ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

Interactive Multimodal Visual Search Using Voice Command on Mobile Device

Mayur Kanawade [1], Unmesh Kadam[2], Priyanka Jadhav[3], Jai Chavan[4]

Student, Department of Computer Science, SIT, Lonavala, India[1,2,3]

Professor, Department of Computer Science, SIT, Lonavala, India[4]

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 content very well.

Keywords: multimodal search, visual search, mobile phone, interactive search, information retrieval

I. INTRODUCTION

Image search is a hot topic in both computer vision and information retrieval with many applications. More consumers use phones or other mobile devices as their personal concierges surfing on the Internet. Along this trend, searching is becoming pervasive and one of the most popular applications on mobile devices. The bursting of mobile users puts forward the new requests for image retrieval. The images are searched based on the query given by the user. Text, Voice, Sketches, Photo and Content of the images are used as query to search images on mobile devices. In the text-based search, the user can type an entity name to find the images. While on the go, consumers use their phones as a personal Internet-surfing concierge. Searching is becoming pervasive and is one of the most popular applications on mobile phones. People are more and more addicted to conducting searches on their phones. It is reported that one-third of search queries will come from smart phones by 2014. However, compared with text and location search by phone, visual (image and video) search is still not that popular, mainly because the user's search experience on the phone is not always enjoyable. On one hand, existing forms of queries (i.e., text or voice as queries) are not always user friendly—typing is a tedious job, and voice cannot express visual intent well. On the other hand, the user's intent in a visual search process is somewhat complex and may not be easily expressed by a piece of text (or text transferred from voice). For example, the query like "find a picture of a person with a straw hat and a spade" will most likely not result in any relevant search results from existing mobile search engines.

II. PROPOSED SYSTEM

It is designed for the users who already have pictures in their minds but have no precise description or names to address them. By describing it using speech and then refining the recognized query by interactively composing a visual query using exemplary images. Directly applying keyword-based search to mobile visual search is straightforward yet intrusive. As we have mentioned, typing a long query is not always user-friendly on mobile devices. This is the reason that mobile users type on average 2.6 terms per search, which can hardly express their search intent. Compared with text-to-search, capture to-search is becoming dominant in mobile visual search. It is more convenient for mobile users to take a photo and use it to search , Point and Find are recent visual search applications in this area .There exist efforts on mobile visual search in the computer vision community. Most of these efforts have focused on the exploration of different visual descriptors. The search procedure of our proposed system consists of the following phases: 1) the user speaks a natural sentence to describe the intended images, 2) the speech is recognized and further decomposed into keyword(s) which can be represented by exemplary images, 3) the user selects preferred exemplar(s) and composes a schematic collage as a composite image, 4) the composite image is then used as a visual query to search for similar images, and 5) if possible, further information like GPS locations and image descriptions are provided to the user. The first step is speech recognition. Speech recognition is now a much more mature technique than image recognition. Especially for speech to text technique, the state-of-art accuracy achieves an accuracy of 98% in quiet environment. Actually it is still infeasible for machine to understand natural human language. However, there are feasible tools of data mining to partially solve the restricted problem. Fortunately, there are a lot of works and resources available to do such mining, such as ImageNet and various online search engines. Since the task is just to extract words in the sentence which can be represented by images, ImageNet can be a code book to find such useful words. For example, the user says "find an iron tower on

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

the grass," and then the system will recognize "tower" and "grass." With these two concepts, the system can further refer to the Internet to understand what these concepts might look like. The Internet images with each keyword will be fetched and analyzed based on their text information and image content. Through mining and clustering, a few proper exemplary images are selected and shown to the user. Many approaches are also available in this step like visual query suggestion .More search intent will be revealed once the user selects exemplary image(s). Moreover, through resizing and positioning the exemplar(s) on a blank canvas on the screen using multitouch input, the search intent is vivid and obvious. Given such a composite image, similar image search is then performed.

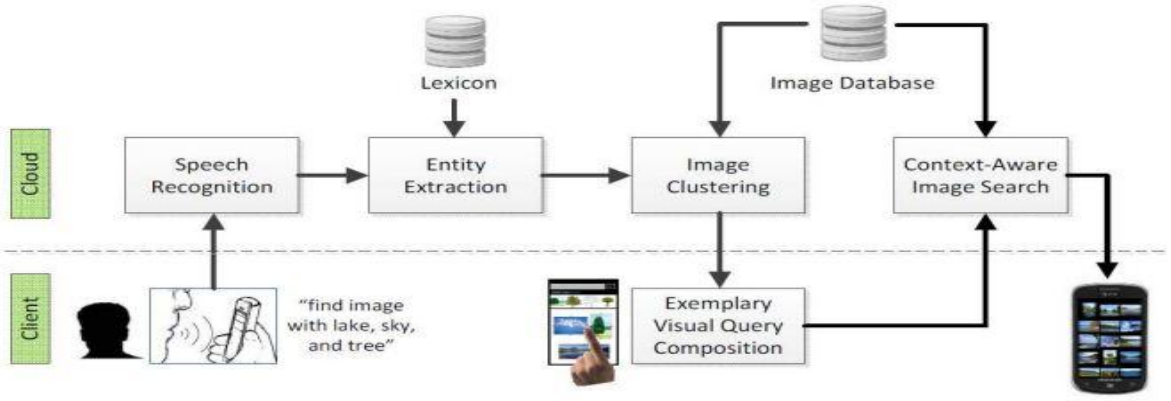

Fig. 1 Architecture of a proposed system

III. DESIGN AND IMPLEMENTATION CONSTRAINTS

The most related work on generic visual search to JIGSAW is interactive search, in which users specify their search intent interactively. The advanced functionalities in Google and Bing's image search engines enable user to indicate search intent via various filters, e.g., "similar images," color, style, face, and so on. Tineye supports the uploading of an exemplary image as a query example for search, while Xcavator even enables users to emphasize certain regions on the query image as the key search components. In a more advanced search engine prototype, such as GazoPa and MindFinder, the search is performed by sketching a shape image. The "Concept Map" uses the position and size of a group of tags to filter the top text-based search results, while the "Color Map" enables the selection of multiple color hints on a composite canvas as a visual query. However, user interaction on desktop is not as natural as that on mobile device. Therefore, an interactive mobile visual search system which takes advantage of multi-touch and multi-modal functionalities is desirable. [1]

