ミックスド・メディア・リアリティ（MMR）
：紙と電子世界をつなぐ新しい方法

Mixed Media Reality (MMR) : A New Method of eP-Fusion†TM

ジョナサン ハル* バーナ エロル* ジェイミー グラハム* シーファ ク* 岸 秀信*
Jonathan J. HULL Berna EROL Jamey GRAHAM Qifa KE Hidenobu KISHI
ホーヘ マラレダ* ダニエル バン オルスト*
Jorge MORALEDA Daniel G. Van OLST

要　旨

紙文書と電子情報とをつなぐ新しい方法について説明する．これは紙文書の見た目やフォーマットを一切変える必要がない画期的なものである．我々がミックスド・メディア・リアリティ（Mixed Media Reality：MMR）と呼んでいるこの技術は，商業印刷物及びPC等から印刷される紙文書のどちらにも適用可能であり，紙の使い方／価値を大きく向上させる可能性を有している．本論文では，これを可能にした認識技術やいくつかのアプリケーション例，そして紙文書を認識し電子情報のリンクにアクセスする機能をカメラ付き携帯電話に実装した例についても述べる．実際に毎秒4フレームレベルでの認識速度をスマートフォンTreo700W上で実現し，いくつかのアプリケーションを実装した．この中には クリッカブルペーパー（“Clickable Paper™”）と名づけたアプリケーションが含まれるが，これはウェブページを通常印刷と一切見た目の変化のない形で印刷するにも関わらず，カメラ付き携帯電話を通じてその各リンク先を印刷物上でクリックできるものである．

ABSTRACT

A new method for linking paper documents to electronic information is described that does not modify the format of the paper document in any way. Applicable to both commercially printed documents as well as documents that are output from PC's, the technique we call Mixed Media Reality (MMR) substantially improves the utility of paper. We describe the recognition technology that makes this possible as well as several applications. An implementation on a camera phone is discussed that lets users retrieve data and access links from paper documents to electronic data. Recognition performance of 4 frames per second is achieved on a Treo 700w and support is provided for several user applications, including “clickable paper™” – printed web pages whose appearance is unchanged but that can be navigated with a camera phone.

† "eP-Fusion™": A World where paper information and electric information are closely linked

* California Research Center, Ricoh Innovations, Inc.

1. Introduction

Linking the physical and digital worlds is a long standing goal of eP-Fusion™. Previous techniques alter the appearance of paper documents with bar codes or textured paper. Often, a special purpose device, such as a bar code reader or a pen with a camera in it, must be employed to recognize an embedded code. These characteristics restrict the use of eP-Fusion™.

Our prior work indicated that images of small patches of text contain enough information to make them as unique as a fingerprint [1]. We showed that it was possible to distinguish a small rectangular region (one inch square in our previous research) from among thousands of other text image patches. At that time, we leveraged this characteristic to identify the electronic original for a given paper document. However, the same results indicate that patches of text can be used as links to electronic data in an eP-Fusion™ system.

This paper proposes a new method of interacting with documents termed Mixed Media Reality (MMR) that links patches of text to electronic data and uses a camera phone as the recognition device. This brings a new level of interactivity to paper documents and allows them to be updated without reprinting them. New links to electronic data can be easily created and the content of old links can be changed without modifying the original document.

2. Algorithm Outline

The operation of an MMR system is illustrated in Fig.1. MMR-enabled documents are created (a) by choosing a bounding box, applying the text patch feature extraction algorithm to the image data within it, and storing the relation between that data and some electronic information in a database. A simple example of electronic data is a URL that points to a web page. However, it could just as easily be a video file, an audio clip or even an electronic version of the original document itself.

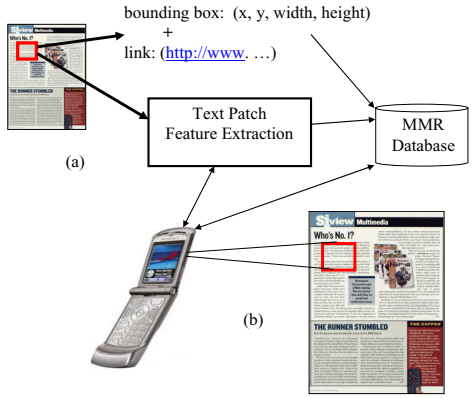

Fig.1 Creating mixed media reality documents (a) and using them (b).

At a subsequent time, Fig.1 (b), a user captures an image of a similar region with a camera phone; the system applies the same feature extraction to it and determines whether the database contains an association between those features and some electronic data. The data is returned to the phone and the appropriate rendering application is applied to it. If the data is a URL, a web browser could be invoked.

