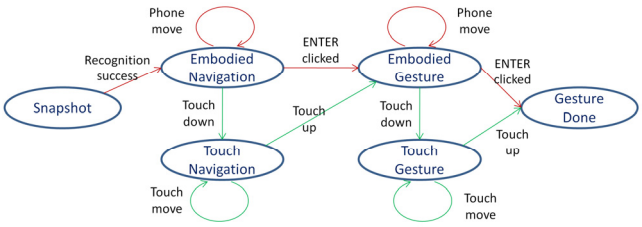

Figure 5. State transition of the PACER interface

Tackling Hand Jitter

For more spatial flexibility, we prefer one-handed phone operations, so that the free hand can hold and arrange the paper document with a high degree of freedom. However, hand jitter makes it difficult to manipulate small targets. This is an issue with all direct freehand pointing techniques including the handheld projector and laser pointer interfaces [6]. Zooming with a large control-to-display factor could relieve the problem to a degree, but, when zoomed to a small document region, the user may lose the contextual global view that is important to the on-going tasks [2], such as navigating a map or copying a large image region.

We address these challenges with a three-stage processing. First, we borrow the idea of Zoom-and-Pick [6], filtering out phone movement below a certain threshold. This may result in imprecise alignment of the screen content and the background paper document, but this is not an issue due to loose registration. Second, we render the beautified ideal gesture shapes that best fit the original input instead of the raw gesture strokes, or simply highlight the currently selected content. This design is motivated by REXplorer [13], which suggests that gesture beautification can improve the user experience with low fidelity motion-based gestures on handheld devices. Lastly, we exploit the application-specific semantic, snapping the crosshair to the pre-defined objects like words (Figure 4-3), icons or map routes (Figure 8-2). Pilot tests confirmed the effectiveness of this design.

Integrating the Embodied and Touch Interaction

The increasingly popular touch screens on cell phones provide new design space for the PACER interface. In the initial system deployment, we noticed that, to select a target within the reach of a user's thumb, direct finger manipulation on the screen is often faster and more intuitive. As a follow-up, we conducted a series of tests, and investigated the pros and cons of the PACER embodied interface (simply *embodied interface* hereafter) and the touch interface.

With paper documents, the embodied interface preserves the context for phone operations. It is quick and natural for browsing across a large area, e.g. searching multimedia annotations on a brochure [5], exploring points of interests on a map [22], or selecting a figure spanning two columns. It has no finger occlusion issue. However, its dynamic peephole mental model [24] may not be familiar to novices.

	Embodied	Touch Screen
Mental model	Dynamic peephole	Static peephole
Context Information	Rich	Poor
Large Area Browsing	Easy	Hard
Paper-Phone Coordination	Hard	No
Input Sampling Rate	Low	High
Finger Occlusion	No	Yes
Transmission Bandwidth	Low	High

Table 1. Comparison of the embodied and touch screen interaction in PACER (Note: all the ratings are relative)

Users need to carefully coordinate the phone and paper (in case of strict registration). The input sampling rate is relatively low (e.g. 10~15 Hz with PACER). This leads to low performance in small target acquisition. As shown in TinyMotion [29], the information transmission bandwidth is just 0.9 bits/sec with a pointing device building on camera-based motion detection.

On the other hand, the touch screen operation is free of hand jitter, demands no phone-paper coordination, and has high input sampling rate (100+Hz). Its static peephole mental model [24] is more familiar to average users. However, the small screen limits navigation across a large area and constrains global context. The finger occlusion is an issue for small target selection [27]. The comparison of the two types of interfaces is shown in Table 1.

Therefore we propose a hybrid interface, which employs the embodied interface for global and coarse document navigation, and touch screen interaction for local and fine-granularity target acquisition. The interaction is very similar to the basic one, except in the *embodied navigation* and *embodied gesture* modes the user can directly move her/his thumb on the screen to refine the crosshair location with high precision. To avoid confusion, the motion tracking is disabled once the screen is touched. Accordingly, we add two states for the screen operations, *touch navigation* and *touch gesture*. Upon finger lifting, the gesture point is confirmed. The state transition is illustrated in Figure 5, in which the green paths denote the new state transitions.