A. .Jigsaw

JIGSAW is an interactive mobile visual search application that enables users to naturally formulate their search intent in an interactive way and combines different visual descriptors (e.g., SIFT, color, and edge) for visual search. Figure 2 shows the framework of JIGSAW. On the client-side, a user first speaks a natural sentence to initiate a voice query, e.g., a sentence like"find an iron tower on the grass." On the cloud side, the system employs speech recognition (SR) to transfer the speech to a piece of text, and then extracts entities from the text. As a result, "tower" and "grass" are recognized as two entities that can be represented by two exemplary images. Directly using those entities as textual queries may not return relevant results, as it only searches the surrounding text and neglects the position and size of these exemplary images on the query canvas. Therefore, we propose to enable users to further specify search intent by touching the screen and dragging their preferred exemplary images, and then formulating a composite visual query. Those exemplary images are automatically generated using a clustering process according to the extracted entities. Finally, we exploit both the text and the composite visual query to search for relevant images, by considering the position and the size of the exemplary images. In the next sections, we will describe the details of each component.

1) The user speaks a natural sentence to describe the images,

2) The speech is recognized and then decomposed into keyword(s) which can be represented by exemplary images,

3) The user selects preferred exemplar(s) and then composes image,

4) The composite image is then used as a visual query to search for similar images.

Compared with JIGSAW, JIGSAW+ the algorithm has been improved in three aspects:

1)Segmentation-based image representation;

2) relative position checking and

ISSN (Online) 2278-1021
ISSN (Print) 2319-5940

International Journal of Advanced Research in Computer and Communication Engineering
Vol. 4, Issue 3, March 2015

3) inverted index is constructed for matching.

Segment Based Image Representation:

The images are retrieval based on features of images such as color, texture and more. Uniform grid partitions are used to break the original image into smaller pieces of image.[3]

B. Color Feature Extraction:

Before color feature extraction, the images are over-segmented. The over-segmentation methods use graph based algorithms that segment an image into many homogeneous regions. Each node in the graph stands for a pixel in the image, with undirected edges connecting its adjacent pixels in the image. The weight of each edge between two pixels reflects their similarity. The similar pixels are merged. Moreover, the similarity of RGB-color space is used instead of gray level, so that inside each piece the color is close.

C. Texture Feature Extraction:

The most widely used local feature of SIFT (http://pointandfind.nokia.com) to capture local texture patterns in the image. To reduce computing and noise, only prominent regions are used instead of homogeneous regions, prominent regions are detected using Difference of Gaussian (DoG) detector.

D. *Relative Position Checking:*

The position of each exemplar in image should be consistent with the composite visual query. The existence scores for all exemplars are obtained for Image.

IV. CONCLUSION

Text-based search engines are still available on mobile devices. But it is neither user-friendly on phone, nor machine-friendly for search engine. Voice queries must need general idea of expected pictures such as color configurations and compositions. Sketch-based search is difficult to use for users without drawing experience. Photo-to-search needs exact partial duplicate images in their database for search similar images. Thus user's search experience on mobile device is significantly improved by interactive mobile visual search system compare to all other techniques, which allow the users to formulate their search through multimodal interactions with mobile devices. The visual query generated by the user can be effectively used to retrieve similar images. Mobile visual search (Ystad and Sweden 2011) takes the advantages of multimodal and multi-touch functionalities on the phone.Our future works include the following issues. First, we will try to use the visual structure within each exemplar, which may further improve the similar image search results. Second, we will further develop the usability of our system and improve the user experience. For example, we may deploy the visual search system on other mobile devices with larger screen such as tablets. Thus more powerful interactions and functions can be introduced into the system. Third, we will focus on is combining low-level features into mid-level features. Because a relatively small vocabulary size degrades the searching speed and large vocabulary size is too sensitive to feature variances, multiple low-level features can be combined into more robust and discriminative visual words.

ACKNOWLEGEMENT

We express our sincere thanks to our guide who have provided us valuable guidance towards the completion of this paper. We hereby take this opportunity to record our sincere thanks and heartily gratitude to **Prof. Jai Chavan** without whom we would have not started authoring this paper.

REFERENCES

[1] Houqiang Li, *Senior Member, IEEE*, Yang Wang" Interactive Multimodal Visual Search on Mobile Device" IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 3, APRIL 2013

[2] Anushma C R "Multimodal Image Search System on Mobile Device" International Journal of Computer & Organization Trends – Volume 7 Number 1Apr2014

[3] JIGSAW: Interactive Mobile Visual Search with Multimodal ueries_Yang Wang †, Tao Mei ‡, Jingdong Wang ‡, Houqiang Li †, Shipeng Li † University of Science and Technology of China, Hefei 230027, P. R. China‡ Microsoft Research Asia, Beijing 100080, P. R. China

[4] H.Bay,T.Tuytelaars,andL.VanGool,"SURF:Speeded-up obust features," in Proc. ECCV, 2008, vol. 110, no. 3, pp. 346–359.

[5] 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.

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

[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