FACT: Fine-grained Cross-media Interaction with Documents via a Portable Hybrid Paper-Laptop Interface

Chunyuan Liao[1], Hao Tang[2], Qiong Liu[1], Patrick Chiu[1], Francine Chen[1]
[1]FXPAL, 3400 Hillview Ave, Bldg 4, Palo Alto, California 94043, U.S.A.
[2]Dept. of ECE, University of Illinois at Urbana-Champaign, Urbana, Illinois 61801, U.S.A.
[1]{liao, liu, chiu, chen}@fxpal.com, [2]haotang2@uiuc.edu

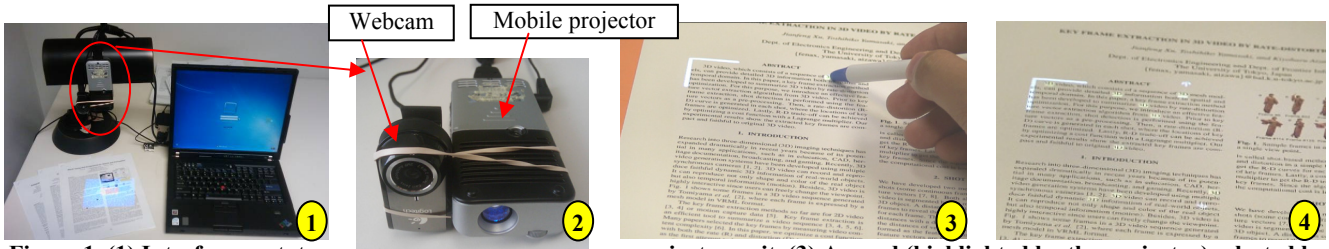

Figure 1. (1) Interface prototype, (2) Close-up of the camera-projector unit, (3) A word (highlighted by the projector) selected by a pen tip for full-text search, (4) The resulting occurrences of the word highlighted by the projector.

ABSTRACT
FACT is an interactive paper system for fine-grained interaction with documents across the boundary between paper and computers. It consists of a small camera-projector unit, a laptop, and ordinary paper documents. With the camera-projector unit pointing to a paper document, the system allows a user to issue pen gestures on the paper document for selecting fine-grained content and applying various digital functions. For example, the user can choose individual words, symbols, figures, and arbitrary regions for keyword search, copy and paste, web search, and remote sharing. FACT thus enables a computer-like user experience on paper. This paper interaction can be integrated with laptop interaction for cross-media manipulations on multiple documents and views. We present the infrastructure, supporting techniques and interaction design, and demonstrate the feasibility via a quantitative experiment. We also propose applications such as document manipulation, map navigation and remote collaboration.

ACM Categories and Subject Descriptors
H.5.2 User Interfaces (D.2.2, H.1.2, I.3.6): Interaction Styles

General Terms
Human Factors, Design

Keywords
Paper interface, camera, projector, fine-grained, cross-media

MM'10, October 25–29, 2010, Firenze, Italy.

1. INTRODUCTION
Paper and computers have unique and complementary advantages [26]. Paper is comfortable to read and annotate, light to carry (to a degree), flexible to arrange in space, robust to use in various settings, and well accepted in social settings. Computers are powerful in multimedia presentation as well as document editing, archiving, sharing and search. Even though new computers have improved robustness, readability, and flexibility, there are still many technical difficulties and cost efficiency concerns about completely replacing paper with computers. Thus paper and computers are extensively used in parallel in many scenarios. We believe this situation will continue in the foreseeable future. It is therefore important to seamlessly combine the advantages of paper and computer.

We focus on *simultaneous* use of paper documents and a computer side by side on a table, which is a very typical workstation setting for knowledge workers. With this setting, people may, for example, read articles on paper and write a summary on a computer. In conjunction with the read-write activities, users often need to search on the Web for extra information about specific content, quote a sentence or diagram from an article, or share interesting document segments with others via email or IM.

The problem, however, is that the existing mixed use of paper and computers does not provide a convenient transition between the two media. The content on paper is insulated from digital tools such as remote sharing, hyperlinks, copy-paste, Web search and keyword finding. This gap between paper and computers causes low efficiency and degrades user experience. For example, it is nearly impossible for a person to perform a Web search of an unknown foreign word in a book if she does not speak that language; it is difficult to copy a picture from paper to a Word document on a nearby computer; and there is no easy way to annotate a technical term in a paper document with web pages.

Much has been done to address these issues. However, the prior work still does not bridge the paper-computer gap completely. First, most of the existing systems such as PlayAnywhere [30], DocuDesk [8] and Bonfire [14] focus on interaction with a whole page or document rather than supporting fine-grained manipulation within the pages (e.g. individual words and user-specified regions). Second, these systems only support limited digital functions on paper, typically page-level hyperlinks [8, 30] , spatial arrangement tracking [15], and text transcribing [23, 29], and lack a general framework to address the paper-computer gap. Third, these systems may interfere with existing workflows due to their inflexible hardware configurations [29] or the requirement for specially marked paper [28].

In response, we propose a novel vision-projection based interactive paper system called *FACT* (Fine-grained And Cross-media inTeraction). As illustrated in Figure 1-1,2, the FACT system features a portable hybrid user interface consisting of a compact camera-projector unit, a laptop and ordinary printed paper documents without any barcodes or markers. Using the camera-projector unit in conjunction with precise content-based image recognition and coordinate transform software, the system allows users to use pen gestures to specify ***fine-grained*** paper document content (e.g. individual Latin words, Chinese and Japanese characters, symbols, icons, figures, and arbitrary user-chosen regions) for digital operations. For instance, to find the definition of an unfamiliar word in a paper document, a user can point a pen tip to the word (Figure 1-3) and issue a "Keyword Search" command. In response, the occurrences of that word on paper are highlighted by the projector (Figure 1-4), so the user can easily browse the occurrences for the definition.

This fine-grained paper interaction is well integrated with the laptop for ***cross-media*** interaction, combining the complementary affordances of paper and computers. For example, users can easily extract an arbitrary chunk of text or image region from the paper, and insert it into a Word document on the laptop (Figure 12-1,2); they can create a hyperlink between a specific term on paper and a Wikipedia page as annotation (Figure 12-3); and they also can navigate Google Street View on the laptop by pointing to a place on a paper map (Figure 12-4,5).

The contributions of this paper are four-fold, including 1) the use of a robust and precise content-based approach for fine-grained physical-digital interaction mapping, 2) the quantitative analysis of the mapping performance, 3) the design of the pen gesture command system that is geared towards the camera-projector interface, and 4) the interaction techniques for fine-grained multiple document manipulation across the paper-computer boundary.

