Multimodal Image Search System on Mobile Device

Anushma C R
Department of computer science,
Pondicherry University,
Pondicherry, India

Abstract— Mobile phones have involved into powerful image and video processing devices equipped with built-in cameras, color displays, and hardware-accelerated graphics. These more features allow users to give multimodal queries for searching information on the go from the world wide web. In this paper, we propose a multimodal image search system that fully utilized multimodal and multi-touch functionalities of smart phones. The system allows searching images on the web by using an existing image query or a speech query with the help of existing image search engine. If the user doesn't have an existing image query or captured photo, they can input a speech query that clearly represents a picture description in the user's mind. The proposed system enhances the mobile search experience and increases relevance of search results. It involves a natural interactive process through which user has to express their search intent very well.

Keywords— multimodal search, visual search, mobile phone, interactive search, information retrieval.

I. INTRODUCTION

Now a day's camera-equipped mobile device with increased mobile computing power and network property results in a very rising demand for mobile image search. Most of the existing image system services use text query, but it's difficult for users who have to convert their search intent to text input. Typing a text query, is a difficult job on the phone and also inconvenient because of its limited display screen. So other services introduced voice as a query that provides better interaction in the search process. But the voice-to-search requires a good background, sometimes which gives garbage search results that often happen in a noisy environment. Apart from this camera phone can support differing types of queries like images, contextual information and even videos. Because of these additional features, users more addicted to conduct their search process on mobile phone and input multimodal queries to get more relevant search results. A number of systems are developed, taking advantage of inbuilt cameras on the mobiles. Earlier image search applications such as Goggles [1], iBing, and SnapTell [2] used photo shots(using the built-in camera) as a visual query to search information on the go. It is easier to get a query image using camera phones. In this paper, we facilitate an interactive visual search on mobile phones by taking its full multi-modal and multi-touch functionalities. If the users have an image in their hand, they can use it directly as a query and find matching images on the world wide web like some existing search engines. Otherwise users can easily formulate a composite image as their search query by naturally interacting with the phone through voice and multi-touch. So the proposed system allows users to express their implicit and explicit search intent well. We have designed a multimodal image search system to carry out different types of queries from mobile phones and expressing user's information needs in a better way. Contextual information also added to the system to improve search results. By the use of multimodal queries, the system always gives more satisfactory search results to user than existing systems.

II. PREVIOUS WORK

Recently, many systems developed that used multimodal query for image retrieval. Quickset was the first system that applied multimodal interaction with mobile systems [3], developed by the US Marine Corps. Speak4it local search application is another example [4], [5], where users generate mobile search queries by using multimodal commands that combines speech and drawing. (Xie et al, 2008) proposed client-server architecture for mobile device to perform multimodal search. The architecture mainly consists of four parts mobile Client, a carrier and forward server, a storage server, a media search server [6]. After that, many multimodal systems developed following client-server architecture. Xian Fan et al. proposed a system named Photo-to-Search to search information from the web on the go by using the captured photo [7]. When the user is attracted by any advertisement, they have captured image using camera and start searching process. To get such a query image is easier for camera phone users.

Sometimes when a user performs searching, example images will not be always at hand, which motivates sketch based image retrieval (S B I R) research that uses simpler hand-drawn sketches as a query image. Among various query modalities, the sketch is the most challenging one. Compared with traditional search, sketch-based search is more accurate and convenient when a user's needed information is specific and complex, for example, if you want to find some picture of a beautiful pendant

that you once saw in a shop. That query is usually too ambiguous to properly convey your search intention, but use sketch of pendant is simple. Users have to express their visual intent through sketches, but it's difficult for users without drawing ability. Yang Cao [8] proposed MindFinder system, which is the first interactive sketch-based multimodal image search engine. It enables users to sketch major curves of the target image in their mind; Tagging and clearing operation are also added for higher search results. An image raw curve-based algorithm applied to calculate the similarity between the salient curve representation of natural images and a user's sketch query. The different visual search application applied different types of image matching techniques.

