

# Photo-to-Search: Using Multimodal Queries to Search the Web from Mobile Devices

Xin Fan<sup>\*1</sup>, Xing Xie<sup>2</sup>, Zhiwei Li<sup>2</sup>, Mingjing Li<sup>2</sup>, and Wei-Ying Ma<sup>2</sup>

<sup>1</sup>University of Science and Technology of China

Department of EEIS  
Hefei, 230027, P.R. China

van@mail.ustc.edu.cn

<sup>2</sup>Microsoft Research Asia

5F, Sigma Building, No.49, Zhichun Road. Beijing,  
100080, P.R.China

{xingx, zli, mjli, wyma}@microsoft.com

## ABSTRACT

Nowadays, mobile phones with the digital camera are getting more and more popular. With necessary technologies, they are possible to become a powerful tool to search the Web on the go. Most Web search engines only support text queries. Therefore, users have to convert their information needs into words. However, it is sometimes difficult to describe the needs in text and the text input is inconvenient on small devices. To solve the problem, we propose a system named Photo-to-Search which allows users to input multimodal queries. Particularly, we study queries with captured images and optional text messages in this paper. For example, the user can simply take a photo of the flower and input a few terms like “flower”. Textually relevant Web images are retrieved according to the query terms. Afterwards, the snapped picture is compared with these images by the CBIR (Content Based Image Retrieval) method. According to the context of the visually similar images, related key phrases are extracted. Finally, the search results are returned in multiple forms. Our system can also search for very similar images on the Web, such as movie posters or photos of film stars, to find related information. Experimental results on the large scale data showed our system achieved satisfactory efficiency and performance.

## Categories and Subject Descriptors

I.4.9 [Image Processing and Computer Vision]: Applications

## General Terms

Algorithms, Performance, Human Factors

## Keywords

Multimodal interactions, web image search, mobile search, content based image retrieval, duplicate image detection

## 1. INTRODUCTION

Current mobile search engines like Google Mobile [7], Google SMS [8] and Yahoo! Mobile [21] are mostly using text-based input and output. However, users sometimes feel difficult to find suitable words to describe their information needs, especially

facing present unintelligent search engines. On the other hand, owing to the constrained input/output modalities, it is still inconvenient to use the search service on mobile devices. In our opinion, instead of the current flat query mode, camera phones can support richer and hybrid queries, not only text but also images, voices and even videos.

Mobile phones with the embedded camera already have millions of users and their growth potential is enormous. Therefore, focusing research into technology and user interaction in mobile image is increasingly important. Moreover mobile cameras are considered becoming a promising HCI manner for mobile devices, just like the emergence of the mouse for desktop computers. However, the value of camera phones on daily information acquisition has not been sufficiently regarded. This situation motivates us to build up a system to make the camera phone turn into a powerful tool to search the Web in daily lives.

In this paper, we propose a system named Photo-to-Search to provide an easy and effective approach for mobile users to acquire information on the go. The information query process can be described in an exemplary scenario. When a user is attracted by an outdoor advertisement and desires more information about the product, he can take a photo of the advertisement poster and then send the image to our system. Our system will find copies of duplicate or very similar images in volumes of Web images and return related web pages for reference. Afterwards, he is curious about some beautiful flowers in the roadside terrace. So he sends the flower photo with a word “flower” to our system. Since there are no duplicate Web images, the captured image is used to retrieve visually similar ones within a set of Web images related to the flower. From the web pages where these images are located, we would extract relevant information and return several illustrative entries from online encyclopedias or provide the most similar images with links to original web pages.

In our preparatory user study, when users provide images as a query, the information needs fall into following three categories:

- ◆ Expected results are exact ones or the copies of the query image, for example, the capture image is a famous painting, a product casing or a book cover.
- ◆ Expected results are with similar visual features to the query image, for example, the query image is a certain flower, an animal or a building.
- ◆ The query image contains certain characters or signs for the purpose of character recognition like OCR.

Permission to make digital or hard copies of all or part of this work for personal or classroom use is granted without fee provided that copies are not made or distributed for profit or commercial advantage and that copies bear this notice and the full citation on the first page. To copy otherwise, or republish, to post on servers or to redistribute to lists, requires prior specific permission and/or a fee.

MIR '05, November 10-11, 2005, Singapore.

Copyright 2005 ACM 1-59593-244-5/05/0011...\$5.00.

---

\* This work was performed at Microsoft Research Asia.

In this paper, we focus on the first two kinds of information needs. The issue on the first scenario is usually able to be addressed by the approaches of duplicate/near-duplicate image detection [3][11][12] while the second category is much more related to Content Based Image Retrieval (CBIR) approaches. As far as we investigated, most of work [9][14][18][22][23] concerning searching with pictures captured by build-in mobile cameras casting their tasks into a CBIR context. Considering the computational costs and the limited accuracy of CBIR methods, they are generally designed to aim at a specific task and based on a small-scale local image database or an amount of Web images from several predefined domains. To carry out visual queries on a large-scale data set which is more approximate to the real World Wide Web, the computation time and the algorithm complexity are critical considerations. For such large scale data, traditional CBIR methods are computationally inhibited and not accurate enough.

