Mobile Visual Search Using Image and Text Features

Sam S. Tsai[1], Huizhong Chen[1], David Chen[1], Ramakrishna Vedantham[2], Radek Grzeszczuk[2] and Bernd Girod[1]
[1]Department of Electrical Engineering, Stanford University, Stanford, CA 94305, USA
[2]Nokia Research Center, Palo Alto, CA 94304, USA

***Abstract*—We present a mobile visual search system that utilizes both text and low bit-rate image features. Using a cameraphone, a user can snap a picture of a document image and search for the document in online databases. From the query image, the title text is detected and recognized and image features are extracted and compressed, as well. Both types of information are sent from the cameraphone client to a server. The server uses the recognized title to retrieve candidate documents from online databases. Then, image features are used to select the correct document(s). We show that by using a novel geometric verification method that incorporates both text and image feature information, we can reduce the missed positives up to 50%. The proposed method can also speed up the geometric process, enabling a larger set of verified titles, resulting in a superior performance compared to previous schemes.**

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

I. Introduction

Mobile visual search has gained major interest as mobile devices become equipped with powerful computing resources and high resolution cameras. Location recognition [1], [2], product recognition [3], [4] are some of the many creative applications that have been developed based on mobile visual search techniques. These systems extract local image features, such as SIFT [5] or CHoG [6], from a query image a user takes and use the bag-of-visual-features (BoVF) matching approach [7] to find a match within a large image database for recognition.

To perform visual search on objects with text, such as paper documents, the BoVF approach can also be used [8]. However, with the knowledge that the query image consists primarily of text, document visual search can be performed by using descriptors that are built upon characters and words. In the works of Nakai et al. [9] and Moraleda [10], they used spatial distributions of words to generate descriptors for search, while Moraleda additionally generated alphabetical codes from the descriptors and used the generated text code to perform matching. Visual features or word features based approaches typically yield good accuracy, but are difficult to scale.

In [11], we proposed a new architecture for mobile visual search where we can utilize both text and image features. On the cameraphone, text is detected and the title text is recognized while image features are also extracted. The two types of information are transmitted to a server. The title text used to perform an online search and the image features are used to verify the search results. Compared to previous schemes, the system has the advantage of being capable of performing a web-scaled database search while still achieving the accuracy of the visual feature based approach.

We further improve the system by incorporating text information within the geometric verification framework in this work. The geometric verification process typically involves the following two steps. First, the features of one image are matched to the features of the other based on their similarity where the ratio test [5] is applied to filter out ambiguous matches. Then, RANSAC [12] is used to estimate a geometric transformation model between the locations of the matching features.

Two issues arise when the query image and document image depict paper documents. For query images that contain the document title, the view typically does not contain the full document title page. Thus, all features that are not contained in the query image act as background clutter. Furthermore, text typically contains repetitive structures which would cause the ratio test to reject potential matches. We address these two issues by incorporating the title text information within the geometric verification framework. From pairwise image matching experiments, we find that we can improve the matching performance, reducing the missed true positives, and reduce the total time needed for verification. While incorporate in a web-scale document search, our system is capable of verifying a larger list of candidates and improving the recall.

The rest of the paper is organized as follows. In Section II, we explain our proposed algorithm that incorporates both text and image features in a mobile visual search system. In Section III, we describe how in detail how we can further improve the system using a geometric verification scheme that utilizes text information. We then demonstrate the superior performance of our system by the experimental evaluations of pairwise matching and web-scale document retrieval in Section IV.

II. System Description

The pipeline of our mobile visual search system is shown in Figure 1. Title text recognition and image feature extraction are performed on the client side to reduce the data sent over the network. We only perform OCR for the title text to reduce the total processing time on the client phone. Searching for the document in an online database and geometrically verifying the match is done on the server where processing power is

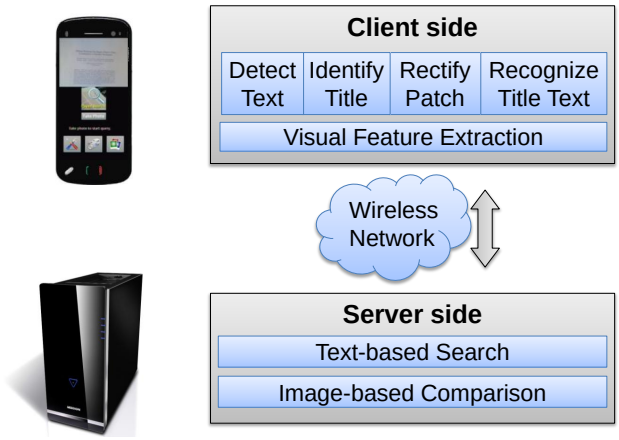


Fig. 1. Diagram of the mobile visual search system that use both text and image features.

abundant. We describe the title text recognition pipeline and image feature extraction on the client phones, and explain how the title text is used to search from an online database in following subsections.