To facilitate the screen interaction, we borrowed ideas from SHIFT [27] , showing a callout of an enlarged view of the thumb-occluded area (Figure 6-1). In contrast to SHIFT we adopt a rectangular callout to fit the word shapes. Once the finger is close to screen borders, automatic scrolling is triggered at a speed reversely proportional to the finger-border distance (Figure 6-2).

Menu Management

PACER allows users to specify, via a menu, an action for the document content selected by a gesture (Figure 6-3). We adopt a hierarchic pie menu system, so users can simply select a menu item by indicating its direction, which requires less visual attention and facilitates smooth skill transfer from novices to experts [14].

Users can browse the menu items by touching and sliding the thumb on the screen. This is preferred if the thumb is over the screen, for instance, after setting a gesture ending point. A menu item is selected once the finger lifts. In pure embodied interaction, the thumb often rests on the ENTER button, so it is inconvenient to change hand pose to reach the menu. Thus in this case, we use phone tilting in the item directions for menu browsing (Figure 6-3). Clicking the ENTER button confirms the menu selection. To avoid confusion, once the user touches the screen, the tilting input is disabled.

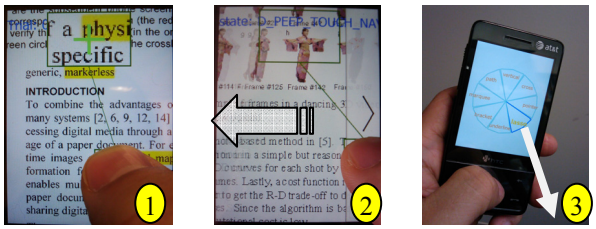

Figure 6. 1) Occlusion callout. 2) Auto-scrolling. 3) Tilting the phone to select a menu item "lasso"

Command Execution

Once the gesture and the action are specified, the command system passes the selected document content to the specific application to perform the action. There are two execution modes. In Sequential Execution mode, the command execution is held until the gesture is finished, which is common for many tasks like copy & email. In Parallel Execution mode, the action is performed while the gesturing is going on. For example, while a user issuing a path gesture along a route on a map, the street view of the route is played simultaneously (Figure 8-2).

APPLICATION SCENARIOS

The PACER command system eases the development of camera-touch phone based interactive paper applications. We explored several use scenarios to demonstrate the highlighted features of PACER.

While reading work-related documents, people often google an unfamiliar word, check the authors' web site, contact the authors via email, download citations, look up a word in the document, or share a diagram with remote colleagues. However, these functions are not available when people read a hardcopy without a computer nearby, e.g. in a mobile setting. We developed Mobile Reader Assistant (MRA) based on PACER, which enables users to select arbitrary text and image content on paper for various commands, including Google, Wikipedia, Citation Downloading, Copy&Email, Dictionary and Keyword Finding.

MRA facilitates capturing and transcribing paper content. As Figure 1 shows, with a simple marquee gesture over a paper document, one can easily copy a high resolution document region based on the low resolution phone-captured picture; a short bracket gesture can extract a long paper title for search at Google Scholar (Figure 4). MRA is especially useful for selecting math symbols (e.g. Ω, Φ and θ) and foreign words, which are usually hard for users to type in. Although some of these functions (e.g. googling a