Earlier researchers focus on searching for visually similar objects related to traditional Content Based Image Retrieval (CBIR) Technology [9]. On CBIR systems, image retrieval depends upon the content of images such as color, shape, and text information. However, CBIR-based approaches provide low precision because there is a big gap between high-level semantic concepts and low-level features. Most of the web images are also not tagged with text information. Other two image matching methods based on, key-points are SIFT and SURF. Scale Invariant Feature Transform (SIFT) selects major points on the image, and then compares those points within the given image. This technique is a slow process, takes a long time to check all the points in the image. SURF (Speeded up Robust Feature) is another local feature detector method quicker and stronger than SIFT.

In recent year, I-SEARCH project [10] developed a multimodal search engine provides a novel unified framework for multimedia and multimodal content indexing, sharing, search and retrieval by using the concept of content objects (COs) (Zahariadis et al., 2010) [11]. A "tap-to-search" is other multimodal approach. It helps user to select only interested regions via "tap" actions on the mobile touch screen. Interested region candidates selected by using an Automatic segmentation technique [12]. Visual vocabulary tree based search adopted by incorporating rich contextual information which collected from mobile sensors. Their proposed approach use GPS contextual information, so it gives satisfactory results to users.

III. PROPOSED SYSTEM

This paper proposed a new multimodal search system for image retrieval that helpful in two different situations. Consider an example a user move to an unfamiliar place and takes food from one unknown restaurant. To visit the same restaurant next time. 1) He simply takes a picture of that one. 2) Another situation is that he forgets to take pictures of the restaurant. After returning only he remembered. He also has no idea of the name of the restaurant, but can describe its particular appearance such as "a Chinese restaurant with white door, two red lamps lions, and many red pillars in front.

The proposed system handles these two situations. In first case system uses an image query such as a captured photo of the restaurant and start searching process and retrieve similar images from the web. In the second case, user's doesn't have an existing image, but the user can generate an image query by giving speech input to the system, that represent picture described in the user's mind. Earlier developed sketch based search engine to express user's visual intent in the form of a sketch, but it's difficult for users without drawing experience. But our system helpful for all users they can simply express their information needs via speech input. It uses Google speech recognition engine service to convert user speech to text input. Then keywords extracted from the text. Based on these keywords, users can start searching process, but text- based image retrieval does not give more satisfactory results, which produce a lot of garbage in the search results. So the proposed system generates exemplary images corresponding to each keyword from the back-end search engine (i.e., Google). The images correspond to each keyword, then arranged on canvas and generate a composite image query that used as searching query. The image results depend on the position and resize of exemplary images in the composite query decided by the user. To improve search results, location information also provided to the user. The Architecture of the system shows as in figure 2.