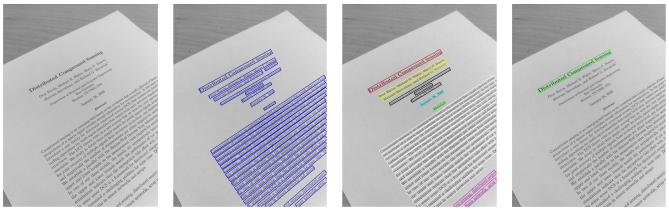
A. Text Detection

A standard OCR engine typically only recognizes characters from high-resolution text images under good lighting conditions and aligned horizontally [13]. For the OCR engine to work on cameraphone images, an additional step of text detection is needed to locate the text [14]. In our work, we use a text detection algorithm that is based on edge-enhanced Maximally Stable Extremal Regions [15]. Maximally Stable Extremal Regions (MSER) are extracted from the query image and pruned using Canny Edges [16]. Stroke width information of the regions are calculated using the distant transform of the pruned MSERs. MSERs are considered as character candidates and are pairwise connected based on their geometric properites, i.e., size, location, stroke width. Finally, straight text and word lines are extracted from the connected sets of character candidates. The output of the text detection algorithm are bounding boxes on the text within the query image, see Fig. 2 (a) and (b).

B. Title Identification

From all the detected text boxes, we identify the text boxes that correspond to the title. Since the title text typically contains the most informative keywords to search for a document, we only run OCR on these identified title text patches. This reduces the processing time on the client phone.

The title text typically stands out in terms of size and stroke width, i.e., the font is typically large and emphasized so the size and stroke width, compared to the other parts in the document, are typically larger and wider. Based on this observation, we first group text lines that have similar orientations and are close to one another, see Fig. 2 (c). Groups that have more than three lines are typically the document body and are not considered as a title candidate. Then, for the remaining groups, we calculate a score, S, that is based on the character height and stroke width, i.e., $S = h + \alpha \times w$, where h, w are the height and with respectively and α is a constant which is experimentally determined. The group with the highest score is the title box, see Fig. 2 (d). We extracted the image patch from the title box and pass it to the next stage of the character recognition pipeline.



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

Fig. 2. (a) Original image. (b) Detected text boxes. (c) Groups of text boxes. (d) Identified title patch.

C. Text Patch Rectification

Although the title text patch is extracted, an additional rectification step for the image patch is needed before running OCR on the text patch. Because the query viewpoint is unconstrained, there is typically perspective distortion and the characters are typically not in their upright position. Thus, if we pass the original text patch into an OCR engine, the recognition accuracy is not good, as shown in Fig. 3 (a) and its caption. Thus, we rectify the text patch, Fig. 3 (b).

Our rectification process consists of two steps. First, we unify the character heights throughout the text line by projecting image so that the *x-line* and the *base-line* of the text patch is parallel. The *x-line* is found by fitting a line on the upper pixels of the characters while the *base-line* is found by fitting a line on the lower pixels of the characters, see Fig. 3 (a). Second, we undo the shear by warping the image so that the characters are not slanted. We estimate the slant angle by using the gradients of the text patch between the *x-line* and the *base-line*, Fig. 3 (c). The strokes between the two lines typically consists of the vertical strokes so the gradients should be oriented in the horizontal direction. Thus, if the characters are slanted, the dominant orientation would change accordingly. We estimate the slant angle by simply finding the maximum count bin of the orientation histogram of the gradients, see Fig. 3 (d).

D. Image Feature Extraction

In parallel with the title text recognition, we extract local image features from the query image. We perform interest point detection on the query image. We code the locations of the interest points using Location Histogram Coding [17] and the orientation of the features using a simple quantization scheme and entropy coding. We discuss the how quantizing the orientation affects the matching peformance in Section IV-A. From each interest point, we extract CHoG descriptors [6],