3. Text Patch Recognition

The objective of the text patch recognition algorithm is to correctly determine the identity of a page and the x-y position in the page of a small patch of text. The technical challenge is illustrated by the image in Fig.2 that shows the typical quality of images produced by commonly available camera phones. Characters are so blurry that "OCR" is basically impossible. However, it is still possible in almost every case to identify the bounding boxes around words since the spaces between words and lines can still be distinguished.

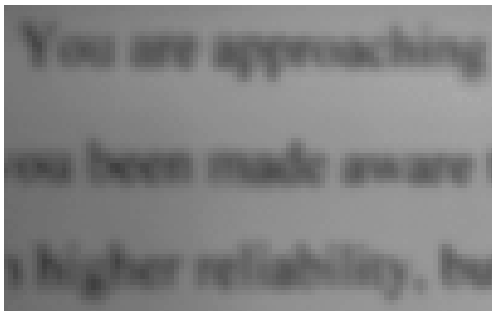

Fig.2 Typical camera phone image.

We developed a new text patch recognition algorithm based on arrangements of word bounding boxes. We detect bounding boxes by first applying a Laplacian edge detector to the image. Then a morphological dilation operation (with a structuring parameter that's wider horizontally than vertically) is used to smear the Laplacian image. The smeared image is then binarized with an adaptive global threshold. Using connected component analysis, word-like component boundaries are identified.

After word boundaries are identified, we assign a numerical value to each bounding box that is the normalized word length. The normalized word length, also called number of the nubs, is scale invariant. The number of nubs of each word is estimated by dividing the width of word in pixels by its height in pixels. We also look at the fraction of pixels remaining from this division to see if the word length could be one higher or one lower. For example, let's say that a word is 10 pixels high and 12 pixels wide. The word is (12/10)+1=2 long in nubs. The fraction is 2 pixels (12 mod 10), which is 20% of the word height ((12 mod 10)/10). If the fraction is smaller than a lower bound percentage (in our experiments this is 30%), then the word is assigned an alternate value that is one less than its length. So in this case it is equal to 1. The word is then said to be length "2" in nubs with "1" in alternate length. Similarly, if the number of normalized fraction pixels is larger than a higher bound percentage (in our experiments this is 70%), then the word is assigned an alternate length of one higher than its length in nubs.

After the length in nubs is computed, descriptors are calculated based on word cluster constraints. In our implementation, the word clusters are composed as shown in Fig.3. For a given word box, a word cluster is composed of the vertically overlapping word boxes. In order for a word cluster to be valid, the given word should have both above horizontally overlapping and below horizontally overlapping words, and at least two above or below horizontally overlapping words. Fig.3 shows two such feature points (not all of them are shown here). Then a descriptor value is computed as the weighted sum of the feature value of the current word box, the feature values of the above word boxes and the feature values of the below word boxes. For example, if the descriptor value is xxyyyzzz, it is composed of :

xx=length of the current word, up to 99

yyy=lengths of the above words, up to 3 words and each word <=9

zzz=lengths of the below words, up to 3 words and each word <=9

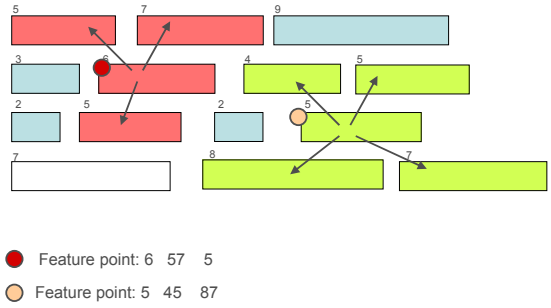

Fig.3 Computation of descriptors.

It is possible to have more than one descriptor associated with each word box. In such cases, new feature values are extracted containing all the alternates. For example, let's say the word box feature is its quantized length. Because of quantization, in some cases a word box can be assigned two lengths. Let's say the length of the current word box is 4, above lengths are 5

and 2(with alternate length 3) and the below word length is 8 (with alternate length 9). The extracted feature values would be : 4528, 4538, 4529, and 4539.

The computed descriptor values for the query image are looked-up in a hash table containing descriptor values of document patches in the MMR database. The document patches that contain the highest number of matching descriptor values are sorted and the first N patches are determined to be the matching candidates.

The similarity of relative locations of descriptors in the query image to those of patches in the MMR database are measured by computing a score that is based on the matching angles in two sets. This is shown in Fig.4. Angles from each descriptor to other descriptors (using the x,y coordinates) in the query image are computed. These are then compared to the angles between each descriptor and other descriptors in each database document patch candidate. If any angles for matching descriptors are similar (i.e. the L1 norm is smaller than a threshold) then the similarity score is increased by one.

Once the scores are computed between the query patch and each retrieved patch, the retrieved patch that provides the highest score is selected. The image, location of the image, link to the image, the source file, page, and location related to this image can be output. Alternatively, the first N retrieved images can be used to compute the output of the retrieval process. For example, if the first N retrieved images are linked to the same source file, page, and neighboring x,y locations, the retrieval process can output the source file, page number, and the average (or median) x,y locations of the first N patches retrieved from the database.

