MET: Media Embedded Target for Connecting Paper to Digital Media

Qiong Liu[1], Andreas Girgensohn[1], Lynn Wilcox[1], Frank Shipman[2], Tony Dunnigan[1]

[1] *Fuji-Xerox Palo Alto Laboratory, Palo Alto, California, U.S.A.*

[2] *Department of Computer Science, Texas A&M University, College Station, Texas*

[1] {liu, andreasg, wilcox, tonyd}@fxpal.com , [2] fmshipman@gmail.com

ABSTRACT

Media Embedded Target, or MET, is an iconic mark printed in a blank margin of a page that indicates a media link is associated with a nearby region of the page. It guides the user to capture the region and thus retrieve the associated link through visual search within indexed content. The target also serves to separate page regions with media links from other regions of the page. The capture application on the cell phone displays a sight having the same shape as the target near the edge of a camera-view display. The user moves the phone to align the sight with the target printed on the page. Once the system detects correct sight-target alignment, the region in the camera view is captured and sent to the recognition engine which identifies the image and causes the associated media to be displayed on the phone. Since target and sight alignment defines a capture region, this approach saves storage by only indexing visual features in the predefined capture region, rather than indexing the entire page. Target-sight alignment assures that the indexed region is fully captured. We compare the use of MET for guiding capture with two standard methods: one that uses a logo to indicate that media content is available and text to define the capture region and another that explicitly indicates the capture region using a visible boundary mark.

Keywords

Cross-media interaction, Paper interface, augmented paper, barcode, camera phone, document.

INTRODUCTION

Applications for linking paper to digital media are increasingly common. Although QR codes can provide this functionality, they take up space on the printed material and can disrupt the aesthetic design of the page [7]. More recent techniques use features of the printed image for identification and linking [5]. The user captures a region of the page using a smart phone application and a recognition engine matches the image with known images in its index. If a match is found, media associated with the matched image is displayed on the phone.

(a) (b)

Figure 1. Logo on magazine cover (a) and text in (b) indicate capture regions.

(a) (b) (c)

Figure 2. Different marker designs, EMM (a, b) and MET (c).

A basic problem with image matching systems is indicating to the user that there is linked content. Typically, a logo is printed on the paper to indicate to the user that there is media linked to the page. An example from the cover of Computer magazine is shown in Figure 1a). Features of the entire page are indexed, so that the user needs only capture a significant part of the page for accurate recognition. When multiple links on a page are desired, there are two approaches. One is to index the entire page, present the user with all available links, and allow the user to select the desired link. A preferable option is to index sub-regions of the page separately. This allows links to specific objects on the page. However, there must be a way to indicate to the user which regions have linked content. Current systems typically use the logo, with additional text to explain which regions to capture. An example from Time magazine is shown in Figure 1b).

An alternate solution is provided by EMM [23]. An EMM is a semi-transparent marker that is overlaid on the region of the page to be captured. It contains an iconic symbol that indicates the media type of the link (e.g. video, web). Although this solution clearly indicates to the user the region to be captured, in practice publishers and designers did not like the mark being placed on top of their content. A subsequent attempt replaced the semi-transparent overlay with corner brackets, but even this caused problems when the brackets had to be placed on top of content to specify the region. Examples of EMM are shown in Figure a), b).

In this paper, we introduce Media Embedded Target (MET) to indicate a media link and to provide capture guidance for the specified region of the page with minimal disturbance to the content on the page (see Figure c). A MET usage example is shown in Figure . Unlike Embedded Media Marker (EMM) [23], which places a semi-transparent marker around the entire region to be captured, the MET approach uses a small icon in a blank margin near the region to be indexed. To

Figure 3. Use a MET to guide big paper region capture by aligning the target with a small on-screen sight.

use a MET, users align the target on paper with a sight of the same shape on the cell phone display. This alignment determines the distance of the cell phone from the page and its orientation, thus defining a capture region on the page. In appearance, the target is similar to logos used by other systems to indicate linked content, but in addition the system defines an interaction. The user moves the cell phone to align the sight on its display with the target, and when the system detects correct alignment the image is captured by the phone's camera. This game-like interaction guides users to linked content and specifies the capture region. Figure illustrates use of the MET marker as a target for the sight on the cell phone display.

When authoring the content and creating links, the designer has in mind an object or region on the page that correspond to the link. However, recognition is based on image features and the object or region itself may not have sufficient features. Thus for recognition a larger portion of the page may need to be captured. The MET authoring tool helps the user place the MET on the page with the proper size and orientation to capture a recognizable region of the page.

In this paper, we explore paper-to-digital link designs that balance the trade-offs of aesthetic impact and usability. We compare the MET technology with two other approaches to indicate paper-to-digital links; overlays of the linked region and icons plus text describing the link in the margins. We perform a within-subject assessment of how these three techniques compare in terms of the accuracy of link region capture, the time it takes to acquire the link, and user-experience. Our results indicate clear trade-offs among the three approaches.

The rest of the paper is organized as follows. In Section 2 we discuss related work. In Section 3, we will talk about MET design. Section 4 covers target-sight alignment and Section 5 describes capture region identification. In Section 6, we describe an algorithm for semi-automatic arrangement of targets and give a comparison between EMM and MET based on their effectiveness and intrusiveness in Section 7.

Section 8 describes user experiences and is followed by possible MET extensions in Section 9. In Section 10, we provide some conclusions and discuss future work.

RELATED WORK

According to [21], there are seven approaches to identify a document region. The first approach is to print a barcode or QR code [36] in the region for identification. The second approach is to use micro-optical-patterns such as Dataglyph [11] in document regions for identification. The third approach is to modify document content to encode hidden information for identification [9]. The fourth approach is to index underlying paper fingerprint for document identification [6]. The fifth approach is to use OCR or character recognition outputs for identification [14][18]. The sixth approach is to index printed document content features and use these features to identify document [8][12] [22][24][25]. The seventh approach is to put an RFID in the document for identification [13][32].

Barcode and barcode like markers [33] are now well-known to provide additional information either by encoding the information directly or by providing document identification. The difficulty with this barcode approach is that barcode is opaque and therefore can only be printed in blank space in the document. If the barcode is small, users have to move their cameras very close to the barcode to get enough capture resolution for barcode decoding. When a camera gets very close to a paper surface, it may be hard to focus and may block necessary lighting. If the phone has sufficient camera resolution, it is not necessary to move the phone too close to the paper surface for barcode resolution. However, using a small portion of a high-resolution image to capture a barcode will make the barcode detecting process more difficult and thus reduce media link detection speed and detection rate. Moreover, a small barcode also makes it difficult to encode sufficient linking information. Additionally, since barcode does not provide media type information, users may confuse barcode-like media link with price tag and library management tag.

