International Journal of Advanced Research in Science Management and Technology
ISSN(Online)-2454-4159 Volume 3, Issue 9, September 2017

Survey on Interactive Visual Search

Manisha M. Kundle[1], Shital Jadhav[2]

PG Scholar, Dept. of Computer Science and Engg., G.H. Raisoni Institute of Engg. & Management, Jalgaon, M.S.,India[1]

Professor, Dept. of Computer Science and Engg., G.H. Raisoni Institute of Engg. & Management, Jalgaon, M.S.,India[2]

ABSTRACT— Mobile phones have evolved into powerful image processing devices equipped with high-resolution cameras, color displays, and hardware-accelerated graphics. This survey paper describes a novel multimodal interactive image search system on mobile devices. The system, the Joint search with ImaGe, Speech, And Word Plus (JIGSAW), takes full advantage of the multimodal input and natural user inter-actions of mobile devices. It is designed for users who already have pictures in their minds but have no precise descriptions 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, the user can easily find the desired images through a few natural multimodal interactions with his/her mobile device.

KEYWORDS- Mobile Visual Search, Multimodal Search, Interactive Search, Mobile device.

I. INTRODUCTION

Image search is a hot topic in both computer vision and information retrieval with many applications. The traditional desktop image search systems with text queries have dominated the user behavior for a quite long period. However, while on the go, more and 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 [10]. It is reported that one-third of the Internet search queries will come from phones by 2014 [11]. The bursting of mobile users puts forward the new requests for image retrieval.

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, as shown in Fig. 1(a), 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.

Fig.1Three modes of mobile visual search: (a) voice/text-to-search, (b) photo-to-search, (c) interactive multimodal search

To facilitate visual search on mobile devices, the work described in this paper aims at a more natural way to formulate a visual query, taking full advantage of multi-modal and multi-touch interactions on mobile devices. As shown in Fig. 2(c), users can easily formulate a composite image as their search intent by naturally interacting with the phone through voice and multi-touch. Although similar applications exist, such as Goggles[1] , iBing, and SnapTell[7] , which support photo shots (using the built-in camera) as a visual query for instant search, as shown in Fig. 2(b), our work represents a complementary mobile visual search by which users can compose an arbitrary visual query (not necessarily an existing image) through natural user interaction.

It is known that visual search on a mobile device is different from that on a desktop. Compared with a desktop PC which predominantly supports text-to-search mode, a mobile phone provides a richer set of user interactions and thus achieves a more natural search experience. For example, beyond the traditional keyboard and mouse inputs, mobile phones are usually enabled to receive multi-modal inputs. The most common interface of this kind combines a Figure 1: The main modes for mobile visual search: (a) voice/text-to-search, (b) capture-to-search, (c) Modular search. Visual modality via the built-in camera with a voice modality via speech recognition. In addition, the multi-touch phone screen, which recognizes multiple simultaneous touch points, provides rich interaction between users and devices. All these advantages provide for a more natural interaction to formulate search intent and thus achieve a better search experience via mobile phone.

There exist some visual search applications for mobile devices. Table 1 is a survey of the recent visual search applications on various mobile platforms. All of them require users to first take a picture and then perform similar image searches in various vertical domains (i.e., capture-to-search mode). However, in many cases, the user's search intent is implicit and cannot be represented through capturing the surroundings. The user nevertheless, can express his/her intent via a piece of voice description. For example, a user is looking for a restaurant with a red door and two stone lions in front of the door, however s/he forgot the name of the restaurant. Therefore, a client-side tool that can transfer a long textual query into a visual query with user interaction is required to determine the restaurant's name and location.

Table.1 Recent Mobile Search Application

App	Features
Goggles [1]	product, cover, landmark, namecard
Digimarc Discover [27]	print, article, ads
Point and Find [2]	place, 2D barcode
SnapTell [7]	cover, barcode
SnapNow [28]	MMS, email, print, broadcast
Kooaba [3]	media cover, print

Photo-to-search is becoming pervasive as the development of the computer vision and content-based image retrieval. This enables the user to capture photos using the in-built camera on the phone and then initiate search queries about objects in visual proximity to the user . This advance offers various applications such as identifying products, comparison shopping, finding information about buildings, movies, compact CDs, real estate, print media, artworks, etc. First deployments of such systems include Google Goggles [1], Nokia Point and Find [2], Kooaba [3], Ricoh iCandy [4]–[6], and Amazon Snaptell [7].