Therefore, instead of pure image queries in most existing work, our system support multimodal queries, i.e. the query can be images with hint words attached. Firstly, identical image detection is performed for the information need of image copies. A hash signature indexing approach is used to reduce the computational cost. If no identical images are found, a hybrid image matching approach is designed for the information need of similar images. According to the attached words, a set of textually related images are collected by a text-based Web image search engine [4]. The design of this image set is partly inspired by the idea of bootstrapping database in [22]. Then the CBIR method is employed to find visually similar images in this set and key phrases are extracted from the web pages where these images are located. In fact, the pure image query can also be supported in our system if only identical image search was performed. We will demonstrate in Section 4 that this ensemble approach ensures fast search speed as well as accuracy. In summary, our system mainly comprise four components: identical image search, hybrid image matching based on text and image content, key phrase extraction and search result presentation. Details of these components are respectively presented in Section 3.

## 2. RELATED WORK

Text-based Web image search is the usual image search method provided by common search engines like Google, Yahoo! and MSN. Generally, the textually related images are retrieved according to the matching degree between the query words and the keywords in the text features of Web images. However, the contents in an image are multiple and some queries are difficult to be characterized by a few words. Therefore, content based image retrieval approaches [6][17] have been widely researched in the last decade. Early work on CBIR is mostly based on global descriptors, while in the recent work much effort has been drawn to the research on local descriptors and their associated semantics. There are mainly two types of approaches derived from this aspect: one [2] is to perform image segmentation and then link the segmented image regions to semantic conceptions. In the other type of approaches [16], salient point/region detection is in place of the image segmentation and feature vectors or local descriptors are derived around the salient areas.

As a matter of more and more attention to image copyright protection [11][12] and the consideration of removing the redundant copies of consumer digital images [10], duplicate or

near-duplicate image detection has been extensively studied in recent years. Although the duplicate/near-duplicate detection can be regarded as one aspect of image retrieval, it lays more emphasis on searching for copies of a given image instead of the similar images as traditional CBIR methods do. The ordinal gray-value measure [12] is a common method for duplicate image detection. Moreover, the approaches based on salient point detection and local descriptor extraction [10][11] are also employed in many researches and achieve better results. The main shortcoming of those methods is relatively expensive computation.

A common problem for the visual search is the difficulty for getting a query image, but it is much easier for camera phone users. A few systems have been proposed taking advantage of built-in cameras on the mobiles. Most of them [9] [14][18][23] are based on a limited and specific image database and only accept pure image queries. We notice in the work of IDEixis [22] more Web images can be retrieved by a text-based image search engine like Google. However, its initial image database for CBIR is still specialized and limited in scale. This solution may not be suitable for a generic purpose search on large scale data. Comparatively, our system differs significantly in several aspects. First, the query in our system can be multimodal inputs such as an image with some hint words. Second, the duplicate image detection is introduced. If no duplicate images are detected, we will retrieve a textually related image set instead of using a specific image database. Third, in our system, we focus on more extensive applications and large scale data. Therefore, the search efficiency is also a critical consideration in addition to the precision. Fourth, though CBIR's performance is not good enough, the performance of our hybrid image matching method is much better due to the data-driven mechanism based on a large amount of Web images and the adoption of key phrase extraction.

## 3. SYSTEM ARCHITECTURE

In this section we introduce the implementation details of our system. To illustrate the whole search process of our system, a flowchart is given in Figure 1.

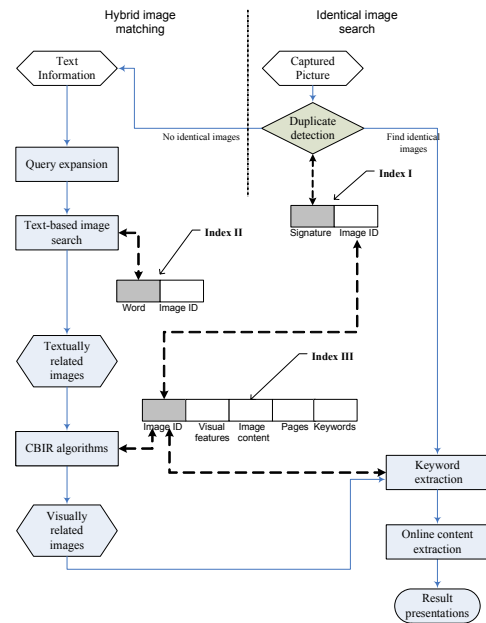


Figure 1. The flowchart of Photo-to-Search system

The multimodal query of a captured picture and some hint words is supported in our system currently. For the captured picture, we first carry out the identical image search. If identical images are found, a few key phrases will be extracted from the related web pages for obtaining further information. In Figure 1, the part on the right of the dashed line illustrates this process. If no identical images are found, we will turn to a hybrid image matching module to find textually and visually relevant images. Afterwards, key phrase extraction is carried out to acquire further information.

### 3.1 Identical Image Search

Before dealing with the image query for identical images, we need to extract features of all images in our database and construct the index for fast search.

As we have mentioned in Section 2, the common approaches for duplicate/near-duplicate image detection can be roughly separated into two categories by the extracted features. One is based on the ordinal descriptors which is mostly the intensity value. The other is the salient point/region based approach. Although the second approach shows better robustness, its computational cost is too expensive for millions of images in our system. Therefore we design a light-weight measure method based on the ordinal intensity value in this module. In previous work [12], the measures based on DCT coefficient features in the image intensity channel achieve satisfying retrieval performance in the precision, recall, robustness and computational cost. Therefore, we firstly extract similar features from images in our database.