To overcome the barcode issues, researchers proposed semi-transparent barcode links [22] that can be overlaid on content and still can be recognized by barcode readers. Figure illustrates this kind of barcode. To motivate users for more multimedia explorations, envelops of those semi-transparent barcodes are also changed to reflect the linked multimedia type. Even though the semi-transparent barcode link overcomes many barcode issues, publishers' are not yet familiar with this technology and may not want to adopt it.

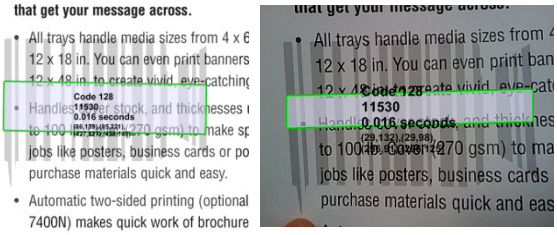


Figure 4. Semi-transparent barcode. Excerpt from [22]

Figure 5. Use text explanations (highlighted with red circles) to guide users' captures.

Besides the barcode approach, all other document region identification methods require some kind of user guidance for a proper capture. Early prototypes in this field explored Augmented Reality (AR) approaches for document capture guidance. However, AR technology alone does not provide any indication of the existence of multimedia links. Moreover, without any guidance on paper for a proper capture, many users may not know where and how to move a camera to get the AR overlay. This may cause users to miss links on a page. Since most current printed pages do not have multimedia links, helping users to easily identify linked regions and perform proper capture is critical important for reducing users' link-search frustration.

After realizing this issue, interactive paper prototype builders [4][17][20] started to use text explanations for page region capture guidance. These examples can be found in Figure . Even though text explanations are helpful, this approach also has some disadvantages. Firstly, text explanations require users to read and therefore are not very intuitive for regular users to follow. Secondly, text printing is not very attractive for getting users' attentions. Thirdly, text printing normally requires a large printing area that occupies precious space on magazines and newspapers. This extra space requirement makes it hard to insert these links at many places. Beside text explanations, people also explored small icons near a page corner [2][4][34] for indicating that a page has a link. However, these page marking icons do not provide camera capture guidance. On the other hand, from these text explanations, we see clear demand for giving users more clear guidance.

To avoid the need for text explanation and page marking, researchers [23] developed Embedded Media Markers (EMM) to guide page region capture. EMMs are nearly transparent iconic marks printed on paper publications that signify the existence of multimedia associated with that

part of the paper. It also provides associated multimedia type to motivate users' for further exploration. Additionally, it reveals a visible boundary that encloses all indexed visual features. With this visible boundary overlay on an original document, users only need to align camera-capture view with this boundary to ensure enough and just enough visual features for reliable paper region identification. On the other hand, by restricting users' capture within the EMM guided capture range, back-end machines do not need to index all visual features and thus demands much less CPU, memory, and disk resources. These resource reductions lead to cheaper, faster, and more accurate dynamic information to users. In this human-machine win-win interaction, the interaction medium, paper, is slightly modified. Even though EMM tried to minimize the hardcopy modification by using thin and semi-transparent markers, publishers were still not happy with the overlay.

Here we propose using MET as that shown in Figure . In this approach, small sights on the display region of a cellphone are used. By setting a fixed geometric relation between sight overlays and the overall field-of-view, these small sight overlays may be used to guide a much larger field-of-view capture. Because of this, guidance for a large document capture region can be achieved with a marker that is much smaller than the region and thus much smaller than an EMM marker.

MET DESIGN

Figure 6. Examples of different types of sights/METs.

A Media Embedded Target, or MET is an iconic mark printed in a blank margin of a page near the content to be indexed. It indicates the existence of a media link and serves to guide the user to capture the appropriate region of the document. This guidance is achieved by displaying a sight on the cellphone camera display. The user aligns the sight on the cell phone with the marker on the page, thus defining a capture region. Different capture regions are defined by varying the orientation, location and size of the marker on the page. Figure gives several examples of MET indicating a video link, a web page, an audio link, and a forum.

To define a unique planar relation between the camera view plane and the paper plane, the MET mark and the sight must contain at least three distinctive non-collinear points for alignment. This requirement is satisfied by designing the MET mark and the sight to have the same non-radial-symmetric shape. The shape outline of the target on paper and the outline of the sight on the

cell phone display are used for alignment. The user aligns the sight on the cell phone display with the target on the page by adjusting the cellphone pose or the paper pose to match the camera preview on the phone. This in turn defines the region of the page the cell phone will capture. A symmetrical shape such as a circle cannot uniquely define a page region since a rotation of the phone in a plane parallel to the page will maintain the alignment between the sight and the target but will capture a different region. Similarly, a linear shape will not uniquely define a capture region because rotations out of parallel will result in different regions.

The size of the capture region is determined by the size of the target, which determines the distance of the camera from the paper. Since the sight on the cell phone display is a fixed size, the larger the target on paper, the farther away the camera needs to be to align the target with the sight, and thus the larger the captured region.

The design of the MET is influenced by publishers' current practice to use icons, such as scissors or logos in a blank margin of a page to attract attention or promote a brand. Recently, this marker style has been used for indicating media links [17]. However, these logo-like markers do not explicitly indicate the region to be captured, and are often accompanied by a text explanation describing the region to capture and the link.

We design the MET to be a directional icon, pointing towards the content the link is associated with. Similarly the sight icon on the cell phone points to the center of the capture region on the phone. Alignment of the target with the sight on the cell phone depends only on the shape of the target. Thus, the interior of the target can be used to define branding, or to indicate the type of media associated with the MET, as shown in Figure 6.

TARGET-SIGHT ALIGNMENT DETECTION

In order to retrieve the media link associated with a MET, the user aims a cell phone at the page and tries to align the sight on the phone display with the target on the page. When alignment is achieved, the phone will capture the image, send it to a server which matches the image with the indexed collection. If a match is found, a URL for the corresponding media content is sent back to the phone for playback. For the alignment, we consider two approaches, one manual and one automatic. In the manual approach, the user aligns the sight and target, then presses a button on the cellphone screen to capture the image. However, our experience indicates that users do not like to press a button to capture an image for this type of application.