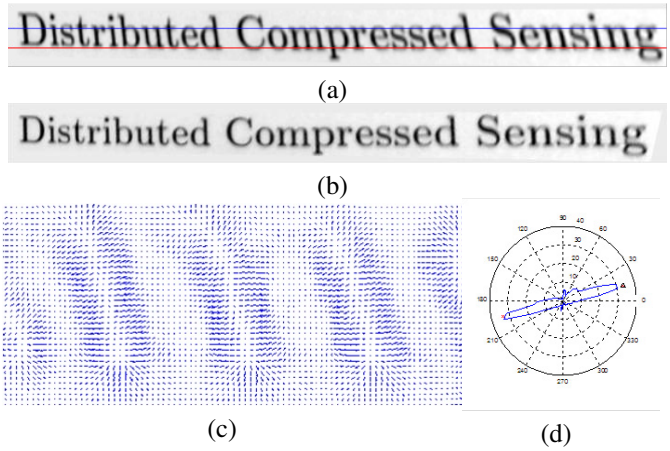


Fig. 3. (a) Perspectively distorted image and the detected x-line (blue) and base-line (red). The patch is recognized as "Distributed Corupressud Sczxmxxxg". (b) Rectified image, which is recognized as "Distributed Compressed Sensing". (c) Gradients of the image between the two lines. (d) Orientation histogram of gradients.

[18]. The geometric information and the descriptors form the low bit-rate descriptors are sent from the client to the server.

E. Online Search and Verification

The server receives the information of the recognized title text and compressed image features. Typically, spelling check is run on the recognized title to improve the recognition performance. The recognized title is used as query keywords and posted to online search engines such as IEEExplore, CiteSeer, and Google Scolar. Search results from different engines are merged to form a list of candidate matching documents and geometric verification is performed on few top entries in the list. We describe in detail how geometric verification is performed in Section III.

III. Geometric Verification with Text and Image Features

Geometric verification of two image feature sets is performed in two steps. First, the descriptors of one feature set is matched to the descriptors of the other based on a distance criteria. For CHoG, the KL divergence is used. To make sure that the match is unique, a ratio test [5] is applied to the rule out ambiguous matches. Then, after finding the matching descriptors, the locations of these descriptors are passed to RANSAC, where an affine model is estimated.

For document images, two issues affect the geometric verification performance. First, when the user takes a picture of the document title, the view typically does not cover the full page. Thus, when we compare the query image to the candidate image, features that are not within the query view act as clutter. Second, text typically contains repetitive structures, which causes the ratio test to fail.

To address these issues, we use the title text bounding box information. The title text bounding box information is available to us from the query image when we perform the title text detection algorithm. Thus, we can estimate a geometric transformation model between the query image title box and candidate image title box. This estimated model can be used as a loose criteria to help in matching the features. We do so in two different ways which we describe in the following two sections.

A. Feature Cropping

We estimate an affine transformation between the two images using the location information of the two title boxes. The affine model can be used to transform the boundary of one image to the other, showing what parts of the one image are not visible in the other. When performing geometric verification, we use only the features that falls within the covered region, as shown in Figure 4. This reduces the amount of image features considered for matching and also removes the noisy features.

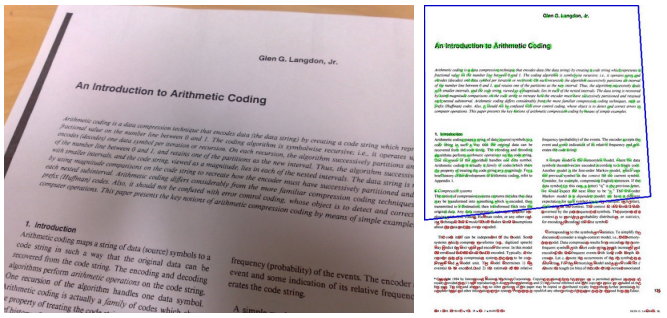


Fig. 4. The query image boundary is transformed on to the candidate document image, showing which parts of the image are covered. Only feature that lie within the region are considered for feature matching.

B. Orientation Check

When estimating a geometric transformation model, the rotation between the two documents is also derived. This information can be used to rule out incorrect matching feature pairs, i.e., the orientation difference between the two matching features should agree with the derived rotation. We add the orientation check within the geometric verification framework in two different locations as shown in Figure 5. The pre-ratio test orientation check helps reduce the number of similarity computations for feature matching. However, it affects the performance of the ratio test because it removes feature pairs. Hence, a conservative threshold is applied. In the second stage of post-ratio test orientation check, we use a tighter threshold to remove the remaining false positives.

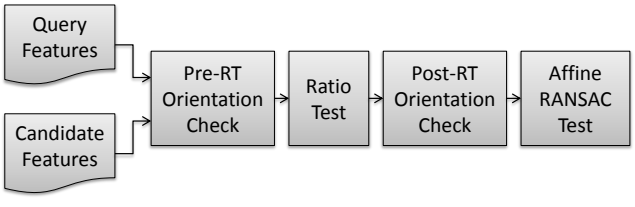


Fig. 5. Orientation check is used before the ratio test with a loose threshold while another orientation check with strict threshold is used after the ratio test in the geometric verification framework.