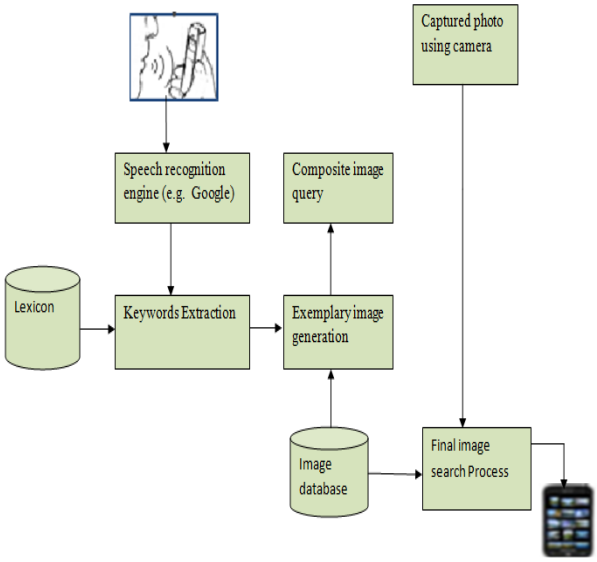

Fig. 1 Architecture of a proposed system

It also solves the It ambiguities in the existing image search engines such as polysemy, aspect, camera view point, and attributes. Polysemy means that a word has multiple meanings. The search for apple results two types of images apple fruit or apple product. Only users can know what information they are looking for, when user search for images related to such ambiguous words, it gives different types of images. The user is not satisfied with these types of image retrieval. The proposed system solves ambiguities because selection of exemplary images depends on user choice, they can express their needed information through exemplary images. Exemplary images corresponding to the keyword is selected from the web or from the gallery itself. Here we designed an application that lets a user search by a taking a picture from a mobile phone's camera or speech query. It would be much quicker and easier. This application allows the user to take a picture of a desired item through a mobile device. The steps involved are 1). The user clicks a picture of the desired item through a mobile camera while on the go or browse images from the gallery itself. An input image sent to an existing search engine and finding matching images. 2) In some cases, the user doesn't have an existing image and have the picture description in their mind, then they prefer to use voice query to initiate search process. Image search using voice involves 4 major components (1) Speech recognition, (2) Keywords extraction, (3) Interactive exemplary visual query composition. Speech recognition is a difficult process than image recognition. It requires 90% accuracy environment. In our application, the Google speech recognition service converts speech to text input after those keywords are extracted from text and forwarded to the Google image search service. It listed exemplary images corresponding to each keyword by applying the clustering method. Here we apply k-means clustering algorithm to cluster images from the web. It groups list of images corresponding to each keyword. From the list of images, the system chooses top most images corresponding to each keyword and display alternately. The user can arrange their selected exemplary images on the canvas and generate a composite image query to start the final search process. Finally, Google retrieves relevant images corresponding to the user composite visual query. The size and position of exemplary images on the canvas effects the search results.

IV. EXPERIMENTS

In this section, we will first introduce the implementations of our system on an Android device in Section IV.A. After that compares our system with existing image search systems in section IV. B.

A. Implementation on the Mobile Phone

We create a system with a phone application to check our multimodal image search engine. The figure shows the application's user interface. Lake, tree, and sky are the three keywords extracted and displayed in the text box. The user has to give speech input through speaker button to describe their picture description in mind. If the user does have an existing image query, they have to use images from gallery to start the search process. If the user is attracted by an object on the go, he has to take pictures of an object at a time and search for similar images from the web. Otherwise users can give speech input by clicking the speaker button and to express their speech description in minds. Depend on the situations, user can start searching process. If the speaker button is tapped application transforms speech to text via Google speech recognition and display text input in a text box. After that keywords extracted and images corresponding to the keywords from web display alternatives. User can also choose exemplary images from gallery corresponding to entities. Changing the position and size of exemplary images, cropping of images affected search results.

B. Comparison of proposed system to Existing system

In the experiments we compare the performance of text search, Mindfinder [8], Photo to Search [7], and the proposed image search system. For example, a user wants to "find an image with several green trees in front of a white house". Text based image search does not handle this long query well. Because of this it may not result in any relevant search, the user is also uncomfortable with typing such a long query on the small screen. Photo to search retrieves some similar images, but the user has to give exactly similar or partially duplicate image search. Using MindFinder [8] user has to express their needed information through the sketch. It gives more relevant results than photo to search. In MindFinder, the use of grid based matching give better result than text, if the database is large enough. The proposed multimodal image search does a better job than all above method, by expressing their visual intent clearly. The Figure shows average precisions of the 15 search results by comparing four methods.

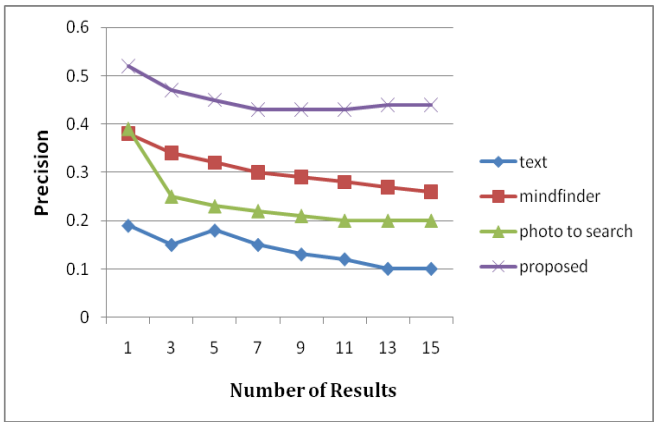

Fig. 2 The precision curve of four different search methods for the "similar image search" task.

V. CONCLUSIONS

Here we propose an interactive multimodal image search system on mobile phone that helps users to express their needed information implicitly and explicitly. Only the user has know what information they are looking for, so expressing their information need play very important role in the search process. Proposed system gives a better way to express their visual intent than other existing systems, it provides more relevant search results, especially in case where users can have a partial picture description in mind. The System provides a cool game-like user interface for query formulation and enhanced user experience on a mobile phone.