The automatic alignment process can make the interaction more game-like. The algorithm continuously analyzes the camera preview image. Its feedback is provided on the phone indicating when alignment has been achieved. This is done by changing the color of the sight, starting with red for no alignment, to yellow for partial alignment, and finally green for good. Figure illustrates the alignment detection approach. The system uses the sight drawn on the phone display as the starting point for an inner mask. To accept close matches, it dilates that mask by 3% of the mask size (e.g., 3 pixels for a 100x100 mask). It than creates an outer mask by scaling up the mask and by removing pixels from the outer mask that are also in the inner mask. The bounding box of the outer mask in the captured image is processed with a Canny edge detector to extract edges. If a MET appears in this processed image, the MET contour will be extracted together with other edges. These extracted edges are then filtered with both the inner and outer masks. If sufficient edge pixels are detected by the inner mask and only few edge pixels are detected by the outer mask, the system will consider the target and sight are aligned. The outer mask is needed to avoid false positives in regions with many edges such as text blocks.

This simple mask-based alignment detection approach can greatly reduce the number of images sent back to server and thus save bandwidth, computation, as well as users' time.

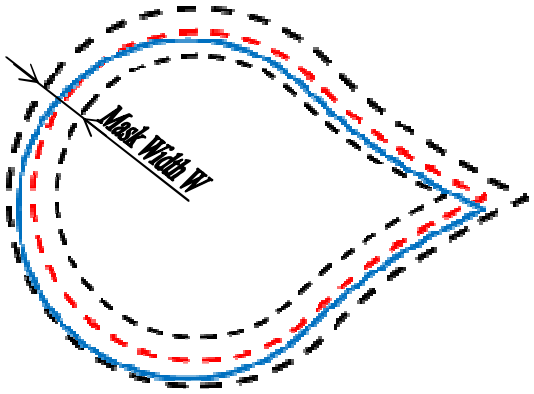


Figure 7. sight-MET alignment detection. All dashed lines are virtual/real traces on a smartphone display. The red dashed line in the middle is the sight skeleton. The two black dashed lines illustrate the boundary of the sight mask. The solid line illustrates the edge detection result on paper surface. When sufficient edge pixels are detected within the mask, the sight and MET are considered aligned.

CAPTURE REGION IDENTIFICATION

Once the image is captured it is sent to the index server. There the system extracts features and uses these features as queries to search candidate images in the indexed image set. In our implementation, we used SIFT-like proprietary image local features for the search. After the system finds matches between the query features and indexed features, the system selects indexed images with sufficient matched features as candidate images and computes accurate

homography transformations[1] between the query image and every candidate indexed image with a RANSAC algorithm[2]. Candidates with too many outliers are excluded. Following this geometric verification process, all candidate images will be re-ranked according to the number of filtered matches and indexed images with sufficient matches will be selected as candidates for next round. In the last identification step, the system measures the offsets between the query image center and each candidate image center in the indexed image coordinate space. The candidate with the smallest offset will be identified as the matched indexed region.

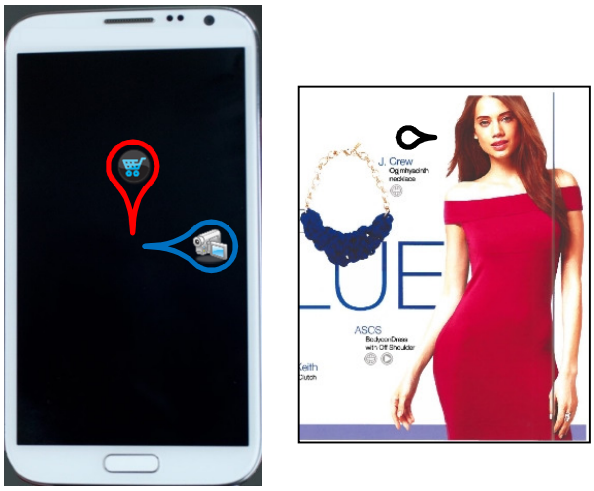
In this indexed image identification process, the number of matches is used to eliminate unrelated regions and select a small number of candidates. The geometric verification process further reduces the possibility of false matches, and the center-offset-based re-ranking is used to identify a user's focus. That is different from the identification process of a captured EMM region [23], which is only based on the number of matched features. With the MET identification process, indexed regions associated with different METs can overlap. This property and the compact size of MET allows publishers to place many more multimedia links on paper.

[1] http://en.wikipedia.org/wiki/Homography

[2] http://en.wikipedia.org/wiki/Ransac

EXTENSIONS

Because geometric verification is used during target recognition, the system can easily figure out the center offset between the indexed image and the captured image in the indexed image coordinate space. It can also figure out the scale difference and rotation difference between the captured image and indexed image. These extra detections can be used to activate more functions on the phone. For example, we can use multiple sights to get different content from same target by detect phone pose. More specifically, assume we arrange two sights on a smartphone screen as that shown in Figure (a) (The sights are enlarged for illustration purpose). The red sight pointing downward is for shopping and blue sight pointing



(a) (b)

Figure 8. (a) sights arrangement at different angles on a phone display. The sights are purposely enlarged to illustrate media icons inside the sights. Real sights are much smaller on a phone screen. (b) MET print example on paper.

from right to left is for video. With this configuration on a smartphone, if a user see a MET on a catalog page as that shown in Figure 8(b), the user can activate a video show of the model by aligning the blue sight with the target. Similarly, the user can activate the shopping webpage by aligning the red sight with the target. This kind of capture-pose-based extension can be done in many different ways. For example, depending on application, the designer may also choose to put fewer sights on the phone display and more targets on the paper. Even though we showed media type icons in sights in this example, designers may also choose to put the media type icon in target depending on the task. The sights arrangement on phone display may also be changed. In Figure 8 (a), we arrange two sights in 90 degree angle. If necessary, the angle for separating sights can be 45 degree etc. With this kind of arrangement, a user can use the cellphone as 'knob' on a hardcopy, and activate different functions associated with a region by rotating the 'knob'.

PLACEMENT OF MET

Although the page designer will have clear ideas on where to place the MET and what part of the page is to be indexed, it is difficult to manually determine the size and orientation of the target that will result in the desired capture region. There are also constraints from the recognition engine, which is based on image local features. In order to accurately recognize a region of the page, there need to be sufficient visual features in the region. For example, a region consisting of a red circle on a white background cannot be accurately identified using common image features such as SIFT[25], SURF[1]. Thus we provide an interactive tool for MET placement that aids the designer in determining the size and orientation of the target while satisfying these constraints.