The remainder of this paper is organized as follows. We begin with related work in section 2, and propose design principles in section 3. Guided by the principles, we outline the system in section 4, and present more details about physical-digital mapping and command issuing in sections 5 and 6 respectively. Based on these techniques, we demonstrate the enabled cross-media interaction and applications in section 7. We then report the performance evaluation of the physical-digital document mapping in section 8, and conclude the paper with future work in section 9.

2. RELATED WORK

Cameras and projectors have been adopted by many systems for interactive paper. These systems vary in terms of interaction granularity, hardware settings and the role of paper. First, systems like EnhancedDesk [16], Augmented Surfaces [24], PlayAnywhere [30], video-based document tracking [15] and recently Sixth-sense [22] and Bonfire [14] focus on page- or document-level interaction, whereas FACT supports fine-grained (e.g. word-level) interaction. Second, many of the systems require special devices or markers on paper to track paper documents. For example, Docklamp [6] and DocuDesk [8] identify paper documents with barcodes. Paper Windows [13] relies on a Vicon motion capture system consisting of 12 cameras and 8 infrared markers on each paper sheet. In contrast, FACT works for normally printed generic paper documents by using content-based document recognition techniques. Third, systems like DigitalDesk [29], CamWorks [23] and PaperLink [3] treat a paper document as a transient information source without persistent linkage to digital information. FACT instead maintains a link from a paper document to its digital version over sessions.

Other sensing technologies like Anoto [2] are also possible for interactive paper. PapierCraft [17] adopts an Anoto digital pen for a pen-gesture-based command system on paper. PenLight [27] extends this idea by using a pen-top projector for direct visual feedback on paper. MouseLight [28] further combines a hand-held spatially-aware mobile projector and a digital pen for bi-manual interaction. Different from FACT, these systems require the dedicated digital pen and paper with Anoto patterns.

Furthermore, mobile phones can be used for digital interaction on paper. HotPaper [7] facilitates creating and retrieving multimedia annotations attached to text-based paper documents. PACER [18] supports hybrid camera-touch gestures for fine-grained interaction on paper. Map torchlight [25] utilizes a projector in a cell phone to augment a paper map. These systems focus on in-air hand-held operations of mobile phones and paper, which are very different from FACT's pen-based cross-media interaction on a table.

3. DESIGN PRINCIPLES

Our ultimate goal is to achieve a practical system that can be easily deployed in real life. Towards this goal, we set up the following principles to guide our design of the system.

Portable and easily configurable user interfaces. The user interface should be portable to support the increasing demands on work-at-anywhere. It should not require non-portable hardware such as ceiling-mounted overhead cameras and projectors. And it should avoid complex calibration and other configurations required to run the system in different locations.

No special paper or reading devices. The system should not alter the original paper documents with any markers (e.g. barcodes, data glyph [11] or RFIDs). It should not rely on any dedicated reading devices like Anoto digital pens [2] for paper interaction.

No special printing. It should accommodate ordinary printing styles, document layouts, page sizes and font sizes. For instance, it should work with letter-size articles in 9 and 10 point fonts from ACM Multimedia and IEEE ICME proceedings. The last two principles aim to ensure that FACT can be seamlessly integrated into existing document processing workflow for real life deployment without much overhead.

Support for general document content and applications. To achieve computer-like user experience with paper through the seamless integration of paper and computers, the system should

be easily generalized for generic paper document content types, including text (possibly in multiple languages), graphics and pictures. And it should be flexible enough to support a wide range of computer functions on paper for various application areas.

Minimal limitations on interaction flexibility. The system should respect the existing flexible use of paper. For instance, users should be able to freely arrange paper sheets on a table, rest their hands and annotate paper documents as they usually do.

Exploiting user adaptation. To be more practical, instead of building a perfect passive vision system, we should take advantage of user adaptation, presenting appropriate system status information to users and guide them to adjust their interaction behaviors (e.g. re-arrange paper or avoid hand occlusion of documents) to improve coordination with the imperfect system for successful operations.

4. SYSTEM OVERVIEW

The FACT system basically bridges a physical and a digital document workspace (Figure 2). It consists of three key components, namely the *camera processor*, *projector processor* and *paper-computer coordinator*. The camera processor captures and analyzes camera video frames to recognize and track paper documents (e.g. a printed Google map in Figure 2-left), and to detect and trace the user's pen tip. It finds the digital version of the recognized paper documents, and interprets the pen tip operations as equivalent mouse pointer manipulations on the digital version. When needed, the paper-computer coordinator interacts, on behalf of the paper documents, with other documents or views, such as Google Street View of the paper map (Figure 2-right). In return, the visual output for the paper document is forwarded to the projector processor, which generates the projection that is precisely aligned with the paper document content for direct visual feedback.

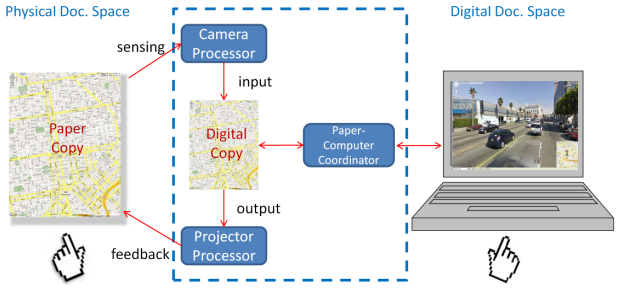

Figure 2. An overview of the FACT system

Although this basic procedure seems similar to that of existing camera-projector-based systems such as DigitalDesk [29], PlayAnywhere [30] and Bonfire [14], FACT distinguishes itself by allowing users to manipulate fine-grained document details (e.g. individual words, symbols and arbitrary regions) with a pen gesture based command system that is well adapted to the camera-projector setting. To realize this unique property and follow our design principles, we have addressed a set of challenges.

From the aspect of system input, we need to choose a document image recognition algorithm that works for ordinary paper documents without any special markers or barcodes. Furthermore, the fine-grained interaction requires the system to precisely determine the small document details pointed to by a user from low quality webcam video (usually of 640x480 pixels). And it should be robust to hand occlusion, paper overlapping and users' freeform marks on paper.

From the aspect of system output, the projected information may interfere with the document recognition and become illegible when mixed with the paper document content. Moreover, the currently available portable projectors are limited by their low brightness and small projection area, which requires some special interface design other than the regular computer GUIs.

4.1 An Overview of the Key Techniques

Towards the goal of fine-grained interaction with paper documents, we address the above challenges through supporting techniques at two levels, namely *content-based* physical-digital interaction mapping and *context-aware* command issuing.

4.1.1 Content-based Interaction Mapping

The physical-digital interaction mapping basically translates a pen-pointing action on a paper document into an equivalent digital counterpart. We take a content based approach, using SIFT-like local visual features of a paper document [19, 21] to identify its digital version and to build precise coordinate transforms between the camera, projector, and document reference frames. This approach does not require any special markers or modification of ordinary paper documents. It is robust to partial occlusion, luminance change, scaling, rotations and perspective distortion (to a certain degree), which is important for retaining the inherent flexibility of paper [19, 21]. For pen tip detection, we currently use a Hue based histogram back-projection method similar to Bonfire's [14], with the assumption that the pen tip color is distinguishable from the background.

