MOBILE VISUAL SEARCH ON PRINTED DOCUMENTS USING TEXT AND LOW BIT-RATE FEATURES

Sam S. Tsai[1], *Huizhong Chen*[1], *David Chen*[1], *Georg Schroth*[2], *Radek Grzeszczuk*[3] *and Bernd Girod*[1]

[1]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA
[2]Institute of Media Technology, Technische Universität München, Germany
[3]Nokia Research Center, Palo Alto, CA 94304, USA

ABSTRACT

We present a novel mobile printed document retrieval system that utilizes both text and low bit-rate features. On the client phone, text are detected using an algorithm based on edge-enhanced Maximally Stable Extremal Regions. The title text image patch is rectified using a gradient based algorithm and recognized using Optical Character Recognition. Low bit-rate image features are extracted from the query image. Both text and compressed features are sent to a server. On the server, the title text is used for on-line search and the features are used for image-based comparison. The proposed system is capable of web-scale document retrieval using title text without the need of constructing a document image database. Using features for image-based comparison, we can reliably match retrieved documents to the query document. Last, by using text and low bit-rate features, we can reduce the transmitted query size significantly.

Index Terms— mobile visual search, image retrieval, document retrieval, document analysis

1. INTRODUCTION

Visual search has gained interest as hand-held devices become equipped with powerful computing resources and high resolution cameras. Applications such as a location recognition [1, 2], product recognition [3, 4], and document retrieval [5, 6] are many but not the only ones being developed. Most of these systems rely on image-based features [7, 8, 9]. While text information is available in a query image, document retrieval systems [10, 5, 11, 6] still use image-based matching techniques because they assume text recognition on the camera phone tends to be unreliable due to photometric distortions, geometric distortions and clutter. However, with the availability of high resolution cameras and increasing compute power on hand-held devices, it is possible to perform character recognition on the phone.

We are primarily interested in applications that involve retrieving scholarly articles. One example is an application that allows researchers to easily share articles with colleagues by taking a picture of the articles using camera-phones. Performing web-scale retrieval using image-based approach requires acquiring the documents and building an image database. The acquisition process is time consuming and the run-time requirements for the document image database is demanding. Thus, we have developed a mobile printed document retrieval system that does not require an image database by combining text-based search and image-based comparison. We use words of the title to find related documents from on-line document databases using text-based search. The title is the description or summary of the article; thus, they are the most efficient keywords for text-based search. Then, image features of the query image are used to perform image-based comparison to reliable find the correct match even when the title text is not fully recognized.

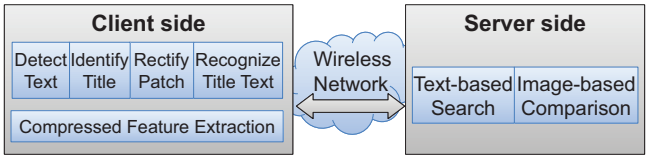


Fig. 1. The proposed mobile printed document retrieval system. Text and image features are extracted on the mobile device and sent over wireless network to a server. On the server, text is used to query an on-line document database and features are used for image-based comparison with the retrieved results.

A block diagram of the proposed system is shown in Fig. 1. A query is initiated when a user takes a picture of the title page of an article. Text patches are first robustly located in camera phone images using a new text detection algorithm based on edge-enhanced Maximally Stable Extremal Regions (MSER)[12]. A rectification method based on gradients removes the effects of perspective distortion within the text patch and the text is recognized using Optical Character Recognition (OCR). Both title text and image features are sent to a remotely located server. The server uses the title text to retrieve a list of possible matching candidates from on-line document databases and the list is re-ranked by image-based comparison of the query image and candidate document images. The best matching document titles are passed back to the client.

There are many benefits of the proposed system. First of all, the server side is capable of web-scale search with the available title text. The overhead of building an image database, as required in [10, 5, 11, 6], is reduced and the run-time requirements of the server is also lowered. Furthermore, by image-based comparison of the documents returned from on-line text search with the query image, we can reliably pick the correct match and improve recall even when the title text is not fully recognized. Last, the amount of query data sent to the server is substantially reduced by using only text and compressed features. This significantly improves system response time and reduces power consumption on the mobile device.

The rest of paper is organized as follows. In Sec. 2, we review related work on camera image text recognition. In Sec. 3, the architecture of our proposed system is described in detail. Evaluations of our system are provided in Sec. 4.