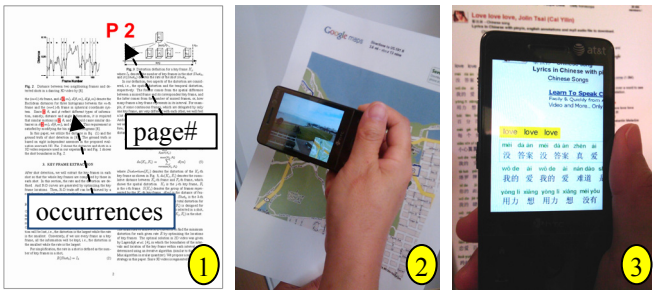

Figure 8. 1) Results of keyword search. 2) Browse street view along a route. 3) Sweep on a music score to play the song

word) could be implemented by a OCR-based system like Paper Link [4], OCR does not work on non-text content, and, more fundamentally, OCR can not provide context information of the captured images.

MRA offers rich context beyond the captured image. As demonstrated in Figure 8-1, by recognizing a single camera image, MRA can retrieve the full text of the whole document and allow the user to search for the definition of a variable γ in the printout. Upon the command execution, the phone interface highlights all the occurrences of the variable and shows the page numbers. With this visual guidance, the user can quickly find more details in the paper. Similarly, Copy & Email can automatically generate the hyperlinks pointing to the original page and full document for the copied document region. This feature is unavailable with the normal camera & email applications.

PACER Map

Cell phone-augmented paper maps are another interesting application area, and much has already been done [20-22]. PACER Map makes advances by offering fine-grained target manipulation of user-specified points of interests and routes. For example, one can point the camera phone to an arbitrary place on a printed Google map, and retrieve the street view around that location. The user can also snap to a route and browse the landmarks' pictures while sweeping the phone roughly along the route (Figure 8-2). Moreover, the user can set the source, destination and intermediate stops on the map for driving directions.

Music Scores

PACER's support for generic content expands the range of paper documents that a camera phone user can interact with. Among them are paper music scores, which are extensively used due to paper's portability, flexibility and tangibility. However, it is hard for an entry level player or singer to perform without instruction and practice. With PACER, we may link a paper score to digital media facilitating self-teaching. As Figure 8-3 shows, a user can draw a bracket gesture to select a section of lyrics, and the phone can play the music of that section. The currently sung lyric word is highlighted while the music is being played. Users can choose different sound effects or play video clips for instruction from a professional.

IMPLEMENTATION

We have fully implemented a working system consisting of a server and a standalone cell phone interface. The server, written in C++/C#, runs as an SOAP service on a Windows Server 2003 PC. The user interface is based on an HTC Touch Pro cell phone with Windows Mobile 6.1. It connects to the server via either Wi-Fi or cell phone network. We adopted the HTC native camera application for snapshots with auto-focusing, and MS DirectShow in C++ for video capture. We also used the Windows GDI APIs to directly access the raw screen buffer for fast rendering, and employed the HTC G-Sensor APIs (http://www.codeplex.com/sensorapi) for tilting detection. The demo applications like MRA are developed in C# with Windows Mobile 6 Professional SDK.

PRELIMINARY USER STUDY ON THE DESIGN IDEAS

To understand the capability and limitations of PACER, we conducted an informal user study with six colleagues (not affiliated with this project). We focused on the high level design ideas of the system, such as document recognition, loose registration and the camera-touch hybrid gestures.

Tasks

The study consisted of two sessions. The participants were first asked to use the pointer gesture to select the designated individual words within a 4-page printed document. Since all other text-related gestures (bracket, vertical bar and underline) build upon the single word selection, this task is representative for word level interaction. In the second session, the participants were asked to select the designated figures in the same document with the marquee gesture. This task requires relatively long distance pointer navigation across at least half page horizontally, so helps us examine the hybrid interaction. There were no constraints on participants' body poses, hand and document positions.

Settings

Figure 7-1 shows a test document page. The target words are highlighted in yellow. All the target words and figures are labeled with a unique ID number in the nearby margin. The targets are evenly distributed through the document. The PACER database contains 400 pages from the proceedings of ICME 2006 and 44 other technical document pages. The testing program is running on an HTC Touch Pro cell phone. It presents the stimuli, indicating the target ID for each trial. The participants read a stimulus, found the target on paper and then clicked a button to start issuing the requested gesture.