The accuracy of the document recognition, coordinate transform and pen tip detection certainly cannot be perfect. To mitigate this inherent limitation of the system, we deliberately project some guidance information on paper, such as the current input focus and coordinate transform result. Accordingly, users can adjust their pen tip, paper orientation and location or hand posture to facilitate the mapping.

Moreover, sometimes the mapping might not work with user input adjustment, due to insufficient visual features or bad lighting conditions. FACT allows the users to switch from paper to the laptop, directly working on the corresponding digital versions with operations similar to the paper interface. This interface switching ensures complete system functionality for realistic deployment.

4.1.2 Context-Aware Command Issuing

Based on the physical-digital interaction mapping, FACT allows users to issue pen gesture based commands on paper. Many pen-computer-like gestures such as Underline, Lasso and Marquee are made available on paper documents for selecting specific words or segments. And popup menus are employed for specifying an action to be applied to the selected content.

Nevertheless, unlike the regular on-screen interfaces, a FACT system renders the on-paper gestures and menus with consideration of its context. For instance, it only renders the outermost contour of the selected document content as feedback. This rendering strategy reduces the interference with the original document visual features, which are the basis of the precise physical-digital interaction mapping. Moreover, we project the menus on a relatively blank area on paper or table, by analyzing the texture distribution of the webcam images to locate such an

area. In this way, we make the projected menus more legible and reduce projection interference with the document features. Additionally, in case where there is no appropriate place to project a menu or other information, the laptop screen can be exploited instead.

These supporting techniques enable the fine-grained interaction with paper document details. This on-paper interaction is then combined with the on-laptop interaction for cross-media multi-document and multi-view manipulation.

4.2 Basic Data Flow for On-Paper Interaction

Figure 3 presents the overview of the data flow in FACT. A camera image is submitted to the image feature extractor to obtain a set of local visual features $F_c = \{F_1,\ldots, F_n\}$. These features are matched against those in a document image feature database. All document pages P_i (i=1…*m*) that has sufficient features V_i matching F_c are taken as the corresponding digital pages of the physical ones showing in the camera image. Based on the feature point correspondence, FACT then derives a homographic transform H_j from the camera image to the matched digital page J (J=1…*m*). This transform is combined with the detected pen tip position T_p in the camera image to determine the specific input focus document page P_f to which the pen tip is pointing. Then the pen tip is interpreted as the equivalent mouse cursor at the position $T_f = H_f * T_p$ in digital page P_f. In the subsequent processing, like a pen-based computer, FACT accumulates the pen point samples as a gesture stroke, and accordingly selects the specific document content $\{T_1,\ldots,T_k\}$ from a metadata database, which stores, for each registered document page, the high resolution version of the page, text, bounding boxes of words and

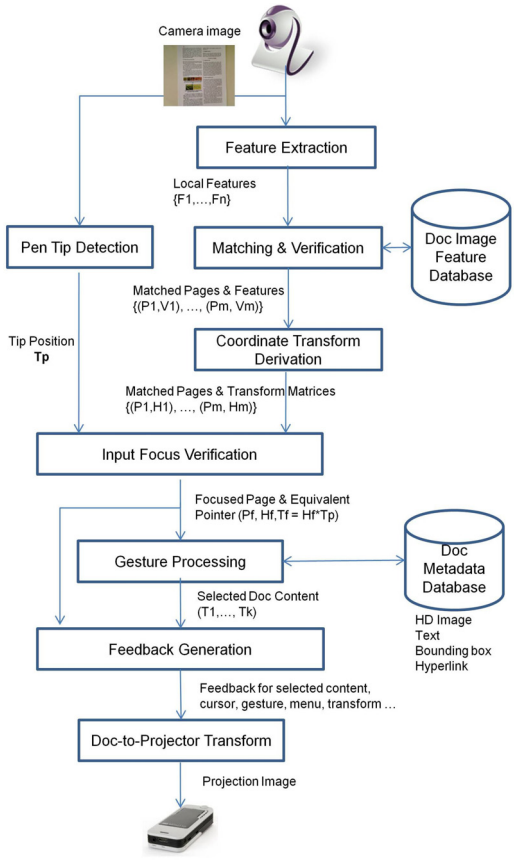

Figure 3. Overview of the FACT data flow

symbols, hyperlinks and so on. In the meantime, FACT generates feedback to indicate the current cursor position, focused page, transform accuracy, gesture and selected document content. This information is then converted into a projection image to overlay the visual feedback directly on paper. We explain the key techniques in detail below.

5. CONTENT BASED PHYSICAL-DIGITAL INTERACTION MAPPING

At the core of FACT is the mapping from physical pen interaction with paper documents to equivalent digital operations. Basically, there are two approaches. One is "transient paper," adopted by systems like CamWorks [23] and PaperLink [3]. These systems use paper as a transient media without persistent linkage to external digital resources. Thus they mainly utilize the information directly from camera images of paper documents, such as text extracted via Optical Character Recognition (OCR). In contrast, other systems like Bonefire [14] and DocuDesk [8] maintain persistent linkage from paper documents to the corresponding digital versions and other contextual information registered with the system in advance. We call this approach "persistent paper."

The first approach does not rely on document registration beforehand, but lacks context information, which is necessary for some applications. For instance, Keyword-Finding often needs to search in *all* pages of a document for occurrences of a word selected in *one* page; Remote-Sharing may need to retrieve the ownership and other related documents for a document page. Moreover, without the paper-digital linkage, a paper document acts just as a transient input media for the system. Users can extract text or image information from it, but cannot add to it persistent digital information such as video annotations that can be retrieved in later sessions. These limitations severely constrain the generalization of the system. Therefore, we opt for the "persistent paper" one as the primary approach, linking a paper document to pre-registered context information through image recognition. If the document is not registered, the system can turn to the second approach with limited functions.

To avoid requiring barcodes, special paper or reading devices, we choose content based document image recognition algorithms to identify a normally printed generic document with its natural visual features. In this way, our system is compatible with existing document processing practices, and end users do not need to do anything special with their everyday printouts. In order to accommodate general document content, we use generic image local feature-based algorithms such as SIFT [21] and FIT [19]. The current implementation uses FIT, as it is more efficient than SIFT in search time and feature storage. The features rely on only *local* information around key points, thus this algorithm is robust to partial occlusion, luminance change, scaling, rotations and perspective distortion (to a degree), which enables FACT to retain the inherent flexibility of paper documents.