Geometric Relations between Predefined Region and MET

The geometric relationship between a predefined region and a MET is defined by the geometric relationship between the target, the corresponding sight, and the camera view. To make our algorithm work for camera views with different aspect ratios, we only use the square center of the camera view and pick the circle inside it. We then put a sight very close to the circle parameter pointing toward the center of the camera view as shown in Figure .

In Figure 9, we denote the radius of a predefined region with R, the center of the predefined region on paper as (X,Y), and the maximum MET dimension as D. With all these parameters defined, we can develop algorithms to assist

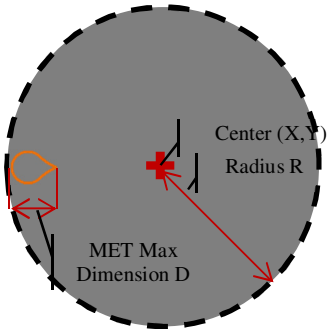


Figure 9. Relationship between a predefined region and a

document editors on adding predefined regions and METs. More specifically, an editor can define an interesting spot on a digital document page via a mouse click. Then, our algorithm will search nearby blank space for a MET. An ideal MET will be placed in blank space, pointing toward the editor defined interesting spot, with a proper size to define a predefined region which encloses enough and just-enough machine recognizable features for indexing. With these predefined region reduced index feature set, the recognition server will use much less computation resources for the recognition task. Moreover, since the indexed feature number is reduced, mismatched features will also be reduced. This will be helpful for increasing the recognition rate. Additionally, with a smaller indexed set, recognition will be faster than using a larger feature index set. All these improvements are beneficial to end user experience as well as server constructions.

Blank Space Search

To place a MET in blank space close to a signified spot, the system first has to find blank space that can hold a target. This can be achieved by putting a target at a simulated position and check if the target 'collides' with original contents. Letting human users to try different location with various size target shapes is tedious. Our system uses edge detection to tell if there is 'content' at a particular pixel. Thus, our system can find blank space that has no content within a simulated target shape. Since target shapes are normally artistic curves that are hard to be described with equations, checking the number of features in simulated target shapes for every position at various sizes on a page is time consuming. To solve this problem, we use a square whose side is larger than the low limit of the target maximum dimension size D and smaller than D's upper-limit to perform the check.

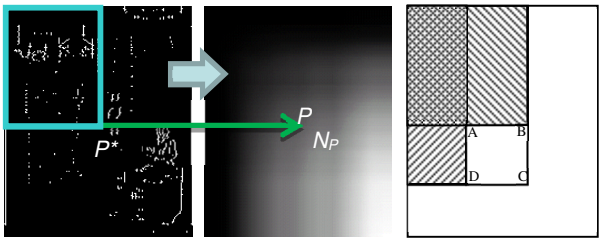


Figure 10. Compute an integral image map. Left: edge point distribution in an image. Middle: N_P equal to the number of edge points in the rectangular box in the left image. Right: Count edge-points in a square ABCD.

There are many ways to check the collision between a target square and contents. If the system counts the number of edge points within a target square for collision detection at each tested location in a brute force way, the computation cost will be in the order of (number-of-pixels in the whole image x number-of-pixels in the target square x number of different target size). To reduce the computation, we used an edge point integral map to perform the counting. Figure illustrates the computation of an integral image map. In this figure, the left image is an edge-point distribution map and the middle image is an edge-point integral map. In the edge-point integral map, there is a randomly selected point P with a value N_P. Corresponding to the point P in the edge-point integral map, there is a point P* in the edge-point distribution map. Even though P and P* are in different images, they have the same coordinates in these two images (have the same distances to the image-top-edge and image-left-edge). In the edge-point distribution map, there is also a rectangular box between the top-left corner and P*. The value N_P of the pixel P in the integral image map is computed by adding all pixel values in the rectangular box of the edge-point distribution map.

The right image of the Figure illustrates the edge point counting speedup with an integral image map. In this image, the estimated edge points NABCD in the square ABCD can be achieved with the following equation:

$$N_{ABCD} = N_C - N_B - N_D + N_A \qquad (1)$$

This equation allows the system to estimate the number of edge points in a target square with three operations. It is much cheaper in computation than going through all pixels in the bounding box.

Blank Space Tolerance to MET Size

With this integral map approach, we may check the maximum MET size that can be tolerated at every location. This maximum tolerance value can be got by trying various target square sizes. If the system tries all possible target sizes in pixels, it will be very time consuming. We designed the following binary search approach to roughly estimate the maximum square size.

```
currentside = (sidehigh + sidelow)/2;
while ((sidehigh-sidelow)>SMALLMARGIN)
    Get the number of edge points in the MET square;
    if (number of edge points  <= 0)
        sidelow = currentside;
```

else

sidehigh = currentside;

end

currentrside = (sidehigh + sidelow)/2;

end

if ((currentside-sidelow) < SMALLMARGIN)

currentside = 0;

end

After going through this process, the system will have an image with highest MET tolerance value for each pixel.

Search Predefined Region

Since the paper patch identification accuracy does not change much if the number of patch-covered feature points is over a certain threshold, the predefined region boundary-circle search goal

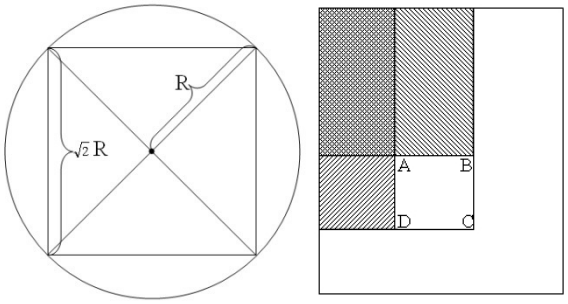


Figure 11. (left) Use the number of keypoints in the square to approximate the number of keypoints in the circle. (right) Use a cumulative keypoint distribution to compute the number of keypoints in a square. Excerpt from xxx et al. [23]

is to achieve good patch-identification accuracy with minimum cost of paper surface area. To get a small predefined region with sufficient keypoints, the algorithm will tend to move to a location with high keypoint density. The algorithm will count the number of keypoints inside a simulated predefined region. If the predefined region has more keypoints than needed, the algorithms will decrease the predefined region size for a new test. If the region has less keypoints than the required threshold, the algorithm will increase the predefined region for a new test. Assume a target is too small to have a large impact to keypoints from the original document, the algorithm may compute all keypoints at the beginning and use their locations for all keypoint counting within a simulated predefined region.