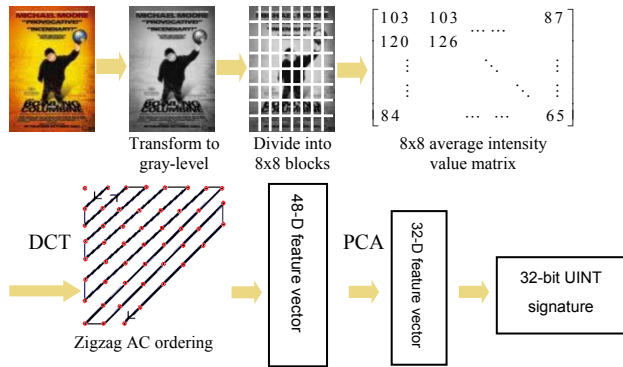


Figure 2. Hash signature generation for an image in identical image search

Since hue and saturation of photos captured by built-in cameras would often shift due to the non-professional exposure, aperture, etc., we firstly transform the input color image to a gray-level one and divide it into 8x8 blocks (image patches) equally. Each block is labeled by the average intensity value in it, which can be represented as a matrix  $I_{ij}$  as follows:

$$I_{ij} = \frac{\sum_{x=0}^{w-1} \sum_{y=0}^{h-1} Int(x, y)}{w \cdot h}, \quad i = 0, \dots, 6, 7, j = 0, \dots, 6, 7 \quad (3.1)$$

where  $Int(x, y)$  is the intensity value at point  $(i, j)$ .  $w$  denotes the block width and  $h$  denotes the block height. Then the 2-D DCT for matrix  $I_{ij}$  is performed and the 8x8 magnitudes of DCT coefficients matrix  $D_{ij}$  is generated. It can be deduced that the magnitudes of DCT coefficients of  $I_{ij}$ , its horizontal flipped version  $I'_{ij}$  and its 180 degrees rotated version  $I''_{ij}$  are all identical.

Thus, the matching in these cases can be achieved at one time.

Since the value  $D_{00}$  called DC coefficient is the average of all 64 values in the intensity matrix, we omit it to eliminate influences of the commonly global variance of luminance in the photo. The remaining AC coefficients are ordered into a "zigzag" sequence. The first 48 AC coefficients in lower frequencies are collected as a 48-dimension feature vector  $A_m$ .

Next we transform the 48-dimension AC coefficients feature vector  $A_m$  into a 32-dimension feature vector  $Y_n$  using Principal Component Analysis (PCA). A training set with 5,500 web images is used to train the transform matrix  $P^T$ . It has been tested that a greater amount of training images would not affect the result matrix much. Accordingly, we update all the image feature vectors by the below formula:

$$Y_n = P^T A_m \quad (3.2)$$

with  $P^T P = I_n$ . Here  $P$  is an  $m \times n$  matrix whose columns are the  $n$  orthonormal eigenvectors corresponding to the first  $n$  largest eigenvalues of the covariance matrix  $\sum A_m$ . The whole process is illustrated in Figure 2.

Instead of comparing vector distances one by one, we design a hash function to map the 32-dimension feature vector into a hash signature for each image. We adopt a 32-bit unsigned integer *signature* to store the hash signature. The hash function is given in the following pseudo codes in Figure 3.

```

Input: feature vector  $Y = \{y_i\}$ ;  $i$  ranges from 0 to  $n-1$ 
For each element  $y_i$ 
    If  $y_i > 0$ , the  $i$ th bit of signature is set to 1
    Else the  $i$ th bit of signature is set to 0
End
Output: signature

```

Figure 3. Algorithms for hash signature generation

Thus, we denote the 32 bits of an unsigned integer to distinguish the identical images. Although one signature may theoretically be mapped into several different images, most of the signatures have the single correspondence as the evaluation in Section 4. So in most cases, only one copy of image will be retrieved by our approach.

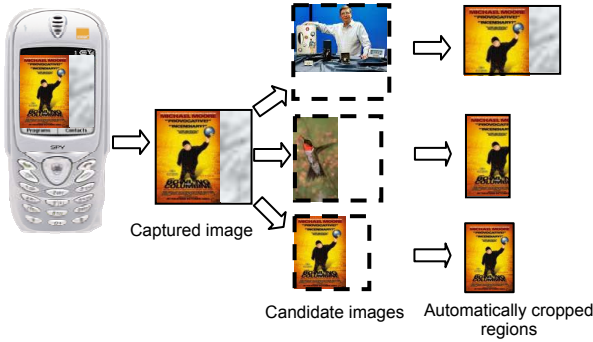
Since the images in our database are indexed by the integer signatures, the complexity of the searching algorithm is only logarithmic to the total number of images in the database. We will give the evaluation to prove its effectiveness in Section 4.



Figure 4. Examples of the detected image copies

For the query image that a user submits, the 8x8 DCT coefficient feature is extracted in the same way and the signature is derived from the transformed 32-dimension feature vector. If the candidate image and the query image are with the same signature,

they are regarded as identical ones as shown in Figure 4. In order to avoid costly image alignment, an appropriate region without unnecessary margins is needed to be cropped in advance. The cropping is easy in some new powerful mobiles like PDAs and smart phones with image editing tools. For the devices without such tools, we adopt a simple approach for aligning. The user only needs to adjust the camera position and the zoom ratio to align both of the top and left sides and either of the bottom or right sides of the desired object with the camera view zone, as shown in Figure 5. Since the width and height of each candidate image in the database are known, we can automatically crop an appropriate region in the query image for matching according to the aspect ratio of the candidate image.