5.1 Precise Coordinate Transform

To interpret fine-grained pen-paper interaction and generate a well-aligned projection overlay on paper (e.g. selecting an individual word in Figure 1-3,4), it is critical to establish the precise coordinate transform $\mathbf{H_{in}}$ for input (from a camera image to its matched digital document page) and $\mathbf{H_{out}}$ for output (from the digital document page to the projection image). The relations between the three coordinate systems are illustrated in Figure 4.

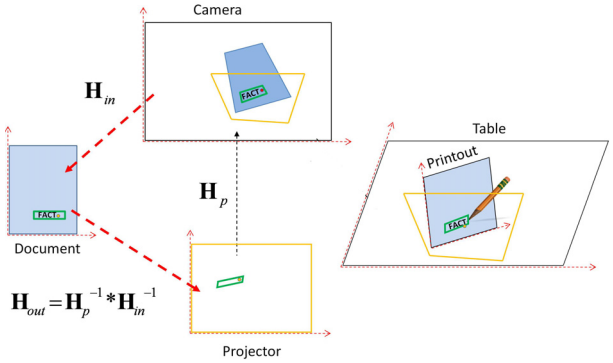

Figure 4. An illustration of coordinate transform in FACT.

To derive $\mathbf{H_{in}}$, existing systems like DigitalDesk [29] and PlayAnywhere [30] detect edges for page boundaries in camera images, and match the enclosed quadrangle to the digital page boundary. This method is sufficient for coarse granularity interaction, such as projecting a video onto a blank paper sheet [30]. However, successful edge detection usually requires enough luminance contrast between paper and background and little occlusion. For instance, this method may fail for the case shown in Figure 5-1. In addition, the boundary based approach does not consider printing margins, which may vary with specific printers. Therefore the paper boundary does not exactly match the original digital page boundary. N-up printing just exacerbates this situation (Figure 5-2).

We instead utilize a content based approach, establishing the homographic transform $\mathbf{H_{in}}$ from the one-to-one feature point correspondence between a camera video frame and the recognized document image in the database. At least four pairs of feature points are required. For N>4 pairs, we use a least-squares method to find the best-fitting transform matrix. To improve the mapping precision, an algorithm similar to RANSAC is applied to remove outliers [10]. With $\mathbf{H_{in}}$, for example in Figure 4, the pen pointing to the word "FACT" on the printout is interpreted as mouse pointing to the word in the corresponding digital version. This approach is immune to the above issues of boundary detection.

For the output transform $\mathbf{H_{out}}$, we take advantage of the hardware setting: The relative positions of the camera, projector and the table planes are fixed, and the table surface is assumed flat. Therefore there exists a fixed homographic transform $\mathbf{H_p}$ (Figure 4) between the camera and the projector reference frame. Thus, $\mathbf{H_{out}} = \mathbf{H_p}^{-1} * \mathbf{H_{in}}^{-1}$. We derive $\mathbf{H_p}$ with a simple one-time calibration: A special image with known pattern is projected to the table, and captured by the camera. By finding the feature correspondence between the projected and captured images (with N >=4 correspondence pairs), we obtain $\mathbf{H_p}$. Note $\mathbf{H_{out}}$ varies with different document pages (multiple document pages can be recognized in one video frame) and different positions of the same document in the camera view. The projection overlay on a specific page automatically follows the page in movement.

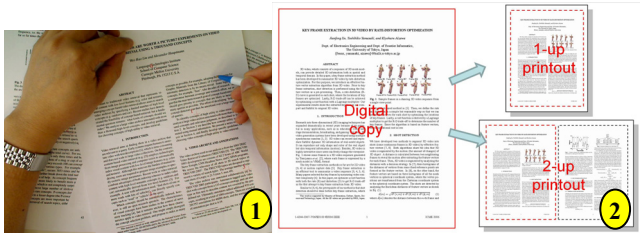

Figure 5. Failure sources of boundary based methods. (1) low contrast, hand occlusion and overlapping, (2) printing margins (the exact page boundaries are highlighted in red)

5.2 Determining the Input Focus Page

When m (m>1) pages are recognized from a camera image, FACT needs to decide which page the pen tip is interacting with. Basically, for each matched page P_i with transform H_i, i=1,…,m, we calculate the equivalent mouse pointer $T_i = H_i * T_p$ in that page. If T_k (k in [1,m]) is out of the digital boundary of page P_k, then P_k is discarded (Figure 6-1). For the special cases like T_p in an overlapping area of multiple pages (Figure 6-2), we assume the input focus page is at the top of the paper stack and it has some features within the overlapped area. Therefore, the page that has the most features in the overlapped area is taken as the input focus page. In the future, we can improve the accuracy by continuously tracking pages' topologic relation like Kim et al. does in [15].

5.3 Feedback for User Adaptation

The accuracy of the physical-digital interaction mapping may suffer from image noise, errors in document recognition, coordinate transform, and pen tip detection as well as fast paper movement. Although it is important to tune each of the above factors to improve the overall accuracy, we believe it is equally important to take into account user adaptation and provide appropriate feedback for users to help the system work correctly.

Based on this design principle, FACT provides two types of basic feedback to be projected on paper, namely the input focus and transform indicator. The input focus feedback includes the cursor and the highlighted region for the currently selected word (for text content), as shown in Figure 1-3. The positions of the cursor and the highlighted region might be different from those of the physical pen tip and its pointed word, but in our early deployment we found that users are able to quickly learn to adjust the pen tip location to get the intended document content selected.

The transform indicator aims to suggest to the user the current accuracy of the document recognition and coordinate transform. A possible solution is to project the *system-recognized* boundary of each digital page on its physical counterpart, so that good alignment of the projected and physical edges indicates the "ready" status of the system. Nevertheless, with the constrained projection area (e.g. quarter of a letter-size page) of the currently available hardware, FACT cannot project over whole pages. We indicate only the orientation of the input focus page. As Figure 7 illustrates, FACT first maps the digital page boundary into the projection reference frame with the derived transform matrix $\mathbf{H_{out}} = \mathbf{H_p}^{-1} * \mathbf{H_{in}}^{-1}$ (Figure 4), which encapsulates both the recognition accuracy in $\mathbf{H_{in}}$ and the projector calibration accuracy in $\mathbf{H_p}$. FACT then generates an L-shape widget in that reference frame, with the widget's two edges being parallel to the vertical and horizontal edges of the mapped page. To reduce the interference

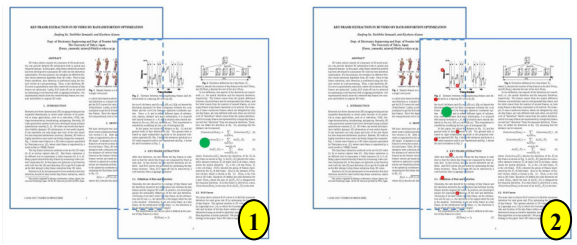

Figure 6. Determining the input focus with (1) basic case, (2) overlap. Pen tip is green and the feature points are red.