2. RELATED WORK ON TEXT RECOGNITION

Performing text recognition on camera images is a challenging task [13]. First, text can be anywhere within the image. Thus, a text detection algorithm is typically applied to locate the text such that text recognition can be performed efficiently. Text detection algorithms can be divided into two categories: 1) Connected component based [12], where character candidates are detected from the image and connected to form text lines based on their geometric properties. 2) Region based [14], where features within a sliding window is used to classify whether the region contains text. We developed and used the algorithm as described in [12] for our system. Second, text in camera-phone images may be taken from a perspective view. The perspective distortion prohibits OCR from performing reliable recognition. Many methods have been proposed to correct the perspective distortion. Pilu [15] propose methods to estimate the perspective transform using paragraph boundaries. In [16], Clark and Mirmehdi find the vanishing points of the document and re-project the vanishing points to rectify the image. Monnier et al. [17] took a similar approach, but they estimate the vanishing points using distances between the text lines. In [18], Zhu et al. detect vertical strokes of text characters using edge maps to estimate the vertical vanishing point. Myers et al. [19] extract text patches and correct the shear using vertical projections of the characters within the text patches. In this work, we develop a novel rectification method that works directly with single line text patches. We estimate the slant angle from the orientation of the gradients and correct the slant of the text patch. Our method requires only taking the picture the single title line of the document and is of lower complexity.

3. MOBILE PRINTED DOCUMENT SEARCH SYSTEM

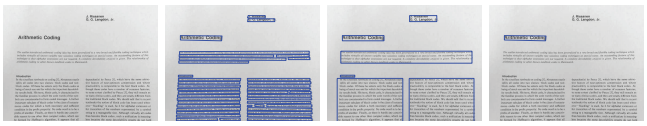
In this section, we describe our mobile printed document retrieval system as shown in Fig. 1. A query is initiated when a user takes a picture of a document using a client phone. We detect the text within the image using an edge-enhanced MSER-based algorithm and identify the title based on a simple layout rule. The image patch containing the title text is rectified using a gradient based algorithm and recognized using OCR. In addition to recognizing the title text, we extract compressed image features from the query image. Both title text and compressed image features are sent from the client to a remotely located server, where the search is performed. The server uses the title text to retrieve a list of possible matching documents from on-line document databases. Then, image-based comparison is performed by comparing the query features to the image features of the possible matching documents to find the best matches. We describe each stage in the following sections.

3.1. Text Detection and Title Identification

The title text is typically of larger font and contains the most effective query keywords. To find text in the image, we run a text detection algorithm based on edge-enhanced MSERs and SWT on the query image [12]. MSERs are detected from the image and pruned using Canny edges, forming the character candidates. Then, stroke widths of the character candidates are found based on distance transforms. The character candidates are then linked together based on their geometric properties to form text lines, Fig. 2(b). Then, text lines are grouped into paragraphs if their orientations are the same and the distances between the text lines are smaller than the text line height, Fig. 2(c). Paragraphs having more than three lines are considered as the document body. Any text line below the first line in the document body is not considered as a title candidate. Then, we mark the score S_i of the remaining paragraphs using the following:

$$S_i = h_i + \alpha \cdot w_i, \quad (1)$$

where, h_i and w_i is the average height and average stroke width of the letters within the paragraph, respectively. α is a weighting and is experimentally determined to be 5. The top scoring paragraph is considered as the title text, Fig. 2(d).



(a) (b) (c) (d)

Fig. 2. Extracting the title patch. (a) Original image. (b) Detected text lines. (c) Text lines are grouped into paragraphs. (d) The highest scoring paragraph is declared as the title patch.

3.2. Rectifying Text Patches for OCR

After the title is identified, the image patch containing the title text is extracted from the original image for OCR. To reliably recognize the title characters, we remove the effect of perspective distortion within the text patch. As a first step, we unify the character heights on the text line by finding the x-line and the base-line, as shown in Fig. 3(a), of the characters and warping the image so that the two lines are parallel. Then, we find the slant of the characters and transform the characters to their upright position to obtain the rectified image, Fig. 3(b).

To estimate the slant angle, we use the image gradients between the x-line and base-line, Fig. 3(c). The gradients are first converted into polar coordinate form,

$$G_r(u,v) = \| (\partial I_x(u,v), \partial I_y(u,v)) \|_2, \quad (2)$$
$$G_\theta(u,v) = \arctan(\partial I_x(u,v), \partial I_y(u,v)). \quad (3)$$

Then, a response of an angle is calculated using the following:

$$H_M(\alpha) = \sum_{(u,v)\in M} G_r(u,v) \cdot I\left(\|G_\theta(u,v) - \alpha\| < \frac{\delta}{2} \right), \quad (4)$$

where δ is a tolerance factor, M is the region between the x-line and base-line, and I is the indicator function. The dominant angle is found as

$$\alpha^* = \underset{\alpha}{\operatorname{argmax}}\, H_M(\alpha). \quad (5)$$

The slant angle is the difference of α^* to the horizontal direction. Intuitively, the region M contains mostly vertical strokes. Thus, the dominant angle will be horizontal if the characters are upright. In the interest of speed, this method can be implemented as histogram voting schemes as shown in Fig. 3(d) where the angle corresponding to the maximum count bin is the dominant angle.