**Figure 5. A simple solution to the alignment of the query image and candidate images**

## 3.2 Hybrid Image Matching

When the identical image search does not find the identical image for the captured image, we first use the text information to find the textually relevant images to build up an image set. The visually relevant images are selected by the image matching process based on the CBIR approach. In order to reveal the underlying information related to the query, key phrases are extracted from the web pages where the selected images are located. Accordingly, we can find the further information from the online encyclopedias or other resources by the key phrases.

### 3.2.1 Textually related image retrieval

It is often difficult for the users to give complete descriptions of the sought information. Especially on the mobiles, users often phrase their hint information only in one or two words. These not well-defined query words only provide ambiguous descriptions. Thus this insufficient information may often result in poor retrieval results. To address this problem and reveal the information beyond the few query words, we expand input words in the query to derive a series of new query words.

Basically, we perform the query expansion using the lexical-semantic relations of the vocabularies [19]. Our basic expansion process is based on Wordnet [13], a large general-purpose lexical system. It covers a vast amount of nouns, verbs, adjectives and adverbs from English language. These words are fundamentally organized in synonym, called *synsets*, and *synsets* are organized by the lexical relations defined on them.

If there are meaningless words in the original input words, such as “of”, “the” and “one” defined in our stop words list, these words will be removed first. The remaining words are transformed by the Porter stemming algorithm [15]. In the stemming process, the

words are represented by their stems. For example, “flowers”, “flowered” and “flowering” are all transformed to their stem “flower”. According to Wordnet, the words listed in the Synonyms and Hyponyms (child) synsets are added in the new query word set  $\{QT_i\}$ . For example, the word “flower” may be expanded to a set  $\{flower, bloom, blossom, heyday, efflorescence, flush, peony, lesser celandine, pilewort, Ranunculus ficaria, anemone, windflower \dots\}$ .

For the specific applications, some knowledge based query expansion approaches [1] can also be employed in addition to the basic lexical approach. Correspondingly, domain-specific knowledge base is needed and the artificial intelligence techniques can be used to provide and supplement the scenario-specific expansion terms.

If we consider this expanded word set as a text feature vector, we can match this feature vector by the word index (Index II in Figure 1) in the Web image search system [4]. In other words, we can send all the words  $QT_i$  in expanded set to the search system in the “OR” relationships. Consequently, the images which are more related to the word set  $\{QT_i\}$  will be given higher ranks. A certain number of images are selected according to the ranks. The selected textually related images constitute an image set  $\{TI_1, \dots, TI_{M-1}, TI_M\}$  from which we will find the visually related images.

### 3.2.2 Visually related image retrieval

We search for the visually similar images by measuring the similarities between the query image  $QI$  and images in the set  $\{TI_1, \dots, TI_{M-1}, TI_M\}$  based on the CBIR approach. The common descriptors for CBIR mainly include color features, texture features, shape features etc. We notice that texture features are not discriminative enough for the pictures captured by mobile cameras. In addition, considering the computational cost, we extract the following three kinds of features from all the images in our system for the image matching:

- ◆ 64-element RGB color histogram. It describes the color distributions in the RGB space.
- ◆ 64-element HSV color histogram. The representation in the HSV color space accords with human perception better than in the RGB space.
- ◆ 192-element Daubechies' wavelets coefficients. It performs better in capturing coherence of image, object granularity, etc.

Accordingly, the values in RGB and HSV color histograms are described as a 64-dimension feature vector  $F_{RGB}$  and a 64-dimension vector  $F_{HSV}$ . The Daubechies' wavelets coefficients in the lower frequency bands and their variances are stored as a 192-dimension feature vector  $F_{Daub}$ . The feature vector values from the whole image database (about 12 million images) are stored in sequence. To facilitate the lookup and reduce the I/O operations, we index these features (Index III in Figure 1) by image IDs.

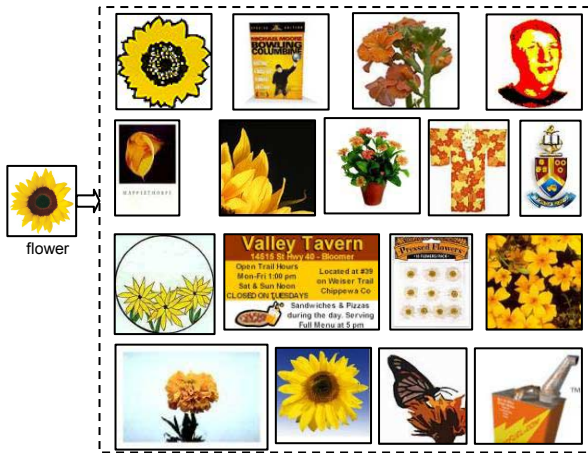
Weighting the performance improvement against the computational consumption, we haven't employed the techniques of dimension reduction and transform like Karhunen-Loeve Transform and Principle Component Analysis before measuring distances of the above feature vectors. The visual distance between an query image  $QI$  and an image in the textually related image set  $\{TI_1, \dots, TI_{M-1}, TI_M\}$  is expressed by the weighted sum of L1 norm of Minkowski matrices between the corresponding feature vectors in Equation (3.3). The distance value of each

feature is normalized to a fixed range (0, 1) to eliminate modality-dependent amplitude differences.