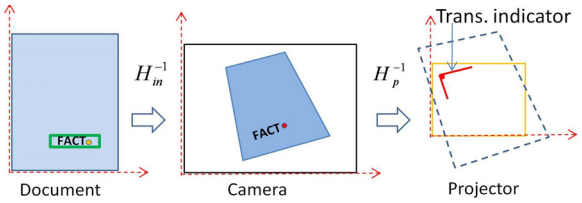

Figure 7. The digital page boundary is mapped to the projector reference frame and indicated by the L-widget

with other projected content, the widget is put at a corner. If the widget's edges are projected in parallel with the physical page's, the accuracy of the recognition and coordinate transform is good enough, so that the user can proceed to interact with the paper document (Figure 1-3,4). Otherwise, the user may withdraw his hands a little for less occlusion, or adjust the paper position and orientation to make the widget appear correctly.

In the same spirit, we can also present application specific feedback towards user adaption. For instance, in Keyword-Finding, some of the resulting occurrences may be out of the projection area (Figure 1-4). We borrow ideas from Halo[4], showing arrows around the projection boarders to indicate more occurrences in these directions (Figure 8). The user can then move the paper document in the reverse direction to reveal more.

6. CONTEXT-AWARE COMMAND ISSUING ON PAPER

The fine-grained physical-digital mapping enables a set of novel computer like interactions on paper. Considering the natural coupling of pen and paper, we borrow ideas from existing pen-based computer interfaces like Scriboli [12], allowing users to draw pen gestures on a paper document to select its fine-grained content, and to specify digital operations from a menu to be applied to the selected content.

6.1 Pen Gestures for Command Targets

FACT currently supports two input modes, namely the "navigation" and "gesture" modes. In the "navigation" mode, the system traces the pen tip and highlights the closest content like a word and symbol for feedback; in the "gesture" mode, the system interprets the pen stroke as a specific gesture and selects the document content based on the gesture type. As illustrated in Figure 9, *Pointer* is suitable for 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; *Bracket* and *Vertical bar* is used for 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. We currently use a 2-second timeout to switch between the two input modes, and users need to manually specify the type of a gesture before drawing it.

Being aware of the projection context, the feedback of the FACT

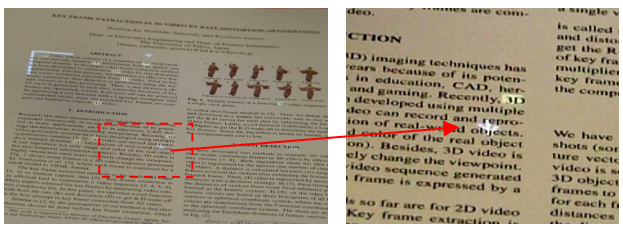

Figure 8. Arrows indicating the off-projection results.

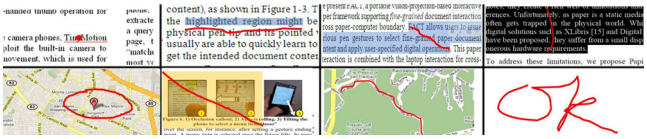

Figure 9. Examples of the FACT gestures (highlighted in red, from left to right and top to bottom): Pointer, Underline, Bracket, Vertical Bar, Lasso, Marquee, Path and Freeform.

gestures is specially designed to limit the possible interference with the original visual features of paper documents. Otherwise, the accuracy of the physical-digital interaction mapping could be compromised. First, we avoid rendering the gesture strokes if possible. For example, we only project feedback for the text selected by the Underline, Bracket, Vertical gestures, but do not render the raw gesture strokes. Second, we use thin straight line segments for projection (except the lasso and free-from gestures) as much as possible, because they generate fewer feature points than complex patterns. Third, we avoid highlighting large areas with solid bright colors, as the resulting glare may distort the original document visual features. Lastly, we only project the outermost contour of the selected content for feedback (Figure 10-1), instead of highlighting individual items separately like regular computer interfaces (Figure 10-2), to further reduce the undesired image features.

6.2 Menus for Command Actions

Upon selecting the command target, the user proceeds to select the desired action from a menu. FACT can directly project this action menu right next to the ending point of the gesture. This "in-place" menu saves movement of the pen, and makes the command gesture fluid and smooth. However, the projected menu may be occluded and thus made illegible by the underlying document content (Figure 11-1). This situation becomes even worse when the environment is bright and the projector luminance is limited (~12 lumens for the 3M MPro 110), which is quite common in realistic settings. Although some adaptive radiometric compensation methods [9] have been proposed to adjust the projection image to make the final projection appear almost the same as the original image, they do not work well for high-contrast and complex backgrounds like text and maps, especially with the low brightness portable projectors. In addition, this occlusion may also distort the original document visual features and affect the accuracy of physical-digital interaction mapping.

Inspired by the environment-aware projection display [5] that changes its interface based on projection geometry, we propose *Adaptive Menu Placement*: FACT projects the menu at a place with minimum occlusion, by searching for a closest projection region with the texture density below a threshold. We approximate the texture density with the number of feature points (found in the document recognition step) in a region (Figure 11-2). In this way, given an anchor point **P** (e.g. the last point of a gesture) and a minimum dimension of a menu **R** in the camera

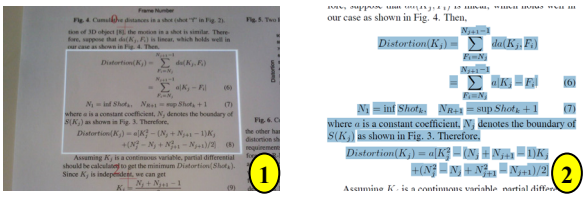

Figure 10. Feedback for selected content in (1) FACT and in (2) Adobe Acrobat

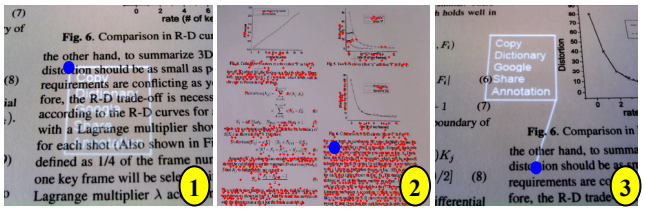

Figure 11. Adaptive menu placement. (1) A regular menu, (2) feature points (red dots) of the camera image. (3) An adaptive menu with less occlusion. (Anchor point P in blue)

reference frame, FACT enumerates, in the order of increasing distance, rectangles with dimension **R** around **P** in the feature point distribution histogram, and counts the feature points inside each rectangle. It stops when a region **M** with less feature points than the threshold is found. **M** is then mapped into the projector reference frame, and the menu is rendered within the resulting area with a callout to help users follow the menu from the anchor point (Figure 11-3). To accelerate counting, we borrow ideas from EMMs [20], using the integral image of the feature point distribution histogram to count the points in any rectangular region with only two subtraction and one addition operations.