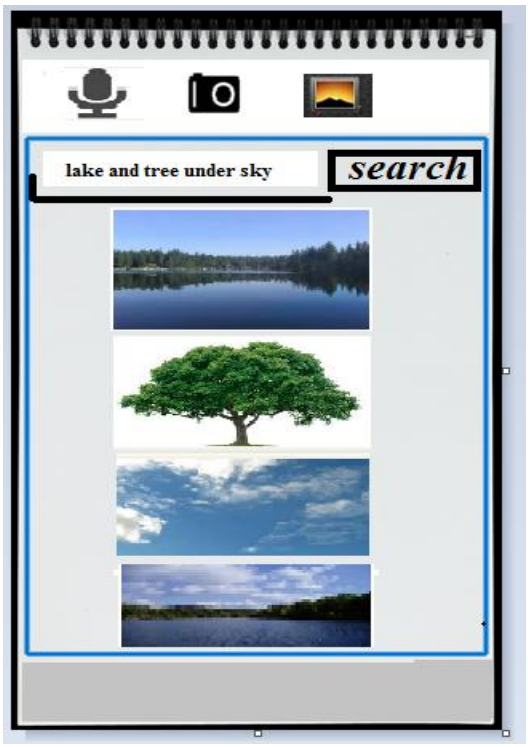

Fig 3 User interface of proposed system

REFERENCES

[1] Google Goggles. http://www.google.com/mobile/goggles/.

[2] SnapTell. http://www.snaptell.com/.

[3] Philip R. Cohen, Michael Johnston, David McGee, Sharon Oviatt, Jay Pittman, Ira Smith, Liang Chen and Josh Clow. QuickSet: multimodal interaction for distributed applications. In *MULTIMEDIA* '97: *Proceedings of the fifth* ***ACM*** *international conference on Multimedia*, pages 31–40, New York, NY, USA, 1997. ACM.

[4] Ehlen, P., Johnston, M., Ave, P., & Park, F. (2010). Speak4it : Multimodal Interaction for Local Search Categories and Subject Descriptors

[5] Michael Johnston and Patrick Ehlen, SPEAK4IT : Multimodal Interaction in the Wild, AT & T Labs Research, AT & T Labs. (2010), 147–148

[6] X. Xie, L. Lu, M. Jia, H. Li, F. Seide, and W.-Y. Ma. Mobile Search with Multimodal Queries. Proceedings of the IEEE, 96(4):589{601, April 2008

[7] Fan, X., Xie, X., Li, Z., Li, M., & Ma, W. (2005). Photo-to-Search : Using Multimodal Queries to Search the Web from Mobile Devices, 143–150

[8] Wang, H., & Wang, C. (n.d.). MindFinder : Interactive Sketch-based Image Search, 1605–1608.

[9] M. Flickner, H. Sawhney, W. Niblack, J. Ashley, Q. Huang, B. Dom, M. Gorkani, J. Hafner, D. Lee, D. Petkovic, and P. Yanker, "Query by image and video content: The QBIC system," IEEE Computer, vol. 28, no. 9, pp. 23-32, Sep. 1995

[10] Axenopoulos, A., Daras, P., Malassiotis, S., & Croce, V. (n.d.). I-SEARCH : A Unified Framework for Multimodal Search and Retrieval, 130–141

[11] Zahariadis T., Daras P., Bouwen J., Niebert N., Griffin D., Alvarez F., Camarillo G., "Towards a Content-Centric Internet", Towards the Future Internet - Emerging Trends from European Research, IOS Press, ISBN 978-1-60750-539-6, pp. 227-236, Apr 2010

[12] Zhang, N., Mei, T., Hua, X.-S., Guan, L., & Li, S. (2011). Tap-to-search: Interactive and contextual visual search on mobile devices.*2011 IEEE 13th International Workshop on Multimedia Signal Processing*,1–5. doi:10.1109/MMSP.2011.6093802.

S. M. Metev and V. P. Veiko, *Laser Assisted Microtechnology*, 2nd ed., R. M. Osgood, Jr., Ed. Berlin, Germany: Springer-Verlag, 1998.

.