Goggles [1] product, cover, landmark, namecard, Digimarc Discover[27] print, article, ads, Point and Find [2] place, 2D barcode SnapTell [7] cover, barcode SnapNow [28] MMS, email, print, broadcast Kooaba [3] media cover, print touch interactions, to help users formulate their (implicit) search intent more conveniently and thus promote visual search performance. The search procedure consists of the following phases: 1) the user speaks a natural sentence as the query to the phone,2) the speech is recognized and transferred to text, and the text is further decomposed into keywords by entity extraction, 3) the user selects the preferred exemplary images (given by an image clustering process according to the entities) that can visually represent each keyword and composes a query image through multi-touch, 4)the composed image is then used as a visual query to search similar images. Therefore, the key component of JIGSAW is the composition of an

exemplary image query generated from the raw speech via multi-touch user interaction, as well as visual search based on the exemplary image. Through JIGSAW, users can formulate their visual search intent in a natural way like piecing together a jigsaw puzzle on the phone screen. The techniques in JIGSAW include speech recognition, entity extraction, image clustering, large-scale image search, and user interaction.

Regarding content-based image search, one kind of famous products, including Google Image [12], TinEye [13] on PC, and Google Goggles [1] on mobile phone, can accept single images as search queries, and return to the user similar images or even with information mined from their databases. With very large databases, these engines are able to achieve impressive results.

However, to initiate such a visual search, the user must have an existed image on hand as a query. Moreover, it needs partially duplicate images or exact the same thing existing in the database. Another kind of image search engines designed for desktop, including GazoPa [14] and some other sketch-based image search researches like [15], [16], [17], use hand-drawn sketches to search for satisfied images. Though sketch-based search allows users to express their visual intent in some way, it can hardly develop complex meanings and is difficult to use for users without drawing experience. MindFinder [15] and Concept Map [18] also provide visual aids to search for images. In these works, visually and semantically similar images are retrieved by multiple exemplary image patches. The user offers lexicons and then composes a visual query using multiple image patches given by the engine according to the lexicons. In these works, images are unnaturally divided into blocks in which features are then extracted. The performance is very sensitive to selections and positions of exemplars. For further information please refer to the papers. Interestingly, in [19] the authors build a Sketch2Photo system that uses simple text-annotated line sketch to automatically synthesize realistic images. They also employ text and sketch to search for templates which are then stitched on a background to generate a montage. However, their work focuses on image composing instead of image retrieval. A small screen limits the presentation of searching results, which requires the top results to be more relevant while on the phone. However, using only text as search query can hardly meet this end. The surrounding texts of the web images are not always correct. Even the tags of the some human-labeled datasets such as Flickr images are unreliable. Moreover, on the one hand, the user must know the exact terms the annotator used in order to be able to retrieve the images he wants. On the other hand, textual annotations are also language-dependent. Actually, there are more images which have no text information on the web repository. All this deficiency can ruin a good user experience of text-based image search system on the mobile phone. Compared with text search, map search, and photo-to-search, visual (image and video) search is still not very popular on the phone, though image search has become a common tool on the PC since 10 years ago, with which the user can input text query to retrieve relevant images. A main reason why such image search applications are not popular on mobile device is that the existing image search applications do not perfectly accommodate to the mobile and local oriented user intent. First of all, typing is a tedious job on the phone no matter whether a tiny keyboard or a touch screen is used.

Mobile image-retrieval applications pose a unique set of challenges. What part of the processing should be performed on the mobile client, and what part is better carried out at the server?[8]

This [9] describes a novel multimodal interactive image search system on mobile devices. The system, the Joint search with ImaGe, Speech, And Word Plus (JIGSAW+), takes full advantage of the multimodal input and natural user interactions of mobile devices. It is designed for users who already have pictures in their minds but have no precise descriptions 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, the user can easily find the desired images through a few natural multimodal interactions with his/her mobile device. Compared with JIGSAW, the algorithm has been significantly improved in three aspects: 1) segmentation-based image representation is adopted to remove the artificial block partitions; 2) relative position checking replaces the fixed position penalty; and 3) inverted index is constructed instead of brute force matching. The JIGSAW+ is able to achieve 5% gain in terms of search performance and is ten times faster.

