PACER: Toward A Cameraphone-based Paper Interface for Fine-grained and Flexible Interaction with Documents

Chunyuan Liao, Qiong Liu
FX Palo Alto Laboratory
3400 Hillview Ave, Bldg 4
Palo Alto, CA 94043, U.S.A
1-650-842-4831

{liao, liu}@fxpal.com

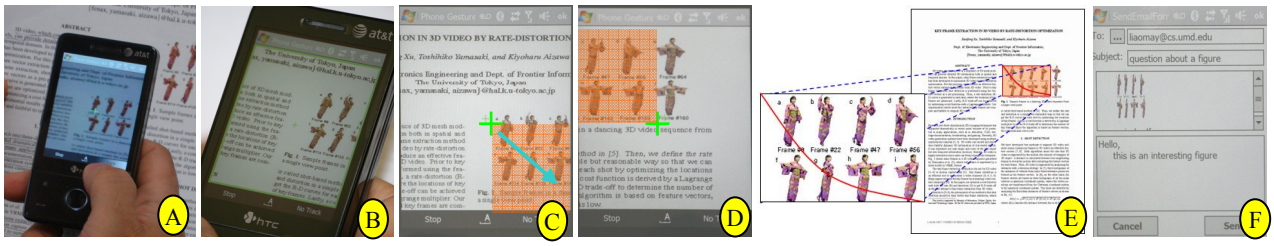

Figure 1 Copy&Email via PACER. (A) Point the phone crosshair to an interesting area on paper. (B) Once the area is recognized, its high-quality version is displayed. (C)-(D) Move the phone (in the arrow direction) over the paper to select a region (highlighted in orange) with a marquee gesture. (E) Overview of the gesture/region within the document. (F) The region is sent via email.

ABSTRACT

Existing cameraphone-based *interactive paper* systems fall short of the flexibility of GUIs, partly due to their deficient fine-grained interactions, limited interaction styles and inadequate targeted document types. We present PACER, a platform for applications to interact with document details (e.g. individual words, East Asian characters, math symbols, music notes, and user-specified arbitrary image regions) of generic paper documents through a camera phone. With a see-through phone interface, a user can discover symbol recurrences in a document by pointing the phone's crosshair to a symbol within a printout. The user can also continuously move the phone over a printout for gestures to copy and email an arbitrary region, or play music notes on the printout.

Categories and Subject Descriptors

H5.2 [**User Interfaces**]: Interaction Styles.

General Terms

Design, Human Factors, Algorithms

Keywords

Camera phone, paper interface, fine-grained, generic document

1. INTRODUCTION

Cameraphone based *interactive paper* systems [1, 2, 4] strive to enable digital interactions on paper and serve rich interactions like a computer GUI. However, existing practices still fall short of that goal. First, most prior systems do not offer sufficient *fine-grained* operations on user-specified document details such as individual words, math symbols or an arbitrary image region. HotPaper [1] focuses on text-patch level multimedia annotation, and [4] augments pre-defined regions of a paper map with dynamic weather information. Second, previous work mainly revolves around point-and-click interaction [2, 4], with a lack of *continuous interactions* like lasso or underline gestures. Finally, document types handled by these systems are *limited*. [4] requires 2D bar codes on paper, and HotPaper [1] does not support content with no or little text, or text in Japanese/Chinese.

We propose PACER (Paper And Cell phone for document Editing and Reading), a cameraphone based interactive paper platform. It supports rich fine-grained interaction styles. As Figure 1 illustrated, a user can move a camera phone like an optical mouse for continuous interaction, for instance selecting an interesting region, at the pixel precision, in a printout and emailing that part to a friend. The user can also specify, via "point-and-click", document details like a word or symbol, and perform language independent text search within the paper document (Figure 2). The document may be any mix of text (possibly in any language), images, graphics and has no special markers. To our best knowledge, PACER is the first one with such properties in the literature.

2. DOCUMENT IMAGE RECOGNITION

A core design of PACER is the choice of the recognition algorithm for the small paper document patch images captured by a camera phone (called *camera image* in following text). It determines the constraints on targeted paper document types. Keeping this in mind, we adopt FIT [3] for this purpose. Unlike text-feature based algorithms [1], FIT extracts pixel level features from images to index document pages in a PACER database, and to match a camera image against the indexed pages. FIT does not require any special markers on paper, works on both text and non-text content, and is language-independent. Also it is robust to image scaling, rotation and occlusion. FIT

ACM Multimedia'09, October 19~24, 2009, Beijing, China.

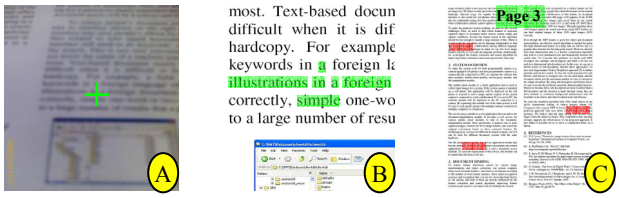

Figure 2 Keyword search. A) The raw camera image. B) The zoom-in view of the high quality image for verification and adjustment. C) The found targets highlighted in hit pages

achieved 99.73% page recognition rate for the 2188-page ICME06 proceedings, 99.9% for a 504-page Japanese book, and 99.39% for a personal collection of 545 photos [3].

