

Mobile Visual Search, Linking Printed Documents to Digital Media

Xu Liu, Jonathan J. Hull, Jamey Graham, Jorge Moraleda, Timothee Bailloeuil
 Ricoh Innovations, Inc.
 California Research Center
 Menlo Park, CA
 650-496-5704
 liu @rii.ricoh.com

ABSTRACT

In this paper we present a visual search technology that enables camera phone users to find related digital media by capturing an image of printed documents. We have implemented two architectures: an integrated system which runs solely on the phone and a networked system which recognizes a submitted image on a server. Although the networked system offers a large database capacity, we argue for the integrated system because it enables real time recognition on the device with smooth user interaction. We have implemented a highly efficient feature extraction and matching algorithm targeting resource-constrained mobile devices. The advantage of our system is the complete integral solution on the phone including a language-independent feature extraction and an efficient database lookup, which provides instant response. Experimental results show that our recognition system has an overall error rate of less than 1% and recognition time less than 1s. Our system runs on iPhone and on Windows Mobile phones. A user study shows it takes 5.3 seconds for a user to complete one visual search query running our software on the HTC8282 Windows Mobile phone.

1. Introduction

Visual search has been extensively researched in recent years[1], largely due to the rapid development and penetration of the camera phones. Mobile augmented reality[5] and outdoor coordinate systems[6] have been created with visual search technology. The ultimate goal of visual search is to provide a link between the physical world and the digital world. Visual search allows a user to retrieve digital information by pointing a camera phone at an object. The choice of object, data type and action associated with it can enable a number of commercial applications. TABLE-1 lists five recent visual search products that are publicly available.

TABLE 1 Mobile Visual Search Products

Company	Platform	Product	Targets
Google	Android	Goggles	Landmark/Book/Artwork/Grocery
Amazon	iPhone	Snaptell	Book/DVD/Game covers
Nokia	Symbian	PointAndFind	Landmark/Barcode/Movie poster
Kooaba	iPhone	Kooaba	Book/DVD/Game covers
oMoby	iPhone	oMoby	General Objects

These visual search products all employ the client-server architecture shown in Figure 1. An advantage of such an architecture is its scalability: the database server could in theory host an unlimited number of images. However, this architecture

has two major disadvantages: network latency and lack of user feedback.

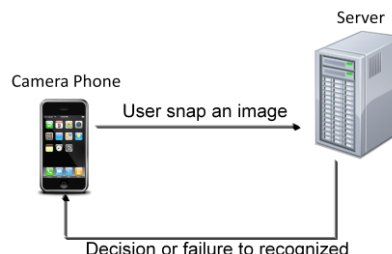


Figure 1. Traditional visual search architecture

Under the client-server architecture, a user has to snap a picture, submit it to the server and then have it recognized. Maintaining an upstream IP connection is still challenging on the phone network. The user might wait too long and lose patience. The recognition could fail for many reasons, for example, the image is too dark, too blurry or simply because the captured object is not in the database. When recognition fails, the user has to snap another picture and submit it again. After a couple of rounds, the frustrated user can believe the system does not work.

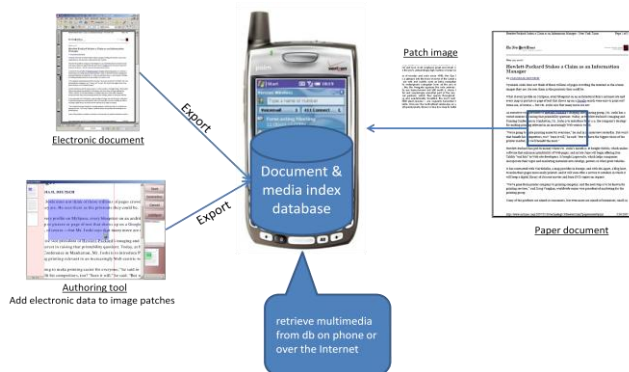


Figure 2. Integrated visual search architecture

In our system, we target a different type of object with a different architecture. Our visual search technique focuses on paper documents, including newspapers, books, magazines and other printed materials. Using our technology, any page or paragraph of the printed material may be linked to online digital media. For example, a picture of a sports column article may return with the video of the game discussed in the article. For document image recognition, Nakai et al.[4] extracted affine invariant features from the bounding boxes around neighboring words but it only works with Latin languages. Similar systems used OCR[2] and server-based recognition[3]. Our system does not rely on OCR

and is language-independent. In order to avoid network latency we introduced the integrated visual search architecture shown in Figure 2. This architecture enables visual search without a network connection because the database is hosted on the handset. Capture and submission is no longer required. With our optimized visual search algorithm, recognition is performed in real time on the camera buffer. Our user simply hovers the camera phone above the document to complete visual search, like using a barcode reader.