In [20]system, user provides input query to the system that can be either text or voice or image and then according to that query, composite images are provided to the users. After that User selects one image among them and retrieves more relevant images. This system is useful in a case where users do not know exact name of an image but by describing it using either text or speech or by providing any other relevant image, users can easily find targeted image. The ANN technique is also added into it to increase the performance of the system. The system works in three phases- 1) Image Composition, 2) Image Processing, 3) ANN In [21] presented a mobile application that allows to perform visual search queries in the fashion domain using a client server architecture. Starting from a picture, the application automatically classifies the object of interest in order to facilitate the user to compose a query, minimizing his interaction with the device. The combination of a SVM and a PHOG feature for the classification task, shows high discriminative capacity and its low computational cost makes it a perfect candidate for the use on devices with low computational power and in a real time context. In particular, process the video stream from the built-in camera of the mobile device, in order to suggest an automatic prediction of the product name of the user's object of interest. The query result consists of a set of commercial products, related to the object of interest, returned using the API of an online visual indexing engine.

The application in order to recognize other types of clothing and at the same time increasing the performance of classification. In [22] a multimodal image searches 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 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.

II. LITERATURE REVIEW

Author proposed an image retrieval framework that integrates efficient region-based representation in terms of storage and complexity and effective on-line learning capability is proposed [23]. The framework consists of methods for region-based image representation and comparison, indexing using modified inverted files, relevance feedback, and learning region weighting. By exploiting a vector quantization method, both compact and sparse (vector) region-based image representations are achieved. How to separate images containing objects from images containing scenes or textures is a crucial and interesting issue for the framework.

Author proposed a system for a query image (mobile image) taken by the mobile phone, our goal is to retrieve its most similar image in a large scale image database where usually the most similar image is the original image taken photos of and are associated with their relevant information [24]. The method exploits spatial relationships between features and combines them with visual matching information to improve features discriminative power. The performance has to be improved further by exploiting better schemes that can utilize both visual and spatial information consistently

Author proposed a system for comparing the performance of MPEG-7 image signature tools, SIFT and CHoG in the context of mobile visual search applications [25]. With the review of MPEG-7 image signatures and CHoG descriptors and discuss feature level Receiver Operating Characteristic (ROC) experiments and pair wise image-matching experiments for the different descriptors. The image matching accuracy for schemes is low in between 70% and 76%.

Author proposed a system using mobile camera, GPS information and PC server to search and recognize buildings without typing any words [26].With quick development of mobile techniques, a large number of people already own smart phones. The effectiveness and efficiency of the system under more diverse cases shall be tested and analyzed.

III. PROBLEM STATEMENT

Mobile visual search became an active filed area in pass few year. It involves methods from several research areas. Due to the complexity of visual information compared with text or voice signals. It is reported that one-third of the Internet

search queries will come from phones by 2014. The bursting of mobile users puts forward the new requests for image retrieval. First, there is a huge gap in user interface between desktop and mobile devices especially for the input methods. On desktop, keyboards and mice are the main input devices while recent mobile devices always provide multimodal input methods including cameras, GPS, microphones, and multitouch screens. Finding information from large image/video sources poses a challenging yet exciting problem. Two well-known barriers preventing successful solutions to date are the semantic gap and the user gap. The former refers to the difficulty in inferring high-level semantic labels from low-level pixel data. The latter reflects user's frustration in expressing his/her information needs of visual content using existing systems.

Integrating multi-modal information to solve many challenging programs, such as

1) The image should contain the entities in the user's query (image description).

2) The entities in the image should be similar to the exemplary images the user chooses (the Comparison approach).

3) The position of single entity or the relative layout of multiple entities in the image should consist with the composite

4) Information description methods, matching and retrieval.

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 proposed system will provides a game-like interactive image search scheme with composition of multiple exemplars by taking the advantages of multimodal and multi-touch functionalities on the phone.

REFERENCES

[1] (2009).Google Goggles [Online]. Available: http://www.google.com/mobile/goggles/

[2] (2006). Nokia Point and Find [Online]. Available: https://en.wikipedia.org/wiki/Nokia_Point_%26_Find

[3] Kooaba. (2007). Kooaba [Online]. Available: http://www.kooaba.com

[4] B. Erol, E. Antúnez, and J. Hull, "Hotpaper: Multimedia interaction with paper using mobile phones," in Proc. 16th ACM Multimedia Conf., New York, 2008.