3.3. Extracting Compressed Features

In parallel with the title text recognition, we extract local image features from the query image. We use the Compressed Histogram of Gradients (CHoG) [20] descriptor with location histogram coding

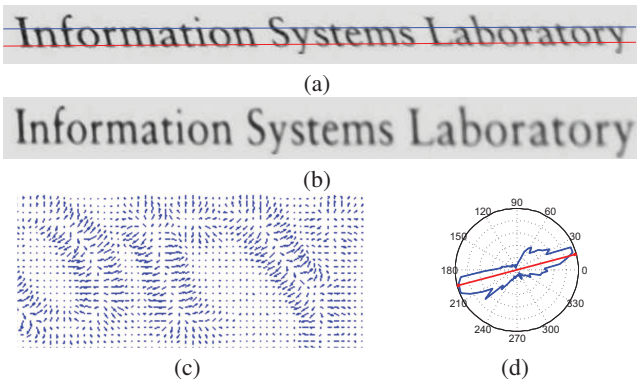


Fig. 3. (a) Original image and the detected x-line (blue) and baseline (red). The patch is recognized as "hxiormatkon Sy stems Labommng". (b) Rectified image, which is recognized as "Information Systems Laboratory". (c) Gradients of points between the two lines. (d) Orientation histogram of gradients. The dominant angle is indicated by the red line.

[21] to minimize the query data size. Since our text detection algorithm uses MSERs, they can also act as the interest points for CHoG to further reduce computation time.

3.4. Text-based Search and Image-based Comparison

The title text and compressed features are both transmitted to the server. The title text is used to search on-line databases. For reliable search, keyword spellings are checked and corrected. We further remove keywords that are not found in the dictionary. The processed keywords are used to query document database search engines, such as Google Scholar, CiteSeer, IEEExplore, and a list of candidate matching documents is retrieved. Then, the query features are used to perform image-based comparison with the image features of candidate matching documents. We assume the databases are able to provide the image features of the title page or simply the image which we can extract features from. We match the query features to the candidate image features using the ratio test [7] and an affine-model RANSAC. The final list is re-ranked according to the number of feature matches after RANSAC. Compared to [10, 5], our system combines text and image-features for search and eliminates need of a document image database. However, our system exhibits greater latency due to text recognition.

4. EXPERIMENTAL RESULTS

In this section, we present evaluation results of our system. We create a document image dataset by picking a set of topics from the engineering literature and taking pictures of printed documents of each topic. We choose 25 topics and take a total of 501 query images using a camera-phone with a 5M pixel camera[1]. The pictures are taken from various view-points as shown in Fig. 4.

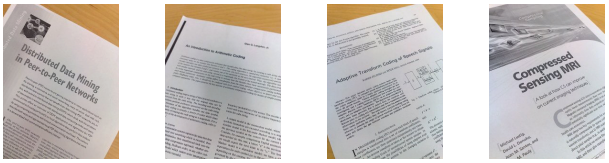


Fig. 4. Samples of the printed document dataset.

[1]http://msw3.stanford.edu/∼sstsai/PaperSearch/

For each query image, text detection is performed on the SVGA-size image while title text patches are extracted from the full-size image. The open source Tesseract recognition engine[2] is used to recognize the text from the patches. The recognized keywords are posted to Google Scholar for on-line search. We combine results of both keyword search and exact phrase search with a mix ratio of 4:1. Compressed image features, CHoG[3], with 1000 features per image are extracted from the VGA-size image. We use documents of the same topic to act as the non-matching results from the returned list for image-based comparison. Experimental results of each part of the system are presented in the following sections.

Text Detection and Title Identification. For each query image, we check whether text is detected and the title is identified. An image is declared to be correct if and only if every character in the title text is included within the detected box. We achieve a high accuracy as shown in Table 1.

Table 1. Title text extraction performance.

	Correct Images	Correct Percentage
Text detection	477	95.2%
Title identification	434	86.6%

Title Text Recognition. OCR is performed on the extracted text patches. We show the recognition performance of OCR on the patches before and after rectification in Table 2. The error count is the editing distance between the ground truth title and the recognized title. After rectification, error is reduced from 16.2% to 12.7%.