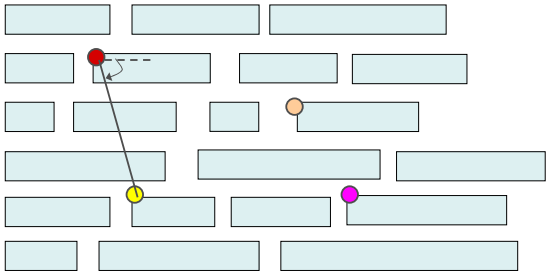

Fig.4 Computing a similarity score based on the relative locations of descriptors.

The current implementation runs on a Treo 700w with a 312 MHz PXA272 processor at about 4 frames per second. The processing times required by each module are presented in Table 1. These results are obtained by performing 2200 queries with video frames that are of size 176x144. The database contained 129 document pages or about 6500 patches. As can be seen from Table 1, the majority of time is spent on bounding box and text graph extraction. The space required for the database files was about 8 MB.

Table 1 Run time per image frame

Task	Time (ms)
Bounding box/graph extraction	172
BWC Retrieval	83
Descriptor computation	11
Other processes	13
Total Processing time	279

The accuracy largely depends on the number of word boxes present in the document patch and the accuracy of word bounding box detection. We performed experiments in order to test the change in accuracy with the increase in the number of word boxes in the query image. The database contains more than 4397 document pages or about 220,000 patches. In this case, the database files require 80 MB of storage. 8276 query patches of size 176x144 were generated with a system [2] that simulates the video output by a Treo 700w as it's moved over a document.

The first step in the experimental evaluation used "perfect" queries so that we could see how many

bounding boxes are needed to identify a patch. Perfect queries contain the exactly correct bounding boxes for each patch. As can be seen from Fig.5 (a), the percentage of correctly retrieved images is almost 100% if the query patch contains more than 40 word boxes.

Fig.5 (a) also shows the retrieval performance when low quality images, with blur and shadow masks, typical of what's output by the Treo, are submitted as queries. In this case, performance increases with the amount of text present in the document patch. When a document patch contains more than 100 text boxes, the percentage of correctly retrieved patches is approximately 60%. If a confidence threshold is imposed on the retrieved document, retrieved patches with low confidence are rejected. This reduces the error rate at approximately a 5% drop in retrieval performance. It's important to note that while a 60% retrieval rate may seem low, in practice it's more than adequate because the MMR system runs in real time on a video stream and the user actively moves the camera over the document, essentially cooperating with the recognizer to improve its performance.

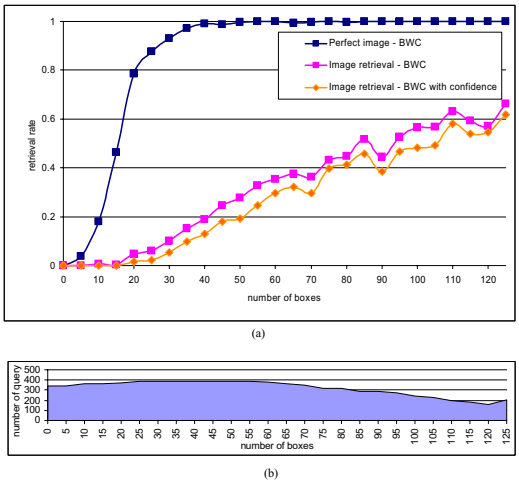

Fig.5 (a) Retrieval performance based on the number of bounding boxes in query images and (b) distribution of number of bounding boxes in the query image set.

4. Applications

There are many possible applications for MMR. Important considerations include whether the MMR database is on the phone or on a server and whether the database is created as a side effect of printing a document on a PC. This section presents two examples, among many we've created and have running, that illustrate the capabilities of the seamless paper-electronic interface we've created.

4.1 Travel Guidebook

Travel guidebooks are almost out-of-date the minute they are printed. The accuracy of the time-sensitive information they contain is always in question. A reader has no way of knowing whether the opening times of the attractions mentioned in such a publication are correct. The only reasonable solution is to call the facility in question. Instead, MMR allows someone to point a camera phone at the passage of text that describes the facility and retrieve the most currently available information about it.

Fig.6 shows an example of an MMR-capable travel guidebook. With the recognition system described previously, it is possible to overlay an indicator (in this example a red dot) that shows there is online information available related to the underlying passage of text. When a button on the phone is pressed, the client application retrieves the menu of choices (Fig.6 (b)) related to the text passage. Based on the user's selection, the appropriate information is displayed, in this case the opening times for the San Diego Zoo.