[5] J. Graham and J. J. Hull, "Icandy: A tangible user interface for itunes," in Proc. CHI '08: Extended Abstracts on Human Factors in Computing Systems, Florence, Italy, 2008.

[6] J. J. Hull, B. Erol, J. Graham, Q. Ke, H. Kishi, J. Moraleda, and D. G. Van Olst, "Paper-based augmented reality," in Proc. 17th Int. Conf. Artificial Reality and Telexistence (ICAT), Washington, DC, 2007.

[7] Amazon. (2007). SnapTell [Online]. Available: http://www.snaptell.com.

[8] "Mobile visual search", IEEE SIGNAL PROCESSING MAGAZINE [72] JULY 2011.

[9] Houqiang Li, Yang Wang, Tao Mei, Jingdong Wang, and Shipeng Li, "Interactive Multimodal Visual Search on Mobile Device", IEEE TRANSACTIONS ON MULTIMEDIA, VOL. 15, NO. 3, APRIL 2013.

[10] K. Church, B. Smyth, P. Cotter, and K. Bradley, "Mobile information access: A study of emerging search behavior on the mobile internet," ACM Trans. Web, vol. 1, no. 1, May 2007.

[11] Online].Available:http://www.pwc.com/gx/en/communications/review/features/mobile-data.jhtml

[12] Google.[Online]Available:http://www.google.com/

[13] TinEye.[Online].Available:http://www.tineye.com/

[14] GazoPa. [Online]. Available: http://www.gazopa.com/

[15] Y. Cao, H. Wang, C. Wang, Z. Li, L. Zhang, and L. Zhang, "MindFinder: Interactive sketch-based image search on millions of images," in Proc. ACM Multimedia, 2010, pp. 1605–1608.

[16] M. Eitz, K. Hildebrand, T. Boubekeur, and M. Alexa, "Sketch-based image retrieval: Benchmark and bag-of-features descriptors," IEEE Trans. Vis. Comput. Graphics, vol. 17, no. 11, pp. 1624–1636, Nov. 2011.

[17] M. Eitz,K.Hildebrand, T. Boubekeur, and M. Alexa, "An evaluation of descriptors for large-scale image retrieval from sketched feature lines," Comput. Graphics, vol.34, no.5, pp. 482–498, 2010.

[18] H. Xu, J. Wang, X. Hua, and S. Li, "Image search by concept map," in Proc. ACM SIGIR, 2010, pp. 275–282.

[19] T. Chen, M.-M. Cheng, P. Tan, A. Shamir, and S.-M. Hu, "Sketch2photo: Internet image montage," ACM Trans. Graph., vol. 28, no. 5, pp. 124:1–124:10, Dec. 2009.

[20] Yojana S. Pawale,Prof. Prakash Devale, "I2MMS: An Interactive Multi Modal Visual Search Technique", International Journal of Advanced Research in Computer and Communication Engineering Vol. 4, Issue 5, May2015

[21] A. Nodari ,M. Ghiringhelli, A. Zamberletti, M. Vanetti, S. Albertini, I. Gallo , "A mobile visual search application for content based image retrieval in the fashion domain" , IEEE JUNE 2012

[22] Anushma C R, "Multimodal Image Search System on Mobile Device", International Journal of Computer Organization Trends-Volume 7 Number 1- Apr 2014

[23] F. Jing, M. Li, H.-J. Zhang, and B. Zhang, "An efficient and effective region-based image retrieval framework," IEEE Trans. Image Process., vol. 13, no. 5, pp. 699–709, 2004.

[24] X. Liu, Y. Lou, A.W. Yu, and B. Lang, "Search by mobile image based on visual and spatial consistency," in Proc. IEEE Int. Conf.Multimedia

[25] V. Chandrasekhar, D. M. Chen, A. Lin, G. Takacs, S. S. Tsai, N. M.Cheung, Y. Reznik, R.Grzeszczuk, and B.Girod, "Comparison of local feature descriptors for mobile visual search," in Proc. IEEE Int. Conf.Image Process., 2010, pp. 3885–3888

[26] Jiemin wang1,2,3, Yuanhai He3, Yujie Zhou3, Yu Qiao1," iGAPSearch: Using Phone Cameras to Search Around the World".

[27] Digimarc Discover [Online]. Available: https://www.digimarc.com/discover/

[28] LinkMe Mobile. [Online]. Available: http://www.snapnow.co.uk/.