$$D_j = w_{RGB} \mathfrak{N}(\|F_{RGB}^{query} - F_{RGB}^j\|_1) + w_{HSV} \mathfrak{N}(\|F_{HSV}^{query} - F_{HSV}^j\|_1) + w_{Daub} \mathfrak{N}(\|F_{Daub}^{query} - F_{Daub}^j\|_1), \quad j = 1, \dots, M \quad (3.3)$$

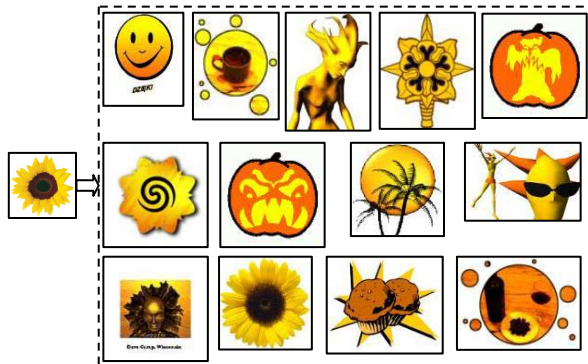
where  $F_{RGB}^{query}$ ,  $F_{HSV}^{query}$  and  $F_{Daub}^{query}$  are feature vectors of the query image  $QI$  and  $F_{RGB}^j$ ,  $F_{HSV}^j$  and  $F_{Daub}^j$  are the feature vectors of the image  $TI_j$  in the set  $\{TI_1, \dots, TI_{M-1}, TI_M\}$ .  $\mathfrak{N}(\bullet)$  is the normalization operator. We adopt constant weights:  $w_{RGB} = 0.3$ ,  $w_{HSV} = 0.5$ ,  $w_{Daub} = 0.2$  for these three kinds of feature vectors in general tasks.

A certain amount  $L$  of images with smaller visual distances are selected as visually and textually related images  $\{VTI_1, \dots, VTI_{L-1}, VTI_L\}$  as shown in Figure 6.



**Figure 6. Textually and visually related images retrieved from 12 million web images by our method**

In most previous work on information search using mobile cameras, only CBIR based matching methods are adopted to find similar images to draw the further information. However, in our system, traditional CBIR based matching is not applicable in regard of computational cost and precision for large numbers of images. Since the CBIR based matching is only implemented on a small amount of textually related images, our solution with multimodal inputs can achieve higher accuracy and faster operation speed in the large scale data. It can be illustrated in an instance of a query image of the sunflower.



**Figure 7. Retrieved images from 12 million web images only by CBIR methods**

In Figure 7, the retrieved images by CBIR method are mainly not semantically relevant to the sunflower or even flower. Most of them are simple decorative graphics since these kinds of images are in the great majority in the web images. Although some visually related images may be found, for example a nearly same sunflower picture, we found there is not any text information related to sunflowers in the web page where the picture is located. Therefore, supposing we have no knowledge of the query image “sunflower”, still no further information would be provided in the retrieved results.

### 3.3 Key Phrase Extraction

CBIR algorithms suffer from the big gap between low level image features and high level semantic concepts. The correct images may not be the top ones in the retrieved image list, while we cannot present too many images to mobile users due to the low bandwidth and small displays of mobiles.

In order to address the above difficulties and reveal the underlying useful information, we carry out key phrase extraction for each web page associated with images and store the key phrases with the image (Index III in Figure 1). The key phrases are extracted offline based on analyzing the statistical and structure characteristics of words in a web page. We will introduce its detail later in this section.

For the retrieved visually and textually related images  $\{VTI_1, \dots, VTI_{L-1}, VTI_L\}$ , a predefined number  $n$  of image key phrases are selected for each image  $VTI_i$ ,  $i = 1, \dots, L$ . A few global key phrases are then sought out from the  $n \times L$  image key phrases according to the occurrence frequencies. Just as we depict in Section 4, in most cases there is only one version of identical image for each query image. Thus for retrieved images from the identical image search module, we skip the second step and simply return the image key phrases.

#### 3.3.1 Image Key phrase selection

Before selecting key phrases, we set out to extract the image contextual information according to the structural layout, which is presented in the HTML DOM (Document Object Model) tree [20]. We retrieve the content between the corresponding explicit separators of the image as the surrounding text. All retrieved surrounding text segments are put together into an article. The following process is based on this article and page key phrases will be extracted from it.

Many concepts are represented not by one term but by several terms. For example, for the flower name “sweet rocket”, any one of the two terms is not related to this kind of flower. We define this continuous term segment as a “phrase”. We set a max number of terms for each phrase, denoted as  $t$ . Here we adopt  $t = 4$  by default. All key phrase candidates will be extracted from the above article using following steps:

- ◆ Phrase enumeration. We enumerate all kinds of phrases with length smaller than  $t$ . The occurrence times of each phrase in the article are calculated.
- ◆ Boundary term check. We remove some words in the phrase which can not be boundary words according to a list of stop words for boundaries, for example, the words “of” and “are” will be removed from the phrase “of mobile devices are”.
- ◆ Phrase recounting. In the first step, we have double computed the number of occurrences of each phrase which

containing other phrases. For instance, the count of phrase “A B” contains the count of “A B C”. We adjust the overlapped counts to get independent occurrence of each phrase.