2. Implementation on the Phone

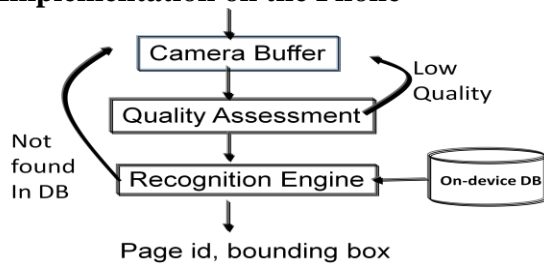


Figure 3. System structure on the phone

Our system differs from existing products (TABLE-1) by continuous recognition of camera buffer frames instead of user submitted snapshots. Recognition is a relatively expensive step (approximately 1s per frame) and there are many "garbage frames" due to motion blur or random objects in view. Therefore it is not optimal to send every camera frame to the recognition engine. We insert a Quality Assessment (QA) module before sending a frame to the recognition engine. As illustrated in Figure 3 the QA module gives very quick (33fps) estimation of the image quality and sends only high quality frames to recognition since they have a better chance to be recognized. According to our experiment, the QA module can filter 60% of the frames that are not recognizable.

The complete recognition algorithm is implemented in C on the iPhone and on a Windows Mobile phone (HTC 8282). Indexed documents are represented with binary tables that are designed for fast download. We tested the current software on a collection of 140 French newspaper pages. Each indexed page contained approximately 4000 features and 23.8 MB were used for the index tables on the phone. With patch images captured on the phone at 320x240 resolution, 0.5~1.2 seconds are needed to recognize each image.

3. Preliminary User Study

We introduced our visual search system to five new users who had no prior experience with visual search. The user was given 50 pages of a French newspaper, a subset of 140-page newspaper database on the HTC8282 phone. The user was asked to perform at least ten visual searches. They could choose any page to search but were instructed to finish the search as quickly as possible.

The time spent on each visual search query is shown in Figure 4. The timer starts as the camera starts rolling and stops when a page is recognized. As shown in Figure 4, a new user requires more time in the first 1~3 trials than in subsequent attempts because they are unfamiliar with the system. But they can quickly learn and adapt and the recognition time stabilizes at 4~6 seconds. Because recognition happens in real time without a key press, the learning cost is very low. Compare this to the client-server architecture where the capture-submit-failure-retry cycle may take

minutes. In our architecture the whole cycle happens within a fraction of a second. Moreover, we display the QA score to assist the user in aiming the camera at areas with richer texture. This helps our users train themselves. The average recognition time in Figure 4 is 5.3 seconds per visual search query.

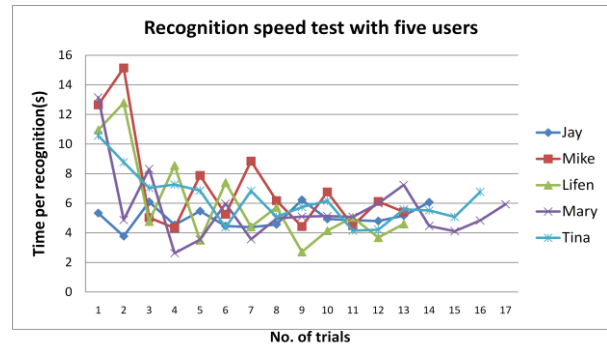


Figure 4. User study with recognition Time

4. Conclusion and Future Work

In this paper we introduced a visual search system that runs solely on camera phones. This system searches for an electronic document using a patch of the document image captured from the phone camera. No OCR is required since the recognition is based on image features. It runs in real time with the preview camera buffer and thus provides a barcode reader-like user experience. We estimate the quality of an input image frame before sending it to recognition and avoid wasting time on low quality frames. Novice users can perform visual search using this system at 5.3 seconds per query. Our future work includes reduction of the database size and enlarging its capacity. At the same time we will further optimize the recognition engine and accelerate the visual search.

References

- [1] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman, "Object retrieval with large vocabularies and fast spatial matching," Proc. of the IEEE Conference on Computer Vision and Pattern Recognition, 2007.
- [2] X. Liu and D. Doermann, "Mobile Retriever: access to digital documents from their physical source," *International Journal on Document Analysis and Recognition*, vol. 11, 2008, pp. 19-27.
- [3] B. Erol, E. Antunez, and J.J. Hull, "HOTPAPER: multimedia interaction with paper using mobile phones," *Proc. of the 16th ACM Intl. Conf. on Multimedia*, Vancouver, Canada, 2008, pp. 399-408.
- [4] T. Nakai, K. Kise, and M. Iwamura, "Use of affine invariants in locally likely arrangement hashing for camera-based document image retrieval," *Lecture Notes in Computer Science (7th International Workshop DAS2006)*, vol. 3872, 2006.
- [5] D. Wagner, G. Reitmayr, A. Mulloni, T. Drummond, and D. Schmalstieg, "Pose tracking from natural features on mobile phones," Proc. of the 7th IEEE/ACM Int. Symp. on Mixed and Augmented Reality (Sept. 15 - 18, 2008), pp. 125-134.
- [6] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W.C. Chen, T. Bismptgiannis, R. Grzeszczuk, K. Pulli, and B. Girod, "Outdoors augmented reality on mobile phone using loxel-based visual feature organization," *ACM International Conference on Multimedia Information Retrieval (MIR'08)*, Vancouver, Canada, Oct. 2008.