Beside keypoint re-computation, counting the number of keypoints in a circle is also time-consuming. If a page has n keypoints, the computational complexity for counting the number of keypoints in a circle is O(n). That is too much computation for estimating one predefined region parameter set. To overcome this issue, we use an approach that is similar to the keypoint counting process of an EMM region. More specifically, we approximate the number of keypoints in a circle with the number of keypoints in a square within the circle. The left image of Figure illustrates this approximation. The right image shows how to compute the number of keypoints in square ABCD. This is similar to the computation described in equation

1 except equation 1 is used to count the number of edge points while we use it to count the number of keypoints at this step.

Optimal MET Position and Size Search

Because the ratio between an MET maximum dimension size D and its corresponding predefined region radius R is fixed, a smaller predefined region needs a smaller MET, and a smaller MET has more flexibility to find a proper location in blank space found by the previous algorithm. On the other hand, small size may also bring MET disadvantage when a target is too small. Technically, when visual keypoints used in a match is too close to each other, equations used for solving the transformation matrix will be very similar. Thus small keypoint position errors may lead to big errors in transformation matrix. This is also true for a human vision system. To an extreme, when a target is too small, it will be hard for user to clearly see its shape and orientation and therefore hard to use it as a capture guidance. To overcome this disadvantage, each MET authoring system should set a minimum size limit for a target.

To achieve a balance between these two issues, we defined the ratio R/D equal to 10, and designed the following algorithm to find an optimal target position and size.

Compute all keypoint locations on the processed image
Get interesting spot point from an editor's mouse click
while(not all blank space pixels are tested)
    Pick a blank space pixel with D_{high} = *maximum MET tolerance value and* D_{low} = *minimum MET tolerance value;*
    $D = (D_{high} + D_{low})/2$;
    Connect the blank space pixel with the editor selected pixel with a straight line L;
    While $((D_{high}-D_{low})>SMALLMARGIN)$
        Find the predefined region center (X,Y) on L by enforcing the distance between (X,Y) and the blank space pixel equal to 9.5D;
        Check the number of keypoints in the predefined region;
        if maximum number of keypoints within predefined region > KEYNUMLOWLIMIT
            $D_{high} = D$;
        else
            $D_{low} = D$;
        end
        $D = (D_{high} + D_{low})/2$;
    end
    record the minimum D in D_{min} *and blank spot position in* (X_b, Y_b)
end

With this algorithm, the system can find the optimal target position (X_b, Yb) and optimal target size with D after the code is executed.

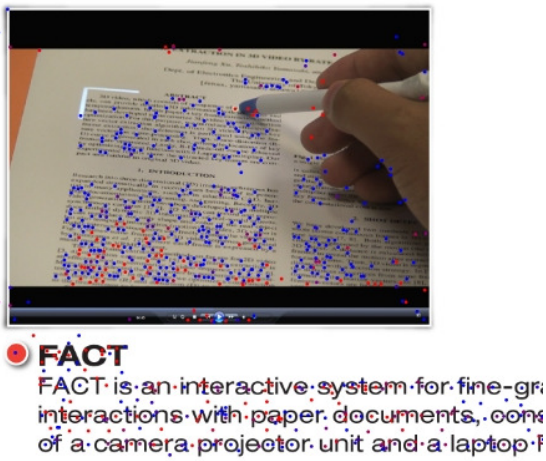


Figure 12. Visual feature density map

Alternate Placement Tool

Based on user experience, we also provided an alternate approach that gives the designer control of MET placement. In this case, the designer first places the target on a blank region of the page and points it towards the content to be indexed. The system then displays the resulting capture region and shows a density map of the visual features (see Figure). The system then counts the number of visual features in the capture region. If this exceeds a required threshold, the target placement is complete. If there are too few visual features in the region, the size of the MET is increased, thus increasing the capture region. The system again counts features in the capture region. The process continues until a capture region with sufficient number of features is found, or the target reaches a maximum size, indicating that this target placement and orientation are not possible. The user is then instructed to change the orientation or location of the target, using the feature density map as a guide.

MET INTRUSIVENESS EVALUATION

Since the work on EMM [23] reported several experiments on paper patch recognition accuracy (nearly 100% accuracy) and the target-sight alignment procedure is very similar to the EMM-camera-view alignment procedure. Thus, we are confident that the MET-SIGHT design will not affect the recognition accuracy. Therefore, we skip patch recognition accuracy assessment in this paper and work on evaluations that differentiate MET and EMM.

The motivations for MET were to enable links while reducing the aesthetic impact on the printed document and improving user experience. The circle of the original EMM had to partially cover the target region. While it was semi-transparent, this was still not acceptable to designers. The use of corners in a subsequent EMM design was intended to address that issue. However, designers still did not like to have corner markers in the middle of the page content. In contrast, the

MET is designed to go into the page margin and does not cover any content.

One way to measure aesthetic impact is to compare the size of the printed regions for the two techniques. To achieve the same functionality, the less occupied area the better. For example, we may use (# of pixels occupied by a SIGHT / # of pixels occupied by a maximum on-screen circle) as a measure. Since a SIGHT is approximately 1/10 of the camera view width, the area covered by MET will be about 1/100 of the viewing area. Thus, a MET will cover about 1/100 of the area required for an EMM, which reduces the aesthetic impact considerably.

USER EVALUATION

Normal interaction with paper containing links to digital content would entail having the user aim the camera at the page region with the link followed by the digital content being displayed on the phone. Although the goal for the user is to capture the link to get the linked content, there are some variables such as lighting conditions and camera auto focus that affect link identification. Thus we studied how well users could identify the page region containing the link and aim the

Figure 13. Designs for indicating links on paper: (top) arrow, (middle) circle, (bottom) text.

camera at it. We used the overlap between capture and target regions as a proxy for the likelihood of a positive match on the server. Rather than display the linked digital content, we showed a thumbs-up image, played a positive sound, and ended the session after the participant aimed the camera at the correct target for a few seconds.

Twelve employees of our workplace who had no connection to this or related research projects were recruited as participants. They included a mixture of managers, research staff, and administrative support personnel. The participants' ages ranged from about 30 to 60.

Fourteen documents (two for training and twelve for the study tasks) were identified with seven coming from a popular architectural magazine (more visually-oriented) and seven coming from ACM magazines, thus including more text and being more academically focused. Each of the fourteen pages was redesigned for each of the three techniques. Examples of links in the three conditions are seen in Figure 13.

Participant sessions were divided into three activities, one for each condition. Each activity included training on the use of the linking technique, the performance of four link traversals with that technique, and answering a questionnaire about the technique. The order of the techniques used was balanced to compensate for training and fatigue effects. Similarly, the assignment of documents to the three techniques and their order was balanced to remove any bias due to the inherent difficulty of the link identification and traversal tasks.