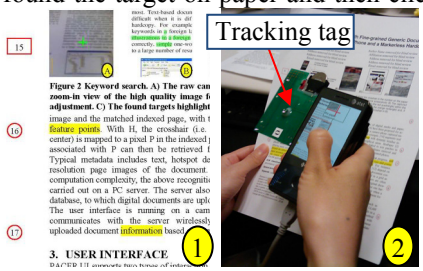

Figure 7. 1) A portion of the test document. 2) The mock-up interface.

We did not use the standalone cell phone interface for the study, but a mock-up one that detects the phone motion through an

attached LED tag board tracked by a ceiling-mounted infrared camera (Figure 7-2). This setting is due to the currently unsatisfactory robustness of the video capture module with the phone's built-in camera, which, although good enough for demonstration, often unpredictably stops working or becomes slow. The mock-up implementation helped us avoid such low-level implementation issues and focus on the evaluation of the higher level design concepts.

We believe the setting did not undermine the validity and usefulness of the study. First, the infrared camera is about 10 feet high, so out-of-tracking-range and body occlusion happened very rarely. Second, based on participant reports and our observation, the tracking board has much less impact on the interaction than other factors like the phone form factor, user hand size, and software design, thanks to the board's light weight and unobtrusive installation position (Figure 7-2). Lastly, the recognition accuracy and loose registration have little to do with specific motion detection techniques.

Procedure

The study was conducted in an author's office. Each session began with an 8-trial training, followed by a 16-trial testing. After the two sessions, the participants answered a user experience questionnaire. Finally, we had an informal interview with the participants about their questionnaire responses. The test lasted about one hour, and participants could abort the test if running out of time. Since the test was an early stage evaluation, we did not quantitatively measure the user performance, but revolved around the subjective feedback and user behavior observation.

Results and Discussion

Overall, the participants' reaction was positive and encouraging. They welcomed the general idea of combining a cell phone with paper documents, and thought PACER is useful especially in mobile settings.

Document Recognition

The document recognition accuracy is not an issue for smooth interaction with PACER. When asked "do you think the recognition accuracy is good enough for the interaction?," participants' response on a 1~7 scale was positive (Mean = 5.83, SD = 0.75). This is confirmed by the interaction logs. Out of all the 96 trials in session one, 77 had the first snapshot successfully recognized. Among the 83 session two trials (due to time limitation, two participants did not completely accomplish the session), 69 succeeded with the first snapshot. In total, the first snapshot success rate is 81.6%. This result is promising, as it reflects the performance of novice users in a realistic setting without special requirements on body poses, hand-device arrangement and lighting conditions.

The recognition failure is attributed to several factors. First, some participants complained about the inconvenience of focusing the phone camera by touching but not pressing the ENTER button, which led to blurred pictures (Figure 9-1).

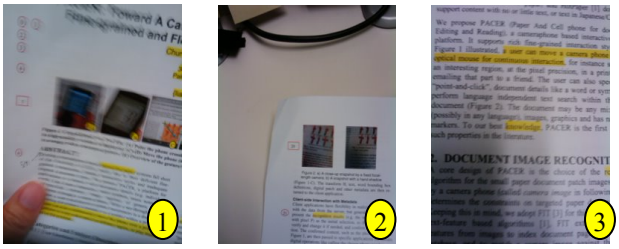

Figure 9. The example snapshots of the three recognition failure sources: 1) blurring, 2) out-of-range and 3) shadow

Second, for the targets close to page margins, the participants inadvertently included too much background while pointing the crosshair to the target (Figure 9-2). This caused incorrect and/or insufficient document visual features in the snapshots. Finally, the phone and the tracking board may create big shadows on paper (Figure 9-3), significantly distorting the original visual features.