- ◆ Stop word removal. We define a list of stop words which cannot be key phrases semantically, for example, numbers, “jpg” and most of the pronouns. If a phrase is in the stop word list, it will be removed from the candidates.

### 3.3.2 Image Key phrase ranking

After acquiring validated phrases from a web page, we rank the page key phrases by several statistical features and structural features.

As to statistical features, we adopt a traditional measure called Term Frequency-Inverse Document Frequency (TF-IDF) and introduce the Mutual Information (MI).

The TF-IDF is defined as follows:

$$tf-idf_i = \frac{n_{id}}{n_d} \log \frac{N}{n_i} \quad (3.4)$$

where  $n_{id}$  is the number of occurrences of a phrase term  $i$  in article  $d$ ,  $n_d$  is the total number of terms in article  $d$ ,  $n_i$  is the number of articles that contains term  $i$  and  $N$  is the number of articles in the whole database. We calculate the  $tf-idf$  for each phrase as the average  $tf-idf$  value of all terms within the phrase.

In addition to term-level statistics, we define a phrase-level feature Mutual Information [5] ( $MI$ ) for each phrase  $P$ . We require only phrases with  $MI$  larger than a threshold can be selected as key phrase candidates to filter some phrases which are not usually used.

Furthermore we use structural features to consider the importance of the page segment where the phrase is located. Phrases locating in an important segment will have additional scores. In our definitions, the importance of the page segment is related to its visual characteristics. We count this factor by the feature Visualization Style Score ( $VSS$ ) in Equation (3.5), supposing the maximal  $tf-idf$  value among all phrases in a web page is  $tf-idf_{Max}$ .

$$VSS(P) = \begin{cases} tf-idf_{Max}, & \text{if } P \text{ is in title, alt text or meta;} \\ \frac{1}{4} tf-idf_{Max}, & \text{else if } P \text{ is in bold;} \\ \frac{1}{8} tf-idf_{Max}, & \text{else if } P \text{ is in a large font;} \\ 0, & \text{otherwise.} \end{cases} \quad (3.5)$$

where  $VSS(P)$  is the  $VSS$  score for phrase  $P$ .

Thus, we propose two statistical features  $\{tf-idf, MI\}$  and one structural feature  $\{VSS\}$  for page key phrase extraction. The three features are combined using a linear formula:

$$Y = b_0 + \sum_{j=1}^3 b_j X_j \quad (3.6)$$

where  $X = \{tf-idf, MI, VSS\}$ . Coefficients  $b_0, \dots, b_3$  are empirically determined by experiments. In our system we select  $n=5$  key phrases with highest score  $Y$  for each image and save them in the index file. A specified number of global key phrases are then sought out from the  $n \times L$  image key phrases from  $L$  matched images according to the occurrence frequencies.

## 3.4 Result Presentation

We notice that there is often only a single identical image, while the number of images retrieved by the hybrid matching method is much greater. Therefore, in our system, the results are presented respectively.

- ◆ For the results detected by identical image search module, the returned page consists of the detected images and page key phrases. The images are with links to original web pages and key phrases are linked to the entries from the online encyclopedia or traditional text search engines.
- ◆ For the results from hybrid image matching, there are two kinds of presentation modes are provided. One is relevant images with links to original pages and the other is global key phrases with links to entries of the online encyclopedia and top relevant pages from the traditional search engine.

## 4. EXPERIMENTS

We carried out a number of experiments to evaluate our system in the aspects of efficiency, effectiveness and usability. The evaluations were done respectively for the identical image search module and the hybrid image matching module.

### 4.1 Settings

Our images and web pages were from the work [4] on the web image search, which consisted of about 12,000,000 images and 28,000,000 web pages. The too small, banner-like and porn images had been removed after being crawled. We implemented a prototype server-side program to receive the query from an email address by the POP3 protocol and to process the query. The prototype was running on a PC with an Intel Pentium 4 3.0GHZ CPU, 2G RAM and MS Windows 2003 Sever operating system. The returned results were stored in a specified folder on the PC. A demo explorer was implemented on the PC to evaluate the usability in the small display, which was with the two-direction scroll bars and the same 320x240 pixels display size as common pocketPC phone screens. A Dopod 818 Pocket PC phone was used to capture photos in our experiments.

### 4.2 Efficiency

We count the average time consumption for one query by 100 samples. We count the process time after queries were received and before the relevant results were sent to users. The durations were recorded stage by stage and no special code optimization is done in each step.

#### 4.2.1 Efficiency of the identical image search

The index of all the signatures in our database was loaded in the RAM before the identical image search. The binary search algorithm was used to find the equal signature and the corresponding image ID.

**Table 1. Time consumption in identical image search**

Module	Average Time (ms)
Signature generation for the query image	16
Searching identical images	<1
Page key phrase extraction	1
Total	18

### 4.2.2 Efficiency of the hybrid image matching

In the hybrid image matching module, only basic query expansion was adopted based on Wordnet. For each query, we retrieved 6,500 textual related images from which 700 visually related images were selected.