Training involved brief instruction in the use of the technique followed by two one-minute sessions with the two training documents. During the one-minute interactions with each training document, subjects could move the phone about to gain an understanding of the interface and feedback. Also during these interactions the images captured by the phone were constantly uploaded to the link server and when the subjects had correctly captured the link for the necessary period the success sound was played and the timer reset. This meant that all subjects successfully targeted the link on the two training pages numerous times. Subjects were thoroughly capable at the end of training. While this might have bored the participants, it did ensure that participants were well practiced in link acquisition in each technique before the data used for comparing the techniques was collected.

Multiple sources of data were captured for each interaction. Timing and accuracy data was collected through the logging of image match results returned by the link server. User experience data was collected through questionnaires completed by participants following their use of each of the three

techniques and notes taken by observers during the tasks.

Performance Measurements

For determining where users aimed on a page, we used a variant of SIFT features to index all the keypoints on each page. We uploaded captured images to the index server via a Wi-Fi connection as quickly as the server could process them (about two per second). For each image, the server determined the keypoints, matched them to the keypoints stored for the page, and determined the homography matrix for the mapping between matching keypoints. The inverted homography matrix was then used to map the square capture region to a quadrilateral on the page and the capture center to a point on the page. Some uploaded images contained insufficient keypoints to match against the stored keypoints, for example, because the auto-focus was still in progress or because the camera lens was covered. With the exception of one participant where many images were out-of-focus maybe due to shaky hands, images without matches only occurred infrequently.

Keypoint-based matching approaches succeed when enough matching keypoints are found to uniquely identify the target with high confidence. If only keypoints from the target region are stored on the server, a large overlap between target and capture regions is required. If keypoints from the whole page are stored on the server, one can project the center of the capture region onto the page even if there is little or no overlap between the target and capture regions. We used a measure relevant for each of those situations.

The first measure is the F-score, the harmonic mean of precision and recall, of the overlap between the target region on the page and the quadrilateral projected onto the page. The area of the overlap was determined with a polygon intersection algorithm. The target region was expanded to the smallest enclosing square centered on the target region. Precision was defined as the fraction of the projected quadrilateral area that was inside the target region. Recall was defined as the fraction of the target region that was inside the quadrilateral. We set the F-score threshold by finding the least square image used in the study (aspect ratio 5:3) and by determining the F-score of the inner square for the image under the assumption that some participants would use the inner instead of the outer square. This produced a threshold of 0.6.

The second measure is the distance between the center of the target region and the projected point of the center of the captured image. This measure would be relevant if the whole page is indexed on the server. This measure does not check if the captured region is much larger or smaller than the target region due to the phone being far away

from the page or close to it. This measure also does not reduce the score if the phone is not parallel to the paper. We used a threshold of 1.5 inches because that provides a clear indication of the aim for typical target regions with dimensions of three by three inches.

To determine if the aim of the study participants improved while the aimed the phone towards the target, we required that they kept the phone on target for several seconds. We used a sliding window of the last seven F-scores and required that five of those would be above the threshold. At that time, the task was ended automatically. After not meeting this criterion for 30 seconds, the task was ended, too. This happened only once in one of the tasks of the participant with the out-of-focus images. Even in that situation, the performance thresholds were exceeded after 11 seconds.

In addition to the F-score threshold, we also used a threshold for the distance between projected center and target center. For both measures we determined the time required for reaching the threshold for the first time.

Performance Results

For comparing the performance of the different conditions, we measured the time it took to reach either of the performance thresholds for the first time. Figure shows the average times for reaching those thresholds for each of the three

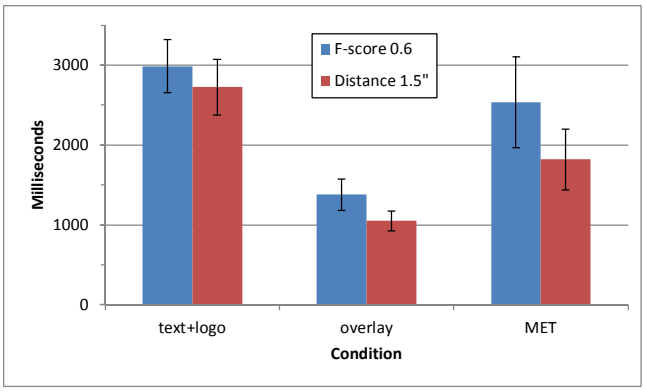


Figure 14. Average times for sufficient matches.

conditions. Average times ranged between one and three seconds across the conditions with a maximum time of 16 seconds. Note, the total task times were a little longer due to the sliding score window. An ANOVA test determined that the effect of the condition on the times was statistically significant for both measures: $F(11,2)=3.92$, $p=0.035$ for the F-score threshold and $F(11,2)=6.01$, $p=0.008$ for the distance-based threshold.

Condition 1 with text describing the capture region was the slowest. Extra time was needed to read and interpret the text. We noticed that some participants seemed to guess what the capture region should be and only quickly confirmed it by glancing at the text. Many of the pages used in the study only contained one prominent visual object such as an image or a chart making this guess possible. In a more realistic setting, Condition 1 may perform even worse.

Condition 2 with a circle indicating the capture region performed better than Condition 3 where participants had to align the sight to the arrow

printed on the page. If Condition 2 were acceptable to publishers, it would clearly be the best choice.

As discussed later, several participants commented that the task of lining up the sight was too difficult. From pilot tests, we determined that the size of the sight plays a large role in the difficulty of the alignment. However, it is constrained by the size of the target region and the available space in the page margin next to it. For the study, we made the width of the sight 12.5% of the width of the target region. As the top of Figure illustrates, that is about the largest size that can fit into the page margin. Based on the participant feedback described in the next section, we plan to explore alternatives that will make the alignment easier.

User Experience

Participants were asked to provide feedback about what they liked and did not like about each interaction method. Comments across all three conditions show that users appreciated that they did not have to manually indicate that the camera was in position to take the link. They also valued the audio feedback confirming a successful capture.

User comments indicated the mixed reactions to text and icons in the margins to indicate links. Four users said this design was “easy” or “clear” and one thought that “words can be more explicit if capture area is complicated”. Three users did not like that they had to read the text to find out where to aim rather than having a visual indicator at or near the link while one user found it initially confusing as to whether the marginal words and icon needed to be part of the image captured. One user compared the text condition to the others by saying they liked that there were “no distracting graphics”.