Embodied Interaction vs. Touch Screen Interaction

In the test, participants performed embodied interaction mainly in two cases, namely when moving the phone toward a target for snapshots and when panning it to set the ends of the marquee gesture. All participants were fine in the first case, as it was natural and did not require precise pointing. But they had mixed feedback to the second case, feeling that it was faster to navigate documents by the embodied gesture than by scrolling the screen, but learning the embodied interaction was more difficult.

The problems mainly originated in the undesired document navigation caused by inadvertent phone movement that happened, for example, while a participant moved her/his thumb from the phone's bottom to the screen center for refining gestures. This issue was especially noticeable with small hands and one-handed operation. As suggested by a participant P6, one solution may be a clutch for users to manually start and stop the phone movement tracking. And a slimmer form factor of the phone is helpful too.

The touch screen operation was more familiar to users. Participants liked the occlusion callout and auto-scrolling for word selection. But if the target was not reachable by the thumb, participants had to change hand pose to get it. Participant P4 pointed out that it may be harder with a wider form factor. Embodied interaction would be useful to move the target toward the thumb. Screen gestures like "sliding" on iPhones is an alternative. Participant P3, as an iPhone user, preferred pure touch screen operations. We think the touch screen advantage is task-specific. For browsing annotations throughout document pages or exploring in a large map, the embodied interaction may win, as suggested by the prior experiments [23]. Deeper study in specific application areas is needed to better understand the two interaction schemes.

Loose Registration vs. Strict Registration

The participants appreciated the high fidelity document images retrieved from the server. When asked "which specific features do you like most?," P1 said "[I like] the

clear display. It is easier than photos, which are not clear enough to select or read words." The responses confirmed the effectiveness of our design.

Our observation suggested that loose registration could lower the mental and physical demand on the participants, because it does not require accurate phone-paper coordination. The participants often relaxed after taking the snapshot, sat back, rested their elbows on the chair arms, and continued the operation on the phone. Then they could just interact with the retrieved digital document without the hardcopy in front of the phone. This was especially common for document manipulations in a small area (e.g. selecting a word or a small figure) for which the cell phone screen is able to serve enough context. The issue of AR mental/physical demand has been reported in some prior work such as [20], which indicates that a strict registration map AR interface, MapLens, attracts more users attention to the device itself than a non-AR cell phone interface, impeding user interaction with the surrounding environment. This issue could be moderated with loose registration. Of course, loose registration has certain limitations. It may not work if some useful information (e.g. handwritten annotations) only exists in the physical copy of the document, or if the misalignment between the screen content and paper documents becomes unacceptable.

More generally, an AR interface involves three basic information layers, namely the physical object (e.g. the paper document in PACER), the digital model of the object (e.g. the corresponding digital document) and the overlay information. Strict registration aims at the perfect alignment of the three layers, while loose registration looses the requirement on the object-model alignment and focuses on the model and overlay. As we discussed above, two approaches have both pros and cons, which should be further explored in the future, especially in context of different use scenarios such as interactive paper documents and 3D object/scene augmentation.

One-handed vs. Two-handed Operations

In the original design, we preferred one-handed over two-handed operation, as the free hand could hold the document, arrange pages on a desk, or perform other tasks with more flexibility. The study suggested that it may not work for some users. For example, some participants with relatively small hands and short thumbs had difficulty in operating the phone with only one thumb. Besides, the HTC Touch Pro phone is relatively thick and heavy, which caused user fatigue after long use of one-handed operations. However, participants could adapt to the new interface. Three participants (P2, P5 and P6) used two hands simultaneously to avoid hand pose changing (Figure 10), even without being told to do so. P5 thought that the second session was actually easier than the first one, partly because she could exploit the embodied interaction for quick navigation.

Figure 10. Two-handed interaction.