Table 2. Title text recognition performance.

	Errors Per Query	Perc. of the Title
Before rectification	8.92	16.2%
After rectification	7.03	12.7%

Document Recall. After the title is recognized, the keywords are posted to on-line search engines. The returned document list is compared using image features and re-ranked according to the number of feature matches after RANSAC. The recall performance is shown in Fig. 5. A recall of ∼71% on the top entry is observed for the system without re-ranking. With re-ranking, we can boost the recall performance of the top entry to ∼82%. We achieve a recall performance of ∼85% for the top 5.

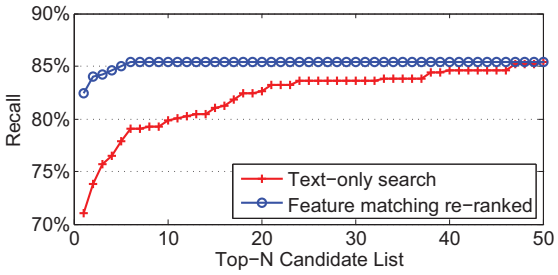
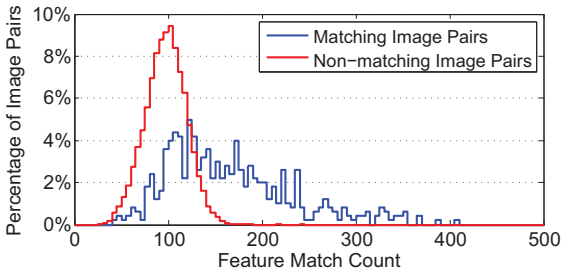


Fig. 5. Recall performance of returned list of the search engine and that after the list has been re-ranked.

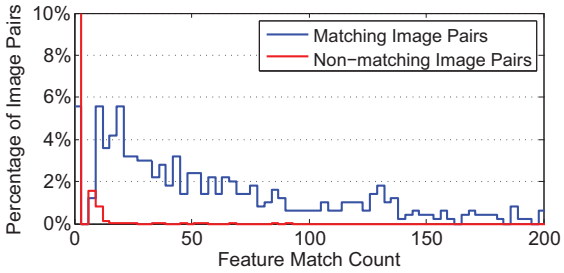
Fig. 6 depicts the distribution of the number of feature matches for matching and non-matching image pairs. We see that after performing RANSAC, most false feature matches are rejected.

[2]http://code.google.com/p/tesseract-ocr/

[3]http://www.stanford.edu/∼vijayc/chog-release.zip



(a)



(b)

Fig. 6. Feature match count distribution for matching and non-matching image pairs: (a) Feature matches after ratio test matching. (b) Feature matches after RANSAC.

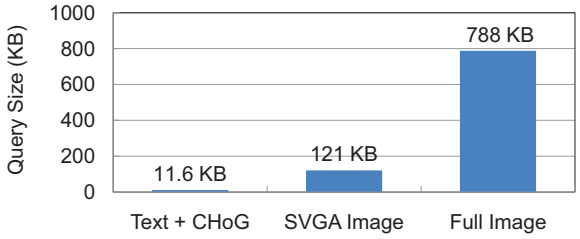


Fig. 7. Size comparison of using text and compressed features, SVGA-size image and full-size image as query. The reduction in query size minimizes network delay and improves the system response time.

Query Size Comparison. We compare the query size of transmitting the text and compressed features to transmitting the query image to the server in Fig. 7. Sending the full-size grayscale image to the server for text recognition requires on average 788 KB. If we trade off recognition performance with query size, we can consider transmitting at least SVGA-size grayscale images, where recognition errors increase by $\sim$12%. The SVGA-size image has a reduced query size of 121 KB on average. For the proposed system, sending the recognized title text and compressed features only requires 11.6 KB per query. Compared to the case of sending a query of full- or SVGA-size image, it achieves $\sim$67.9$\times$ or $\sim$10.4$\times$ reduction in query size.

Query Latency. We implement our algorithm on a Nokia N900 phone with a 600MHz ARM processor and a server with 2.53GHz processor. The processing time for text detection, text recognition, and feature extraction is 6 seconds on the phone or 1 second on the server. The phone and server is connected over wireless 3G, which is typically congested, with a uplink of 60kbps. The total query latency is $\sim$9 seconds for our system. However, for a system that sends the full- or SVGA-size image, the query latency is $\sim$107 and $\sim$18 seconds, respectively.