Participants commented on the overlay condition that the “circle clearly indicates where to point” and that it was “easy to aim without worrying about orientation/rotation angle”. Three participants compared the overlay condition to the text condition in their statements with two saying they preferred the overlay. Participants stated that the overlay provided a “better idea what the target was than (text condition)” and that they “didn’t have to read and figure out what I was supposed to scan, it was obvious”. One participant thought that the overlay design was “reasonably easy and clear, though I preferred the (text method)”. There were also comparisons between the overlay and sight conditions. One participant said of the overlay, it is “much easier than the arrows!” while another commented that there was “no feedback” with the overlays. Two participants also brought up the impact of the overlay on the original content saying the “circle interferes with image on document” and that the

overlay “impacted the design of the page too much”.

The sight condition was the most discussed by the users. Four users liked that the alignment of the mark on the paper with the mark on the phone meant there was a “clear target” and “less guessing about what to capture”. Two commented positively on the value of the feedback supporting alignment, e.g., “obvious feedback (both mark and color)”. And one thought that it “seemed faster but was probably not, just distracted by the task”.

The task of aligning the arrow on the phone with the arrow on the paper was considered challenging by most users. Comments included “hard to aim”, “needs snap to grid”, and “not always clear how far away to focus”. Two users thought that part of the difficulty was due to aligning a sight that is not in the center of the smartphone screen: “not very forgiving; might help if arrow was centered rather (than) on right edge of device” and “having to hold the phone in unbalanced way, with the target not in the center”. Finally, the inability of users to rotate the phone from portrait mode to landscape mode resulted in comments from three users: “phone had to be kept upright/vertical”, “didn’t like that I couldn't rotate the phone — the image/target should rotate with the orientation of the phone”, and “can’t use other orientation; felt locked in”.

Participants were also asked to provide responses on a Likert-scale to questions whether the interaction was frustrating, easy, or fun. They were also asked whether they were sure of what they were aiming for and whether they felt that they could aim at that target.

There was little difference for the three designs with regards to their being fun, conveying the intended target, and affecting the aim of users. Figure shows the responses indicate that use of the sight design was more frustrating than the other two designs. The results approach significance, $p<.06$ between the sight and text conditions and $p<.08$ between the sight and overlay conditions using Wilcoxen signed rank test due to the non-normal distributions. This assessment is in line with comments provided by participants: “it’s really frustrating for me”, “overlapping the arrows requires finer motor control than the other techniques, thus it is slightly more frustrating”. User comments during the performance of tasks echo their answers to the open-ended questions. Users mentioned that there was “no image stabilization, could snap-to-arrow to improve stability”, said it was “harder”, that they would “need a tripod for this”, and that the “arrow is too small”.

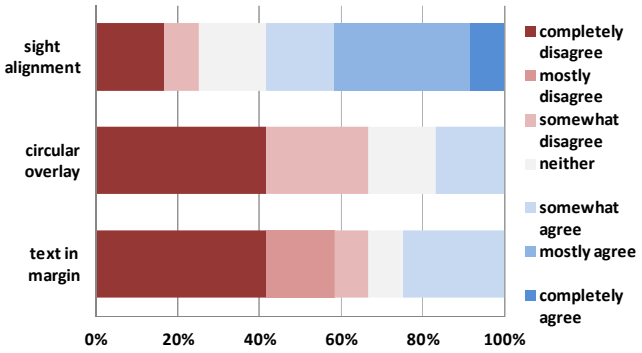


Figure 15. Likert-scale responses to "This interaction was frustrating."

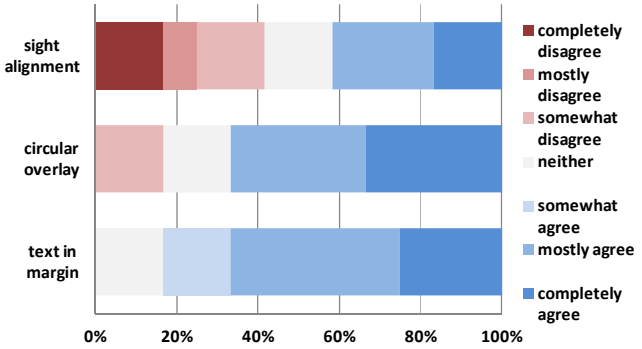


Figure 16. Likert-scale responses to "This interaction was easy."

The difficulties expressed in user comments also can be seen in their responses to the ease-of-use statement. Figure shows that responses indicate that use of the sight design was also perceived as harder to use with the difference between the sight condition and the text condition approaching significance ($p<.1$, Wilcoxen signed rank test).

Discussion

The study results were in line with our expectations. Indicating the target area with an overlay makes it clear what to aim for. It is easy to get such a large target approximately into the camera view. The circular design also allowed users to rotate the phone without having to align it to the page orientation. However, this particular approach is unacceptable to magazine publishers. They do not want to have parts of their artwork obscured. Putting corners around the target area instead of overlaying a circle does not make this option more acceptable to them.

The common practice of indicating targets by describing them with text in the page margin had an impact on performance. We expected the effect to be larger but it appears that the target regions in the study were too easily guessable. Targets also had clear borders so that they could be described sufficiently well with a few words.

The performance with the sight was between that of the other two conditions. We had hoped that more of the participants would have been intrigued by the game-like aspect of that interaction. However, several of the participants were frustrated and only a few really enjoyed this condition. The task might have been easier with a larger sight. As a larger arrow would not fit into the page margin next to the target, a larger sight would not be feasible. To make this condition more palatable, we believe that a better tracking algorithm is needed that outlines the recognized target on the screen and indicates to the user in which direction they should move the phone. The current approach requires too much fine motor

control, in part due to strict alignment requirements, making it especially difficult for people with shaky hands.

CONCLUSIONS

Media Embedded Target, or MET, is a new approach for placing a media link on paper. By placing a target in the margins of a page, it provides a good indication for the mobile phone capture region without interfering with the aesthetics of the page. Like other approaches, this approach links the media by matching keypoints in the images. The main advantage of this approach is that a fairly small target mark can define a much larger capture region.

We compared MET with two other approaches for indicating where links to digital content are available on paper. All three approaches are attempts to make links on paper more attractive than traditional barcode-like approaches while making it easy for users to locate and to follow those links. One of those approaches, indicating the presence of links with text and a logo in the margins, is commonly used in several commercial systems. While the logo has little impact on the visual appearance of the page, the associated text requires space on the page. Also, users must spend time reading the instructions and finding the link on the page. Another approach of placing a circular overlay on top of the link region leads to good performance but interferes with the page design by partially obscuring important content.