IV. Experimental Results

In this section, we show the results of evaluating our proposed scheme using a pairwise image matching experiment and a web-scale document retrieval experiement. We use a document image database [19] of 501 query image which contains images of documents taken using a 5MP cameraphone at various view points, Figure 6.

Fig. 6. Samples of the printed document dataset.

A. Pairwise Image Matching

In the pairwise image matching experiment, we use 501 matching document image pairs and 26K non-matching image pairs. For all image pairs, one is from the query document image set while the other is a clean full page document image of the title page. For each image pair, we perform Geometric Verification (GV) to find the total number of feature matches after RANSAC. We use an exhaustive search when performing the feature matching and ratio test with a threshold of 0.85, and estimate an affine model using RANSAC. We use a threshold on the number of feature matches after RANSAC to determine if an image pair is a true match. By varying the threshold, we can obtain the Receiver Operating Characteristics (ROC) curve of different algorithms.

In Figure 7, we show the ROC performance of a basic geometric verification which consists of only the ratio test and RANSAC test. We can see that the basic algorithm already achieves high precision at low false positive rates. If we add in the Feature Cropping (FC), we can see that the performance of the geometric verification improves; the missed true-positive rate is decreased by almost 20%. By further incorporating the Orientation Check (OC), we can reduce the missed true positive rate by more than 50%. Here we use a threshold of $\pi/2$ for the pre-ratio test orientation check and $\pi/4$ for the post-ratio test orientation check.

The timing of the three different types of geometric verification schemes are shown in Figure 8. By using the feature cropping we can reduce the number of features consider for feature matching, thus we can reduce the total time for geometric verification by almost 40%. By further adding in the orientation check, we can reduce the number computations needed for feature matching, reducing the timing further by a half. The overall speed up is 3x.

The orientation information used in the orientation check is quantized and send from the client. To see how quantization affects the geometric verification, we use different quantization steps to run the experiments and shown the results in 9.

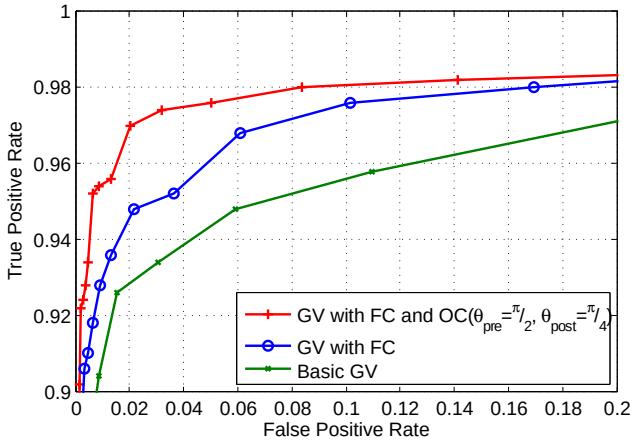


Fig. 7. ROC performance of pairwise document image matching for three schemes.

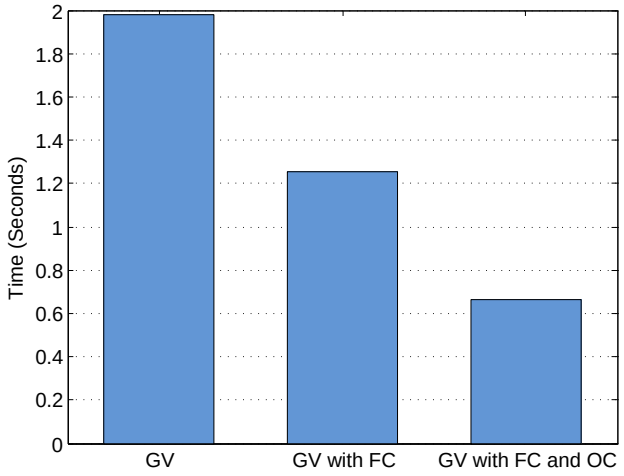


Fig. 8. The timing performance of three difference geometric verification schemes.

B. Web-scale document search

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[20] is used to recognize the text from the patches. The recognized keywords are posted to Google Scholar and CiteSeer for on-line search. Compressed image features, CHoG[21], 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.

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 geometric verification. The recall performance is shown in Fig. 10. A recall of $\sim$76% on the top entry is observed for the system without geometric verification. With geometric verification, we can boost the recall performance of the top entry to $\sim$87%.