**Table 2. Time consumption in hybrid image matching**

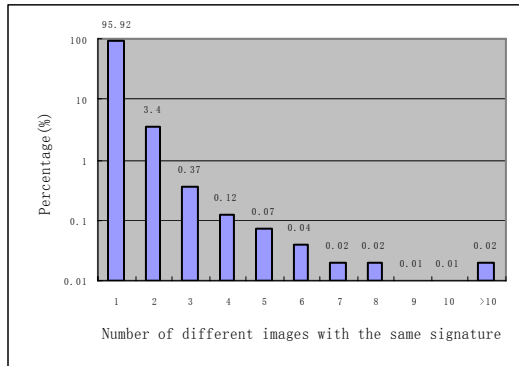
Module	Average Time (ms)
Query word expansion	109
Textually related image retrieval	485
Feature extraction for the query image	188
Visually related image retrieval	516
Page and global key phrase extraction	610
Total	1908

As to the identical image search module, the time consumptions of the processing mainly lie in the query image signature extraction. On the whole, the search speed is much faster than hybrid image matching. Therefore, if the query image is for the information need of identical image, much more computation often can be avoided than the CBIR based work.

## 4.3 Effectiveness

### 4.3.1 Effectiveness of the identical image search

As we have mentioned before, one hash signature may be mapped into several different images theoretically. However, the 32-bit hash signature provides enough distinguishability for the images in our database as shown in the following statistics. Firstly, we counted the amount of images with duplicate hash signatures. However, the duplicate copies of one image also own the same signatures. Since judging the copies in the whole database needs too much computation, we chose 1 million samples to remove the copies with the same signature by the ordinal 8x8 gray-level blocks comparison. The collisions were recounted and results are illustrated in Figure 8. Therefore, we can conclude that in most of cases one signature only represents one image.



**Figure 8. Distribution of the number of different images with the same signature in  $10^6$  samples (logarithmic scale)**

We chose 40 images from our database including film and product posters, book and CD covers, product casing, etc. We recaptured these images as queries from the computer screen. The margins were automatically cropped by the approach in Section 3.1. These recaptured pictures would distort in scale, color, contrast, luminance, geometry, etc.

We can learn from the above conclusion that for most of query images there will be only one copy of retrieved image. Thus, we can evaluate the effectiveness only by the average *precision*. In the 40 query images, 34 returned results were correct. The average *precision* is 85%.

### 4.3.2 Effectiveness of the hybrid image matching

In this experiment, we tested the effectiveness by searching some kinds of distinguishable-featured fruits, flowers and plants. We chose 8 categories of images including olive, strawberry, litchi, sunflower, hyacinth, baby's breath, poppy and cactus. For each category, the top 5 correct images were selected from the results of Google image search and so there were 40 test images altogether. The attached texts were "fruits", "flower" and "plant" and only basic query expansion were carried out based on Wordnet. Our definition of success of the retrieval is that for a query image, there is at least one occurrence of correct names or synonyms in the first 10 extracted global key phrases. In the returned results, 29 of them were successful and 11 of them failed and the success rate is 72.5%.

## 4.4 Usability

Suppose that the retrieval results were correct, we conducted a user study to evaluate the operation time from the different start points of the pages produced by our system and traditional Web search engines. There were seven participants who were graduated students from nearby universities and were proficient in using the web search engine to acquire the information. One subject was for our pilot study and his results were excluded from the final statistics.

As to the results detected by identical image search module, we chose nine correct returned page including posters, book covers, famous paintings, etc. Each page consists of a single detected image with the link to the original web page and page key phrases with links to online encyclopedia. In these nine images, we selected three examples with the illustrative words inside for the purpose of comparison. For these three special examples, users can start their search by the words from traditional search engines like Google text and image search. For all of the nine images, user started the search with the page generated by our system. In the following search process, users can also use traditional search if needed. The aim was to compare the time consumptions from two kinds of start points under the task of finding certain related information the user wanted.

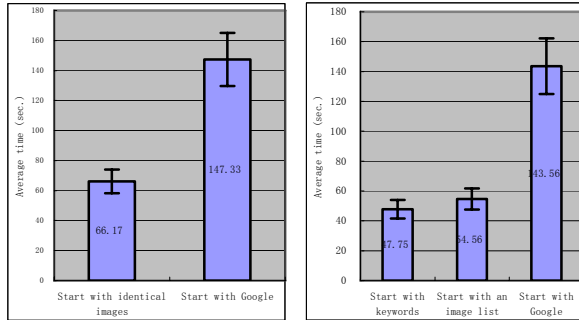
As to the results from hybrid image matching, two kinds of presentations for nine hybrid queries were provided correspondingly. The results were concerned with flowers and fruits. One was nine relevant image lists with links to original pages and each list contains 70 retrieved images. The other was nine groups of global key phrases with links to the entries of online MSN Encarta. Similarly, three keywords among them were provided for the reference. The time consumptions of specific search from the three start points were recorded.

In the pilot study, we asked the user to browse and search on the mobile IE on a Dopod 818 Pocket PC phone. However, we found operation proficiency and rendering time affected the results greatly. In order to eliminate errors caused by operations and rendering, we implemented an explorer in the desktop PC, which was with the same display size as popular 320x240 Pocket PC

phone screens. Moreover, the operations were same as those in PCs and with the two-direction scroll bars.

In our experiment, the kinds of start-point pages were alternately selected. The average time consumptions for one query are illustrated in Figure 9.