MET uses an on-screen sight to guide the user to a link. Rather than placing a marker on top of the link region, it puts it next to the link in the page margin. This approach represents a good trade-off in performance and impact on the page design.

The user study confirmed the relative performance of the three designs. While users were fastest with the overlay design, they also worried about its impact on the printed document. The text description of the link was perceived as slower and as introducing ambiguity into the process. The sight alignment approach was intermediate in terms of performance but the current implementation was considered the most challenging by users. This experimental result gives us clear direction on automating MET detection for further reducing users' frustration.

1. REFERENCES

[1] Bay, H., Ess, A., Tuytelaars, T., Van Gool, L. "SURF: Speeded Up Robust Features", Computer Vision and Image Understanding (CVIU), Vol. 110, No. 3, pp. 346-359, 2008.

[2] Blippar, blippar, Website: www.blippar.com

[3] Costanza, E., J. Huang., Designable Visual Markers, Proceedings of ACM CHI'09, pp. 1879-1888.

[4] DigiMarc, "Enhanced, Interactive Experiences from Media Using

Smartphones," http://www.digimarc.com/, 8/3/2011.

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

[6] Fuji Xerox, "Paper Fingerprint Recognition Technology," Online: http://www.fujixerox.com/eng/company/technology/xaya/, 8/3/11.

[7] Girgensohn, A., Shipman, F., Wilcox, L., Liu, Q., Liao, C., and Oneda, Y. 2011. A tool for authoring unambiguous links from printed content to digital media. In Proceedings of the 19th ACM international conference on Multimedia (MM '11). ACM, New York, NY, USA, 1561-1564.

[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: Algorithms and Systems II, 2008.

[9] Hartung, F. and Kutter, M. "Multimedia watermarking techniques," Proceedings of the IEEE, vol. 87, pp. 1079-1107, July 1999.

[10] Hecht, B., M. Rohs, J. Schöning, and A. Krüger. Wikeye –Using Magic Lenses to Explore Georeferenced Wikipedia Content. PERMID 2007.

[11] Hecht D. L., Embedded Data Glyph Technology for Hardcopy Digital Documents. SPIE -Color Hard Copy and Graphics Arts III, Vol. 2171, pp. 341-352.

[12] Henze, N. and S. Boll. Snap and share your photobooks. In Proceedings of ACM Multimedia'08, pp. 409-418.

[13] Hitachi, F. Develops RFID Powder, http://healthfreedoms.org/2009/10/23/hitachi-develops-rfid-powder/

[14] Holley, R. "How Good Can It Get?", D-Lib Magazine, March/April 2009.

[15] http://www.dlib.org/dlib/march09/holley/03holley.html

[16] Hull, J.J., B. Erol, J. Graham, Q. Ke, H. Kishi, J. Moraleda, and D.G.V. Olst. Paper-based Augmented Reality. Proceedings of IEEE ICAT 2007, pp. 205-209.

[17] IEEE Computer Scociety, Augmented Reality Issue of IEEE Computer Magzine, July 2012.

[18] Iwamura, M. Tsuji, T. and Kise, K. Memory-based recognition of camera-captured characters. In Proceedings of the 9th IAPR International Workshop on Document Analysis Systems (DAS '10). ACM, New York, NY, USA, 89-96.

[19] Ke, Y. and Sukthankar, R., PCA-SIFT: A More Distinctive Representation for Local Image Descriptors. Proceedings of IEEE CVPR 2004.

[20] Kooaba Shortcut, Application, Webpage: http://www.kooaba.com/en/products/

[21] Liu, Q. and Liao, C. "PaperUI," Proceeding of the 4th International Workshop on Camera-Based Document Analysis and Recognition, pp. 3–10, September 2011.

[22] Liu, Q. C. Liao, L. Wilcox, and A. Dunnigan. 2010. Embedded media barcode links: optimally blended barcode overlay on paper for linking to associated media. In International Conference on Multimodal Interfaces and the Workshop on Machine Learning for Multimodal Interaction (ICMI-MLMI '10). ACM, New York, NY, USA.

[23] Liu, Q., Liao, C., Wilcox, L., Dunnigan, A., and Liew, B. 2010. Embedded media markers: marks on paper that signify associated media. In Proceedings of the 15th international conference on Intelligent user interfaces (IUI '10). ACM, New York, NY, USA, pp. 149-158.

[24] Liu, X. and D. Doermann, Mobile Retriever: access to digital documents from their physical source. Int. J. Doc. Anal. Recognit., 2008. 11(1): pp. 19-27.

[25] Lowe, D.G., Distinctive Image Features from Scale-Invariant Keypoints. Int. J. Comput. Vision, 2004. 60(2): pp. 91-110.

[26] Microsoft Tag. http://www.microsoft.com/tag/

[27] Morrison, A., Oulasvirta, A., Peltonen, P., Lemmela, S., Jacucci, G., Reitmayr, G., Näsänen, J., and Juustila, A., Like bees around the hive: a comparative study of a mobile augmented reality map. In Proceedings of ACM CHI '09, pp. 1889-1898.

[28] Nakia, T., K. Kise, and M. Iwamura. Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval. LNCS, Vol. 3872, pp. 541-552.

[29] NODA TSUGIO, MOROO JUN, CHIBA HIROTAKA, Print-type Steganography Technology, Fujitsu 2006, Vol. 57. No. 3, pp. 320-324.

[30] Parikh, T.S., P. Javid, S. K., K. Ghosh, and K. Toyama. Mobile phones and paper documents: evaluating a new approach for capturing microfinance data in rural India. Proceedings of ACM CHI'06, pp. 551-560.

[31] Pixazza, Now a picture is worth more than a thousand words. http://pixazza.com/.

[32] Reilly, D., M. Rodgers, R. Argue, et al., Marked-up maps: combining paper maps and electronic information resources. Personal and Ubiquitous Computing, 2006. 10(4): pp. 215-226.

[33] Rekimoto, J. and Ayatsuka, Y. 2000. CyberCode: designing augmented reality environments with visual tags. In Proceedings of ACM DARE 2000, pp. 1-10.

[34] Ricoh, Tamago Clicker, website: http://www.ricoh.co.jp/software/tamago/clicker/

[35] Rohs, M. Real-world interaction with camera-phones. LNCS, Vol. 3598, pp. 74-89.

[36] Wikipedia, “Barcode,” Online: http://en.wikipedia.org/wiki/Barcode

[37] Wikipedia, Digital paper. http://en.wikipedia.org/wiki/Digital_paper.