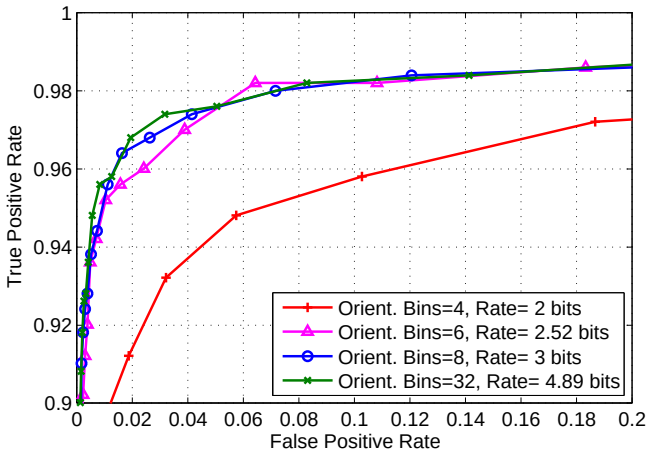


Fig. 9. Geometric verification with quantized orientation.

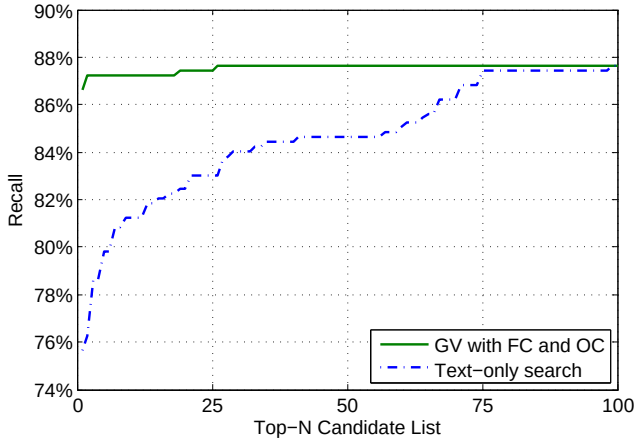


Fig. 10. Recall performance of returned list of the search engine and that after the list has been re-ranked.

V. Conclusions

We propose a mobile visual search system that uses both image and text features. Using a smartphone, a user takes an image of a document. The title text is located and recognized while image features are detected and extracted. We send the title text and the compressed features from the client to the server. The server uses the title text to perform a web-scaled document search and then uses the image features to perform geometric verification. We demonstrate that by incorporating the title text box information within the geometric verification framework, the matching performance can be substantially improved. Although the method requires an additional bit-rate of $\sim$3% of orientation information for the sent features, it reduces the missed true positive matchines up to $\sim$50%. Furthermore, it speeds up the geometric verification up to $\sim 3\times$ and enables verification of more candidate images in the webscale document retrieval framework, resulting in a higher recall of $\sim$87%.

Acknowledgments

The authors would like to thank Vijay Chandrasekhar and Gabriel Takacs for the invaluable discussions and suggestions.

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] D. Lowe, "Distinctive image features from scale-invariant keypoints," *International Journal of Computer Vision*, 2004.

[6] 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.

[7] J. Sivic and A. Zisserman, "Video google: a text retrieval approach to object matching in videos," in *International Conference on Computer Vision*, 2003, vol. 2, pp. 1470–1477.

[8] 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.

[9] 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.

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

[11] S. S. Tsai, H. Chen, G. Schroth, D. M. Chen, R. Grzeszczuk, and B. Girod, "Mobile visual search on printed documents using text and low bit-rate features," in *International Conference on Image Processing*, 2011.

[12] M. Fischler and R. Bolles, "Random sample consensus: a paradigm for model fitting with applications to image analysis and automated cryptography," *Communications of ACM*, vol. 24, no. 1, pp. 381–395, 1981.

[13] S. C. Hinds, J. L. Fisher, and D. P. D'Amato, "A document skew detection method using run-length encoding and the hough transform," in *International Conference on Pattern Recognition*, 1990.

[14] 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.

[15] H. Chen, S. S. Tsai, G. Schroth, D. M. Chen., 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.

[16] J. Canny, "A Computational Approach to Edge Detection," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, 1986.

[17] 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.

[18] 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.

[19] "Document images," http://msw3.stanford.edu/∼sstsai/PaperSearch/.

[20] "Tesseract OCR engine," http://code.google.com/p/tesseract-ocr/.

[21] "Chog release," http://www.stanford.edu/people/vijayc/chog-release.zip.