One-handed interaction techniques like AppLens and LaunchTile [11] could be employed to refine the design. Two-handed interaction could ease embodied interaction, but the participants had to give up some flexibility of document spatial arrangement. These findings urge more studies for a better trade-off.

More Efficient Interaction

Some participants complained that they had to spend time to take still images and wait for recognition results before gesturing. This is because the HTC camera auto-focus is currently only available in still image mode, which thus produces better query images than the video mode. This problem can be solved with newer phones with better video capture capability and/or with more robust algorithms recognizing low quality video. And the future implementation may allow users to gesture on the raw video frames for initial coarse selection, while waiting for the server response. The users can refine the selection after receiving the high fidelity document image.

CONCLUSION

PACER is a cell phone-based interactive paper system that features the fine-grained and flexible interaction through camera-touch hybrid input. PACER allows for interaction with document details (such as individual words, characters and pixels) through a variety of gestures including marquees, lassos, vertical bars, underlines and brackets. It enables a wide range of cell phone-based interactive paper applications for GUI-like user experience on paper.

We examined key design challenges (such as slow image recognition, inaccurate physical-digital document registration and hand-jitter) and proposed several novel techniques including camera-touch hybrid input, semi-real time document tracking and loose registration. We have fully implemented a working system to demonstrate the feasibility of the design. With an alternative mockup implementation, a preliminary user study on the design ideas showed positive user feedback and revealed some usability issues caused by phone form factor, user hand size and undesired motion detection.

Future work will involve more design-test iterations for better understanding of the embodied and touch interaction, as well as one and two handed interaction. We will further explore the pros and cons of loose/strict registration in more applications scenarios. Moreover, we plan to integrate other mobile devices, like a mobile projector and a digital pen, into PACER to explore new interaction with paper.

ACKNOWLEDGEMENT

We thank Don Kimber and Tony Dunnigan for their generous help and useful suggestions. We also thank the user study participants. Lastly, we appreciate the anonymous CHI reviewers for their insightful comments.