In cases where there is no good placement in the projection area, FACT shows the command action menu on the laptop screen. Although away from the command targets on paper, the screen menu does not increase the eye-focus switching much, because after selecting command targets on paper, the user often needs to turn to the computer anyway to view the result of the executed command, such as the URL list of a keyword search on Web.

7. CROSS-MEDIA INTERACTION

The fine-grained paper interaction can be combined with laptop interaction. This feature is very useful to multi-document and multi-view manipulation spanning paper and a computer. As found in previous studies on knowledge workers' manipulation of documents [1], people spend almost half of the time working on multiple documents, for referring, comparing, collating, summarizing and so on. For these interactions, paper and the laptop can support each other. On one hand, the laptop screen has limited size, which makes the multi-document interactions inconvenient. Paper documents effectively extend the screen with more flexibility in spatial arrangement. On the other hand, paper is weaker in rendering dynamic information, which can be complemented by the laptop screen.

Although the basic concept of paper-computer federation has been proposed for years [14, 16, 29], FACT pushes the envelope by achieving computer-like fine-grained paper interaction and thus putting paper and the laptop on a more equal footing for novel cross-media interaction. Therefore FACT can support various applications in information transfer, association, sharing and synchronous navigation across the paper-computer boundary.

7.1 Information Transfer and Association

FACT eases quoting text, images or graphics with contextual information from paper. For instance, to get the explanation for an unfamiliar Japanese word "富士" on paper, the user simply points a pen to the word, and chooses a command "Web Dictionary." In response, FACT forwards the selected text to the laptop and performs a web search on the screen. This visual search is especially useful if the user does not know Japanese and therefore cannot manually transcribe the word through typing or speech.

Similarly the user can easily select a picture on paper and then copy it into a Word document on the computer (Figure 12-1,2). Instead of copying the picture directly from the low quality camera images, FACT copies the same region from the linked high resolution document image retrieved from the metadata database. This copied image is immune to bad lighting conditions, distortion and limited resolution.

The information transfer can be in the reverse direction. Multimedia annotations created on the laptop can be attached to a specific word, phrase, paragraph or arbitrary shaped region. An annotation is indicated by an anchor point and an icon in the nearby margin on paper (Figure 12-3). A pointer gesture on the icon starts rendering the annotation on the laptop. Note the annotated document can be printed again and its annotations can be retrieved via other user interfaces such as a mobile phone that recognizes the paper document (e.g. PACER [18]).

Users can use FACT to link two document segments across the paper-computer boundary. For example, in the above translation scenario, after browsing the results on the laptop, the user can choose the best web page to link to the original Japanese word on paper. Later, she can quickly revisit the translation. In the picture copying scenario, FACT can automatically create a hyperlink in the Word document linking the copied image to the context page and document (in digital form). These cross-page/document associations facilitate organization and navigation of the increasing personalized document information.

7.2 Information Sharing

FACT can be integrated with other computer communication tools to help sharing paper information with remote co-workers, friends or family members. For example in the scenario of trip planning, a user reviews a large paper map of a national park, and chats with a remote friend via IM on her laptop. The user lassos a region in the map and asks "what's cool about this area?" The high definition digital version of the map and the lasso mark are both sent to the remote friend. In the meantime, the rich context information is automatically shared, so that the friend can easily access the park's online brochure and answers "hi, look at page 11 of this brochure, you will find the top-rated spots in that area!" For intuitive discussion, the remote user can mark on the digital map, and the mark is sent back and projected on the paper map.

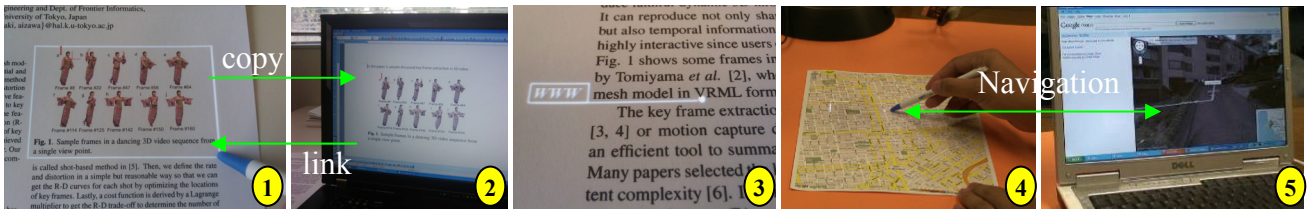

Figure 12. Cross-media interaction. (1) (2) Copying from paper to the laptop, (3) A Wikipedia page annotation for a word "VRML", (4)(5) Synchronous map navigation on a map and a laptop.

7.3 Synchronous Multi-view Navigation

FACT enables users to synchronously navigate different views of the same compound document. For example, paper maps serve large, robust, high quality displays, but they lack dynamic information such as street views and dynamic traffic information. Using FACT, a user can point to an *arbitrary* place on a map to retrieve and navigate the corresponding Google Street View on the laptop (Figure 12-4,5). This is different from existing systems like Map Torchlight [25], which are based on *pre-defined* hotspots. When the user changes the place, the street view is updated automatically. The user can also "drive" in the street view with the current location being highlighted on the map.

8. PERFORMANCE EVALUATION

FACT's fine-grained interaction builds on the accurate document recognition and coordinate transform, which depend on many factors such as image resolution, camera-paper distance, lighting conditions, hand occlusion, annotation and projection interference. To better understand these factors and evaluate the feasibility of the system, we conducted a set of experiments to test the accuracy with different settings of these factors.

8.1 Testing Scenarios

We focus on the performance of recognizing a single non-moving paper document page on a flat table. This is based on our observation that, when people mark on a paper sheet with a pen, the paper is usually held by one hand and remains still on the table. The paper movement normally occurs during page navigation and spatial arrangement, seldom during a fine-grained content operation within a page. Therefore, the poor accuracy of recognition and coordinate transform errors caused by fast moving paper do not really matter for our current applications. Second, we assume the paper is put on a flat table during pen interaction. Although it is possible for a user to hold a paper sheet in air with one hand and draw finger gestures with another hand, we leave this for future investigation. Third, we only tested camera images with just one paper document page, since FACT uses *local* image features and thus the performance on a single page is presumably not affected by additional pages.

We consider two scenarios. 1) *Baseline* has only paper documents in the camera view, and mainly examines the impact of image resolutions. It reflects the basic performance of the physical-digital mapping. 2) *Hand-Occlusion* tests the performance with different levels of hand occlusion over paper documents. Hand occlusion is a major factor reducing the document visual features in the camera images.