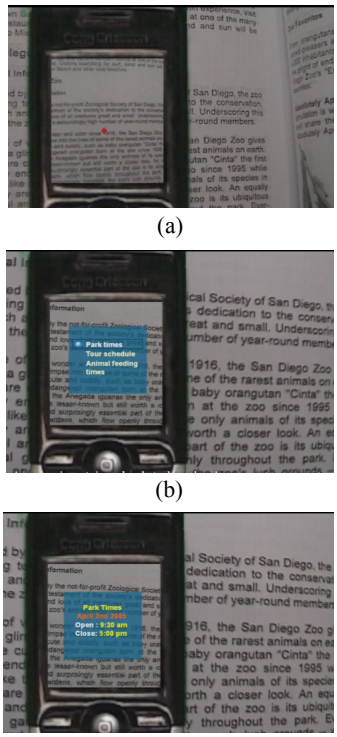

Fig.6 An augmented travel guidebook. An indication that information is present (a), a menu of choices (b), and the selected information (c).

4.2 Self-Printed Documents

Documents that are printed on a desktop PC are typically created by an individual for their personal use. MMR allows a user to customize the interactivity of those documents based on their own needs and permits the database to be under the user's personal control. It could be shared on a networked server, saved on the individual's PC, or pushed to the user's camera phone.

We created the architecture for printing and indexing web pages shown in Fig.7 that automatically captures an image of every document in the print driver, indexes them in the MMR database, and associates the URL's in the document with their physical location on the web page. This includes a plug-in in Internet Explorer that exports URL's to text patch feature extraction software. This system is fully implemented on Windows XP and Vista and can be installed on any PC.

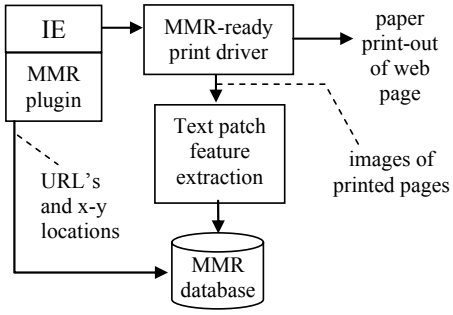

Fig.7 Architecture for automatic creation of augmented web page print-outs.

The version of MMR we call Clickable Paper™ is shown in Fig.8. The paper printout of a web page (a) is imaged and the region surrounding a URL is captured (b). That image as well as the web page corresponding to the URL (c) is shown in the user interface on the camera phone (d). A short history of the last three URL's accessed with this system is also displayed in the user interface.

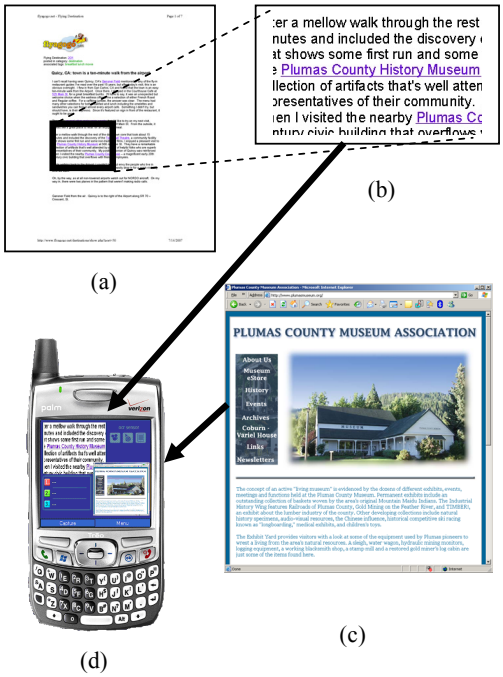

Fig.8 Clickable Paper™ system. A printed web page (a), image of patch containing a URL (b), web page (c), and UI on Treo (d).

5. Conclusions

A new paradigm for eP-Fusion™ was described in which electronic data is added to paper documents without changing the appearance of the paper document in any way. This approach leverages the essential discovery that a unique signature can be derived from an image of a small patch of text and that signature can be linked to electronic data. We described a new algorithm for text patch recognition and presented experimental results that demonstrated it can distinguish an image of a patch of text from a collection of thousands of examples. An implementation on a camera phone that runs at 4 frames per second was discussed. Two applications were presented : an MMR-enabled guidebook and Clickable Paper™. Both of them show the value of eP-Fusion™ and how it can be used in everyday life.

References

1. J. J. Hull and J. F. Cullen, "Document Image Similarity and Equivalence Detection," IEEE Int. Conf. on Document Analysis and Recognition, Ulm, Germany, 308-311, Aug. 18-20, 1997.
2. A. Lookingbill, E. R. Anutnez, B. Erol, J. J. Hull, Q. Ke, and J. Moraleda, "Ground-truthed Video Generation from Symbolic Information," IEEE Int. Conf. on Multimedia and Expo, Beijing, 1411-1414, July 3-5, 2007.