To identify the specific content pointed by the phone crosshair, we derive a homographic transform H between the camera image and the matched indexed page, with their corresponding feature points. With H, the crosshair (i.e. the camera image center) is mapped to a pixel P in the indexed page. The metadata associated with P can then be retrieved from the database. Typical metadata includes text, hotspot definitions and high resolution page images of the document. Due to the high computation complexity, the above recognition and retrieval are carried out on a PC server. The user interface is running on a camera phone, which communicates with the server wirelessly.

3. USER INTERFACE

PACER UI supports two types of interaction, namely point-and-click and continuous phone gestures.

3.1 Point-and-click

Point-and-click is suitable for interaction with pre-defined objects (e.g. individual words, East Asian characters, math symbols, icons). For instance, a user can point a camera phone to a word (Figure 2-a), and then take a snapshot. The snapshot is sent to the server to query the corresponding patch in high resolution, the transform H and the document text. Once the high-quality patch shows on the phone (Figure 2-b), the user can then verify and adjust the results (e.g. the coordinates of P and the words hit by P) by tapping on the touch screen. Upon user confirmation, the selected word is passed to the full-text search application, which consequently highlights the hits within the zoomable hosting pages (Figure 2-c). Similarly, the user can perform functions such as "dictionary", "google", "multimedia hotspot", "email" and so on.

Replacing the original camera image with the high-quality one is an important design for fine-grained operations. Because the camera images often suffer from their low quality, due to phone movement, fixed focal lengths, low image resolutions and/or undesirable lighting conditions. Although many such images (e.g. Figure 2-a) can be recognized by FIT and the crosshair-pointed words can be highlighted, the user can hardly see what the blur small words are. Moreover, due to the approximation of the transformation matrix H, the overlays may be off from the user's targeted details. These issues can be addressed by manipulating the retrieved high-quality page images.

3.2 Continuous phone gestures

With PACER gestures, the phone acts as an optical mouse that the user continuously drags within the document. Besides the

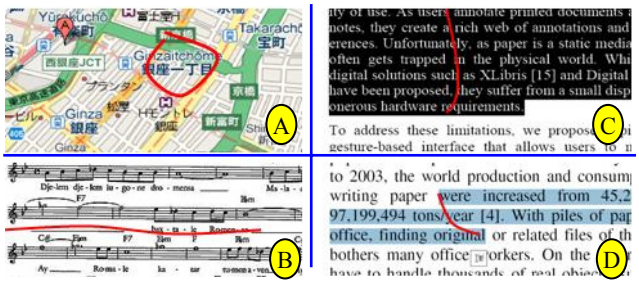

Figure 3. PACER gestures (highlighted in red). A) Lasso, B) Underline, C) Vertical bar, D) Bracket

marquee gesture (Figure 1) for rectangle region selection, PACER also allows the user to, as exemplified in Figure 3, query hotels in a geographic region indicated by a *lasso* on a map (a), play music notes while applying a *underline* to the staff (b), quote multiple lines of text from a paper document via a *vertical bar* (c) or sentences through a *bracket* (d).

The continuous interaction relies on the same server-client communication and high-quality-image substitution as the point-and-click interaction. Yet, it requires real time tracking of the phone crosshair. Despite FIT's high recognition accuracy, it takes a phone about 1 second to accomplish a query session through the wireless link. In contrast, camera-based motion detection is much faster (15fps on a HTC Touch Pro cell phone), but merely serves relative movement with accumulated errors [5]. We opt to combine the two methods, by applying accurate registration at fixed intervals of time (e.g. 1~2 seconds) and estimating crosshair position with the motion detection between two consecutive frames. Every recognition session resets the motion detection to reduce the accumulated errors.

4. CONCLUSION

We encapsulate the paper document recognition and client UIs into a set of APIs, and emulate a mouse-like pointing device for generic applications. The design enables much richer interaction than existing systems. Beyond the conventional point-and-click interaction, users of our system can apply continuous gestures including marquee selections, lassos, vertical bars, underlines and brackets, which effectively promote a GUI-like user experience on paper documents.

5. REFERENCES

[1] Erol, B., Emilio Antunez, and J.J. Hull. HOTPAPER: multimedia interaction with paper using mobile phones. Proceedings of Multimedia'08, pp. 399-408.

[2] 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: Algorithms and Systems II pp.

[3] 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

[4] Rohs, M. Real-world interaction with camera-phones. Proceedings of UCS. IPSJ Press (2004), pp.

[5] Wang, J., S. Zhai, and J. Canny, Camera phone based motion sensing: interaction techniques, applications and performance study, in UIST'06.