8.2 Settings

To collect testing images, we used an interface similar to that in Figure 1, except a desktop PC was used for easy debugging. We utilized a Logitech QuickCam Pro Webcam and a 3M MPro 110 portable digital projector. They connected to the PC via a USB and a VGA port respectively, with the projector as the secondary display of the PC. The camera-projector unit was attached to a lamp stand, and was about 25 cm high above the table. At this distance, the camera roughly covered two letter-size paper sheets, and the projection area was about 13cm x 9.5 cm. The recognition and transform tests were performed on another PC with Intel Pentium (R) III Xeon quad-core 2.82GHz CPU and 4GB RAM.

For the document image database (i.e. training pages), we randomly chose 100 articles (400 pages total) from the ICME 2006 proceedings. For each page, we generated two JPEG images of dimensions 306x396 and 612x792 (hereafter called "small training" and "large training" respectively), but at any given time there was only one version indexed in the database. We believe this database size is reasonable for the working document set of a normal knowledge worker. We do not aim at web-scale image search algorithms, but focus on the interaction techniques for a personal document collection.

For the testing images, we randomly chose 29 documents (116 pages total) from the database, and rendered them into 5100 x 6600 JPEG images (in 600 dpi). Within each of these images, we programmatically selected four random check points and overlaid a cross centered at each check point (Figure 13-1). In the meantime, the document page ID and the check point locations were recorded as the ground truth.

The 116 registered pages were then printed on normal letter-size paper. For each of them, we then used the user interface to captured one video frame (960x720). In the captured images, we used mouse clicks to mark the check points (Figure 13-2). These mouse click locations were later mapped into the digital document reference frame and compared against the ground truth. Note we did not compare the pen tip selected point with the ground truth, because pen tip detection may affect the overall accuracy and we wanted to focus on the document recognition and coordinate transform first.

8.3 Measurement

Recognition Accuracy is the percentage of the pages correctly recognized out of the 116 total testing images.

Transform Error is the average ratio of the point mapping error to the diagonal length of the ground truth documents (8341 pixels). The point mapping error is the distance between a ground truth check point and its corresponding recognized check point $\mathbf{H}_{in} * \mathbf{P}_{m}$ ($\mathbf{H}_{in}$ is the input transform matrix, and $\mathbf{P}_{m}$ is the mouse click location in the camera image). We did not test the accuracy of the output transform matrix $\mathbf{H}_{out}$, because it is also based on $\mathbf{H}_{in}$.

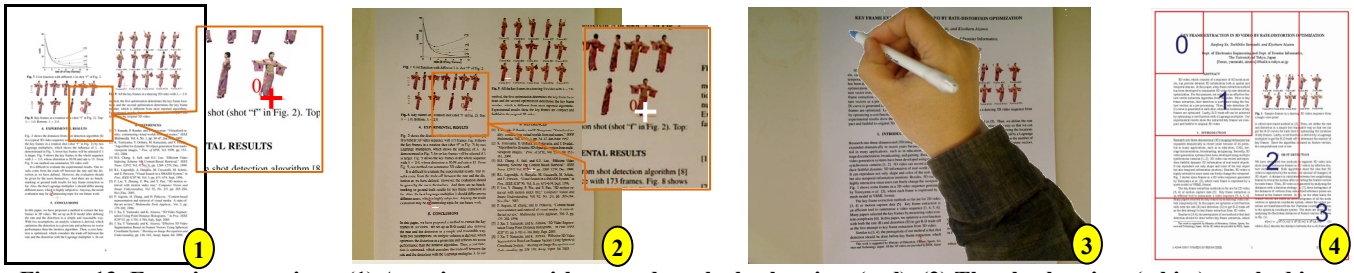

Figure 13. Experiment settings. (1) A testing page with ground truth check points (red). (2) The check points (white) marked in a camera image. (3) A synthesized camera image with occlusion level 0. (4) Four target regions with different occlusion levels.

Recognition Time is the time duration for recognizing a document image and deriving the coordinate transform matrices.

8.4 Baseline Test

In the baseline test, we revolved around the effect of the resolution of the training and testing document images. We tested the performance of the combination of two training resolutions, *low* (306x396) and *high* (612x792), and three testing resolutions *low* (320x240), *medium* (640x480) and *high* (960x720). We generated the low and medium resolution testing images by scaling down the captured 960x720 camera images.

We first examine *Recognition Accuracy*. As shown in Table 1, with the low training resolution, the recognition accuracy is >= 98.27% for all testing resolutions. Especially, the medium (640x480) testing resolution achieves perfect recognition. This result helps confirm the system feasibility.

Interestingly, the high training resolution leads to lower accuracy than the low training resolution. This seems contradictory to the intuition that larger images should produce more features and thus more matched features. Looking into more details, we found that the high resolution training images do contain more features at fine granularity, which however are less distinguishable than the features at coarser granularity and actually interfere with the distinguishable features. For instance, with the low training resolution, one of the testing images has 76 out of 156 feature points matched with the top candidate and 9 with the second one; with the high training resolution, the same testing image has only 7 and 6 matched features with the top two candidates respectively.

We further check *Transform Error* of those correctly recognized frames. The result is very encouraging, especially in the case of the (low training, middle/high testing) resolution combinations, of which the error range is [0.29%, 1.85%] and the average error is <= 0.85% (Table 1). In other words, the average error is within one line for a regular 60-line ACM Multimedia article. And the error can be further mitigated by other mechanisms such as user adaptation feedback. We believe this result confirms the feasibility of fine-grained interaction of FACT.

In terms of *Recognition Time*, with the low training resolution, the *low* testing images cost 433 ms, *medium* 2311 ms, and *high* 3212 ms on average. Taking into account the recognition and transform accuracy and time consumption, we believe the combination of 306x396 training resolution and 640x480 camera resolution is the best one among all the six configurations. Although the processing is not real time, our early deployment shows that it does not prevent user interaction much, because the paper usually remains static during within page fine-grained interaction and the pen tip detection is actually performed in real-time.

8.5 Hand Occlusion

To be efficient, we did not collect separate testing images with hand occlusion, but simulated them by programmatically overlaying a hand on top of the baseline images (Figure 13-3). To examine the effect of different levels of occlusion, we divided a page into four non-overlapping target regions (Figure 13-4), randomly chose a point in each region and aligned the pen tip with the point. For each baseline testing image, we generated four hand-occlusion images with occlusion levels 0~3 (0 for most occlusion, corresponding to the target region 0). In total, we tested 116*4 = 464 hand-occlusion images. We used the low training resolution (306x392) and medium testing resolution (640x480).

Training \ Testing		320 x 240	640 x 480	960 x 720
306 x 396	Recognition	98.27%	100.00%	99.14%
	Transform	19.81%	0.85%	0.68%
612 x 792	Recognition	70.69 %	92.24%	95.69%
	Transform	67.16%	6.62%	7.49%

Table 1. The baseline performance. The setting in grey is the best.