REFERENCES

1. Adams, A., N. Gelfand, and K. Pulli, Viewfinder Alignment. Computer Graphics Forum (Proc. Eurographics), 2008. 27(2): p. 597-606.
2. Albinsson, P.-A. and S. Zhai. High precision touch screen interaction. Proceedings of CHI'03, pp. 105-112.
3. Anoto, http://www.anoto.com.
4. Arai, T., D. Aust, and S.E. Hudson. PaperLink: a technique for hyperlinking from real paper to electronic content. Proceedings of CHI'97, pp. 327 - 334.
5. Erol, B., Emilio Antunez, and J.J. Hull. HOTPAPER: multimedia interaction with paper using mobile phones. Proceedings of Multimedia'08, pp. 399-408.
6. Forlines, C., R. Balakrishnan, P. Beardsley, J. van Baar, and R. Raskar. Zoom-and-pick: facilitating visual zooming and precision pointing with interactive handheld projectors. Proceedings of UIST'05, pp. 73-82.
7. Guimbretiere, F. Paper Augmented Digital Documents. Proceedings of UIST'03, pp. 51 - 60.
8. Hare, J., P. Lewis, L. Gordon, and G. Hart. MapSnapper: Engineering an Efficient Algorithm for Matching Images of Maps from Mobile Phones. Proceedings of Multimedia Content Access'08: Algorithms and Systems II pp.
9. Hinckley, K., P. Baudisch, G. Ramos, and F. Guimbretiere. Design and analysis of delimiters for selection-action pen gesture phrases in scriboli. Proceedings of CHI'05, pp. 451-460.
10. Hull, J.J., B. Erol, J. Graham, Q. Ke, H. Kishi, J. Moraleda, and D.G.V. Olst. Paper-based Augmented Reality. Proceedings of Int. Conf. on Artificial Reality and Telexistence'07, pp. 205--209.
11. Karlson, A.K., B.B. Bederson, and J. SanGiovanni. AppLens and launchTile: two designs for one-handed thumb use on small devices. Proceedings of CHI'05, pp. 201-210.
12. Kim, J., S.M. Seitz, and M. Agrawala. Video-based document tracking: unifying your physical and electronic desktops. Proceedings of UIST'04, pp. 99-107.
13. Kratz, S. and R. Ballagas. Unravelling seams: improving mobile gesture recognition with visual feedback techniques. Proceedings of CHI'09, pp. 937-940.
14. Kurtenbach, G., The design and Evaluation of Marking Menus, PhD thesis, University of Toronto. 1993
15. Liao, C., F. Guimbretière, K. Hinckley, and J. Hollan, PapierCraft: A Gesture-Based Command System for Interactive Paper. ACM ToCHI, 2008. 14(4): p. 1-27.
16. Liu, Q., H. Yano, D. Kimber, C. Liao, and L. Wilcox. High Accuracy And Language Independent Document Retrieval With A Fast Invariant Transform. Proceedings of ICME'09, pp.
17. Liu, X. and D. Doermann, Mobile Retriever: access to digital documents from their physical source. Int. J. Doc. Anal. Recognit., 2008. 11(1): p. 19-27.
18. Lowe, D.G., Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision, 2004. 60(2): p. 91-110.
19. Mackay, W.E., G. Pothier, C. Letondal, K. Bøegh, and H.E. Sørensen. The missing link: augmenting biology laboratory notebooks. Proceedings of UIST'02, pp. 41 - 50.
20. Morrison, A., A. Oulasvirta, P. Peltonen, S. Lemmela, G. Jacucci, G. Reitmayr, J. Näsänen, and A. Juustila. Like bees around the hive: a comparative study of a mobile augmented reality map. Proceedings of CHI'09, pp. 1889-1898.
21. Reilly, D., M. Rodgers, R. Argue, M. Nunes, and K. Inkpen, Marked-up maps: combining paper maps and electronic information resources. Personal Ubiquitous Comput., 2006. 10(4): p. 215-226.
22. Rohs, M. Real-world interaction with camera-phones. Proceedings of UCS. IPSJ Press'04, pp. 74-89.
23. Rohs, M., J. Schoning, M. Raubal, G. Essl, and A. Kruger. Map navigation with mobile devices: virtual versus physical movement with and without visual context. Proceedings of ICMI07, pp. 146-153.
24. Rohs, M. and A. Oulasvirta. Target Acquisition with Camera Phones when userd as Magic Lenses. Proceedings of CHI'08, pp. 1409-1418.
25. Sellen, A.J. and R.H.R. Harper, The Myth of the Paperless Office. 1st ed. 2001: MIT press.
26. Tsandilas, T., C. Letondal, and W.E. Mackay. Musink: composing music through augmented drawing. Proceedings of CHI'09, pp. 819-828.
27. Vogel, D. and P. Baudisch. Shift: a technique for operating pen-based interfaces using touch. Proceedings of CHI'07, pp. 657-666.
28. Wagner, D., G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg. Pose Tracking from Natural Features on Mobile Phones. Proceedings of ISMAR'08, pp.
29. Wang, J., S. Zhai, and J. Canny. Camera phone based motion sensing: interaction techniques, applications and performance study. Proceedings of UIST'06, pp. 101-110.
30. Weibel, N., A. Ispas, B. Signer, and M.C. Norrie. Paperproof: a paper-digital proof-editing system. Proceedings of CHI '08 pp. 2349-2354.
31. Wellner, P., Interacting with paper on the DigitalDesk. Communications of the ACM, 1993. 36(7): p. 87 - 96.
32. Yeh, R.B., C. Liao, S.R. Klemmer, F. Guimbretière, B. Lee, B. Kakaradov, J. Stamberger, and A. Paepcke. ButterflyNet: A Mobile Capture and Access System for Field Biology Research. Proceedings of CHI'06, pp. 571-580.