5. CONCLUSIONS

We propose a novel mobile printed document retrieval system. The system utilizes both text recognition and image-based features. Title text patch is located from the image based on an edge-enhanced MSER text detection algorithm. The extracted text patch is rectified using a gradient based algorithm and recognized using OCR. Rectification of the text patch reduces the errors of character recognition by $\sim$22%. Compressed features, extracted from the query image, and the title text are both sent to a remotely located server. The server uses the recognized title text for on-line search and performs image-based comparison on the retrieved list using the compressed features. The proposed system achieves a recall of over 82% on the top match. Compared to a system that sends the full-size image, the query size is reduced by $\sim$67.9$\times$. The reduction in query size can minimize network delay and lower power consumption. Furthermore, by using the title text, the system is capable of performing web-scaled document search without building an image database beforehand.

Acknowledgements The authors would like to thank Vijay Chandrasekhar, Gabriel Takacs, Ngai-Man Cheung, and Ramakrishna Vedantham for the invaluable discussions and suggestions.

6. REFERENCES

[1] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, and B. Girod, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," in *ACM International Conference on Multimedia Information Retrieval*, Vancouver, Canada, October 2008.

[2] "Google goggles," http://www.google.com/mobile/goggles/.

[3] S. S. Tsai, D. M. Chen, V. Chandrasekhar, G. Takacs, N.-M. Cheung, R. Vedantham, R. Grzeszczuk, and B. Girod, "Mobile product recognition," in *ACM International Conference on Multimedia*, 2010.

[4] "Snaptell," http://www.snaptell.com.

[5] J. Moraleda and J. J. Hull, "Toward massive scalability in image matching," in *International Conference on Pattern Recognition*, 2010.

[6] Q. Liu, H. Yano, D. Kimber, C. Liao, and L. Wilcox, "High accuracy and language independent document retrieval with a fast invariant transform," in *International Conference on Multimedia and Expo*, 2009.

[7] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, vol. 60, no. 2, pp. 91–110, November 2004.

[8] H. Bay, T. Tuytelaars, and L. V. Gool, "SURF: speeded up robust features," in *European Conference on Computer Vision*, 2006.

[9] V. Chandrasekhar, G. Takacs, D. Chen, S. S. Tsai, R. Grzeszczuk, and B. Girod, "CHoG: Compressed Histogram of Gradients," in *In Proceedings of Conference on Computer Vision and Pattern Recognition*, 2009.

[10] B. Erol, E. Antúnez, and J. Hull, "HOTPAPER: multimedia interaction with paper using mobile phones," in *ACM International Conference on Multimedia*, 2008.

[11] T. Nakai, K. Kise, and M. Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval," in *Document Analysis Systems*, 2006.

[12] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen., V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod, "Robust text detection in natural images with edge-enhanced maximally stable extremal regions," in *International Conference on Image Processing*, 2011.

[13] J. Liang, D. Doermann, and H. P. Li, "Camera-based analysis of text and documents: a survey," *Int. Journal on Document Analysis and Recognition*, 2005.

[14] X.-R. Chen and A. L. Yuille, "Detecting and reading text in natural scenes," in *Conference on Computer Vision and Pattern Recognition*, 2004.

[15] M. Pilu, "Extraction of illusory linear clues in perspectively skewed documents," in *Conference on Computer Vision and Pattern Recognition*, 2001.

[16] P. Clark and M. Mirmehdi, "Rectifying perspective views of text in 3D scenes using vanishing points," *Pattern Recognition*, 2003.

[17] C. Monnier, V. Ablavsky, S. Holden, and M. Snorrason, "Sequential correction of perspective warp in camera-based documents," in *International Conference on Document Analysis and Recognition*, 2005, pp. I: 394–398.

[18] Y. Zhu, R. Dai, B. Xiao, and C. Wang, "Perspective rectification of camera-based document images using local linear structure," in *Proc. of ACM symposium on Applied computing*, 2008.

[19] G. K. Myers, Robert C. Bolles, Q.-T. Luong, J. A. Herson, and H. B. Aradhye, "Rectification and recognition of text in 3-d scenes," *International Journal on Document Analysis and Recognition*, 2005.

[20] V. Chandrasekhar, G. Takacs, D. M. Chen, S. S. Tsai, Y. Reznik, R. Grzeszczuk, and B. Girod, "Quantization schemes for CHoG," in *International Workshop on Mobile Vision*, 2010.

[21] S. S. Tsai, D. M. Chen, G. Takacs, V. Chandrasekhar, R. Vedantham, R. Grzeszczuk, and B. Girod, "Location coding for mobile image retrieval," in *Proc. 5th International Mobile Multimedia Communications Conference*, 2009.