Accuracy \ Occlusion	0	1	2	3
Recognition	95.69%	98.28%	97.41%	100%
Transform	2.15%	2.19%	0.98%	0.91%

Table 2. The hand-occlusion performance

As shown in Table 2, both the recognition and coordinate transform accuracy remain high even with severe hand occlusion. For example, in the case of occlusion level 0, although almost half of the pages were occluded, FACT still successfully identified 111 out of 116 pages with average transform error 2.15%. This result confirms the robustness and precision of our content based approach, and suggests that hand-occlusion is not a blocking issue for the FACT interactions.

9. CONCLUSIONS AND FUTURE WORK

We presented FACT, a portable vision-projection-based interactive paper system supporting *fine-grained* document interaction across the paper-computer boundary. FACT allows users to issue various pen gestures to select fine-grained paper document content and apply user-specified digital operations. This paper interaction is combined with the laptop interaction for cross-media operations, such as information transfer, association, sharing and synchronous navigation. We discuss the design challenges for such a system, and propose content based physical-digital interaction mapping and context aware gesture-based command issuing. We also report an experiment, showing FACT can achieve 100% document recognition accuracy and less than 1% transform error with a 400-page database. This result confirms the feasibility and effectiveness of our design.

Future work involves user studies on the interface usability and the effectiveness of user adaptation designs. More investigation will be carried out for content based image recognition, such as how the database distinguishability changes with the amount and resolution of index images, as well as the impact of annotation and projection interference. We will continue to improve the system robustness and explore more interaction techniques such as paper interaction in 3D space and handheld interfaces. Moreover, we will make the camera-projector unit more compact by removing the lamp stand and using mirrors to extend the optical path length just like what MouseLight does [28]. Finally, we will identify some key application areas such as remote collaboration and architecture design to be focused on, and deploy and test the applications in realistic use scenarios.

10. ACKNOWLEDGEMENT

We thank Lynn Wilcox and anonymous reviewers for their suggestions and constructive comments. The research work was fully supported by FX Palo Alto Laboratory (FXPAL). Hao Tang worked on this project during his summer internship at FXPAL.

11. REFERENCES

[1] Adler, A., A. Gujar, L.B. Harrison, K. O'Hara, and A. Sellen. A diary study of work-related reading: design implications for digital reading devices. *Proceedings of CHI'98*, pp. 241-248.

[2] Anoto, http://www.anoto.com.

[3] 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.

[4] Baudisch, P. and R. Rosenholtz. Halo: a technique for visualizing off-screen objects. *Proceedings of CHI'03*, pp. 481-488.

[5] Cotting, D. and M. Gross. Interactive environment-aware display bubbles. *Proceedings of UIST'06*, pp. 245-254.

[6] Do-Lenh, S., F. Kaplan, A. Sharma, and P. Dillenbourg. Multi-finger interactions with papers on augmented tabletops. *Proceedings of TEI'09*, pp. 267-274.

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

[8] Everitt, K.M., M.R. Morris, A.J.B. Brush, and A.D. Wilson. DocuDesk: An interactive surface for creating and rehydrating many-to-many linkages among paper and digital documents. *Proceedings of IEEE TABLETOP'08*, pp. 25-28.

[9] Grundh¨ofer, A. and O. Bimber, Real-Time Adaptive Radiometric Compensation. IEEE Transactions on Visualization and Computer Graphics, 2008. 14(1): p. 97-108.

[10] 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.

[11] Hecht, D.L. Embedded Data Glyph Technology for Hardcopy Digital Documents. *Proceedings of SPIE Color Hard Copy and Graphic Arts III*, pp. 341 - 352.

[12] 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.

[13] Holman, D., R. Vertegaal, M. Altosaar, N. Troje, and D. Johns. Paper windows: interaction techniques for digital paper. *Proceedings of CHI'05*, pp. 591-599.

[14] Kane, S.K., D. Avrahami, J.O. Wobbrock, B. Harrison, A.D. Rea, M. Philipose, and A. LaMarca. Bonfire: a nomadic system for hybrid laptop-tabletop interaction. *Proceedings of UIST'09*, pp. 129-138.

[15] 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.

[16] Koike, H., Y. Sato, and Y. Kobayashi, Integrating paper and digital information on EnhancedDesk: a method for realtime finger tracking on an augmented desk system. ACM Transactions on Computer-Human Interaction, 2001. 8(4): p. 307 - 322.

[17] 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.

[18] Liao, C., Q. Liu, and B. Liew. PACER: A cameraphone-based paper interface for fine-grained and flexible interaction with documents. *Proceedings of CHI'10*, pp. 2441-2450.

[19] 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. 386-389

[20] Liu, Q., C. Liao, L. Wilcox, A. Dunnigan, and B. Liew. Embedded Media Markers: Marks on Paper that Signify Associated Media. *Proceedings of IUI'10*, pp. 149-158.

[21] Lowe, D.G., Distinctive image features from scale-invariant keypoints. International Journal of Computer Vision, 2004. 60(2): p. 91-110.

[22] Mistry, P., P. Maes, and L. Chang. WUW - wear Ur world: a wearable gestural interface. *Proceedings of CHI'09 (extended abstracts)*, pp. 4111-4116.

[23] Newman, W., C. Dance, A. Taylor, S. Taylor, M. Taylor, and T. Aldhous. CamWorks: A Video-based Tool for Efficient Capture from Paper Source Documents. *Proceedings of IEEE Multimedia System'99*, pp. 647-653.

[24] Rekimoto, J. and M. Saitoh. Augmented surfaces: a spatially continuous work space for hybrid computing environments. *Proceedings of CHI'99*, pp. 378 - 385.

[25] Schöning, J., M. Rohs, S. Kratz, M. Löchtefeld, and A. Krüger. Map torchlight: a mobile augmented reality camera projector unit. *Proceedings of CHI'09*, pp. 3841-3846.

[26] Sellen, A.J. and R.H.R. Harper, *The Myth of the Paperless Office*. 1st ed. 2001: MIT press.

[27] Song, H., T. Grossman, G. Fitzmaurice, F. Guimbretiere, A. Khan, R. Attar, and G. Kurtenbach. PenLight: combining a mobile projector and a digital pen for dynamic visual overlay. *Proceedings of CHI'09*, pp. 143-152.

[28] Song, H., Francois Guimbretiere, Tovi Grossman, and G. Fitzmaurice. MouseLight: Bimanual Interactions on Digital Paper Using a Pen and a Spatially-aware Mobile Projector. *Proceedings of CHI'10*, pp. 2451-2460.

[29] Wellner, P., Interacting with paper on the DigitalDesk. Communications of the ACM, 1993. 36(7): p. 87 - 96.

[30] Wilson, A.D. PlayAnywhere: a compact interactive tabletop projection-vision system. *Proceedings of UIST'05*, pp. 83-92.