We can learn from the experiments that our approach can greatly reduce the search time even if users can describe their information needs in some correct words.



**Figure 9. Searching time consumptions under the different presentations for identical image search (left) and hybrid image matching (right)**

## 5. CONCLUSIONS

In this paper, we combined a number of techniques from different domains, including duplicate image detection, content-based image retrieval, text-based Web image search, and key phrase extraction in an ensemble system to provide a feasible solution to support multimodal queries from mobile devices. We demonstrated the efficiency, effectiveness and usability of our solution using large scale experimental results. In our next steps, we plan to implement a client-side prototype on real camera phones and more considerations such as presentation UI, network connection and rendering speed would be concerned in the mobile platform. Additionally, matching methods based on salient point/region detection are promising in some specific tasks like location recognition from mobile images.

## 6. REFERENCES

- [1] R. Bodner and F. Song, Knowledge-based approaches to query expansion in information retrieval, *Advances in Artificial Intelligence*, pp.146-158, Springer, 1996.
- [2] C. Carson, S. Belongie, H. Greenspan, and J. Malik, Blobworld: image segmentation using expectation-maximization and its application to image querying, *IEEE Transactions on PAMI*, vol.24, no.8, pp.1026-1038, 2002.
- [3] E. Chang, C. Li, J. Z. Wang, et al., Searching near-replicas of images via clustering, *Proc. of SPIE Multimedia Storage and Archiving System VI*, vol.3846, pp.281-292, Boston, USA, Sep. 1999.
- [4] Z. Chen, W. Liu, C. Hu, M. Li, and H.-J. Zhang, Ifind: A Web Image Search Engine, *Proc. of the 24th ACM SIGIR conference on Research and development in information retrieval*, pp. 450, New Orleans, USA, Sep. 2001.
- [5] K. W. Church and P. Hanks, Word association, norms, mutual information and lexicography, *Computational Linguistics*, vol.16, no.1, pp.22-29, 1990.
- [6] M. Flickner, H. Sawhney, W. Niblack, et al., Query by image and video content: the QBIC system, *IEEE Computer Special Issue on Content-Based Retrieval*, vol.28, no.9, pp.23-32, Sep. 1995.
- [7] Google Mobile Search, <http://www.google.com/xhtml>
- [8] Google SMS, <http://www.google.com/sms/>
- [9] J. S. Hare and P. H. Lewis, Content-based image retrieval using a mobile device as a novel interface, *Proc. of SPIE Storage and Retrieval Methods and Applications for Multimedia 2005*, vol.5682, pp.64-75, San Jose, USA, Jan. 2005.
- [10] A. Jaimes, S.-F Chang, and A.C. Loui, Detection of non-identical duplicate consumer photographs, *Proc. of the Fourth Pacific Rim Conference on Multimedia*, vol.1, pp.16-20, Singapore, Dec. 2003.
- [11] Y. Ke, R. Sukthankar, and L. Huston, Efficient near-duplicate and sub-image retrieval, *Proc. of the 12th ACM International Conference on Multimedia*, pp.869-876, New York, USA, Nov. 2004.
- [12] C. Kim, Content-based image copy detection, *Signal Processing: Image Communication*, vol.18, no.3, pp.169-184, Mar. 2003.
- [13] G. Miller, WordNet: A lexical database, *Communication of the ACM*, vol.38, no.11, pp.39-41, 1995.
- [14] M. Noda, H. Sonobe, S. Takagi, and F. Yoshimoto, Cosmos: convenient image retrieval system of flowers for mobile computing situations, *Proc. of the IASTED Conference on Information Systems and Databases 2002*, pp.25-30, Tokyo, Japan, Sep. 2002.
- [15] M. F. Porter, An algorithm for suffix stripping, *Program*, vol.14, no.3, pp.130-137, 1980.
- [16] N. Sebe, Q. Tian, E. Loupias, M. Lew, and T. Huang, Evaluation of salient point techniques, *Image and Vision Computing*, vol.21, pp.1087-1095, 2003.
- [17] J. R. Smith and S.-F. Chang, VisualSEEK: a fully automated content-based image query system, *Proc. of the 4th ACM International Conference on Multimedia*, pp.87-93, Boston, USA, Nov. 1996.
- [18] H. Sonobe, S. Takagi, and F. Yoshimoto, Image retrieval system of fishes using a mobile device, *Proc. of International Workshop on Advanced Image Technology 2004*, pp.33-37, Singapore, Jan. 2004.
- [19] E. M. Voorhees, Query expansion using lexical-semantic relations, *Proc. of the 17th ACM SIGIR conference on Research and development in information retrieval*, pp.61-69, Dublin, Ireland, Jul. 1994.
- [20] W3C Document Object Model, <http://www.w3.org/DOM/>
- [21] Yahoo! Mobile, <http://mobile.yahoo.com>
- [22] T. Yeh, K. Tollmar, and T. Darrell, Searching the Web with mobile images for location recognition, *Proc. of IEEE Conference on Computer Vision and Pattern Recognition*, vol.2, pp.76-81, Washington D.C., USA, Jun. 2004.
- [23] T. Yeh, K. Tollmar, K. Grauman, and T. Darrell, A picture is worth a thousand keywords: image-based object search on a mobile platform, *Proc. of the 2005 Conference on Human Factors in Computing Systems*, pp.2025-2028, Portland, USA, Apr. 2005.