Interaction Design for Mobile Visual Search

Jitao Sang, Tao Mei, *Senior Member, IEEE*, Ying-Qing Xu, *Senior Member, IEEE*, Chen Zhao, Changsheng Xu, *Senior Member, IEEE*, and Shipeng Li, *Fellow, IEEE*

***Abstract*—Mobile devices are becoming ubiquitous. People take pictures via their phone cameras to explore the world on the go. In many cases, they are concerned with the picture-related information. Understanding user intent conveyed by those pictures therefore becomes important. Existing mobile applications employ visual search to connect the captured picture with the physical world. However, they only achieve limited success due to the ambiguity nature of user intent in the picture—one picture usually contains multiple objects. By taking advantage of multitouch interactions on mobile devices, this paper presents a prototype of *interactive mobile visual search*, named TapTell, to help users formulate their visual intent more conveniently. This kind of search leverages limited yet natural user interactions on the phone to achieve more effective visual search while maintaining a satisfying user experience. We make three contributions in this work. First, we conduct a focus study on the usage patterns and concerned factors for mobile visual search, which in turn leads to the interactive design of expressing visual intent by gesture. Second, we introduce four modes of gesture-based interactions (*crop, line, lasso*, and *tap*) and develop a mobile prototype. Third, we perform an in-depth usability evaluation on these different modes, which demonstrates the advantage of interactions and shows that *lasso* is the most natural and effective interaction mode. We show that TapTell provides a natural user experience to use phone camera and gesture to explore the world. Based on the observation and conclusion, we also suggest some design principles for interactive mobile visual search in the future.**

***Index Terms*—Interaction design, interactive search, mobile visual search, user interfaces.**

I. Introduction

WITH the ubiquity of mobile devices, people are using their phones as a personal concierge discovering and making decisions anywhere and anytime. For example, they can easily take pictures via their phone cameras and get related information about the pictures on the go. Understanding visual intent conveyed by the captured pictures therefore become important for mobile users. An exemplary mobile application of this kind is LeafSnap[1]: you take a snap of the leaf, search it on Internet with your mobile devices, and the pictures of plants with similar appearances as well as their names and detailed information are returned. LeafSnap belongs to a new kind of mobile applications called *mobile visual search*, which is more natural and user-friendly than text or voice-based mobile search—there is no bothering to know what it is or how to pronounce it; as long as you see it, you can initiate the query. Besides leaf identification, there exists a broad range of applications for mobile visual search, e.g., identifying landmarks, comparing shoppings, finding information about movies, books, artworks, and so on.

Attractive as these applications are, existing systems for mobile visual search work only on limited object categories and under controlled conditions [10], which restricts its potential from a nice-to-have to a must-have. The reason is twofold. First, from the perspective of technology, though many researchers devoted on it, state-of-the-art algorithms find it challenging to perform intelligent image understanding on professional pictures, not to mention on those captured by phone cameras with huge illumination variation and complex surroundings. Second, just like an old saying, "one picture is worth thousands of words," the picture usually contains multiple objects. The ambiguity nature of user intent in the captured picture makes it difficult to understand user intent without user-specified region-of-interest (ROI). We are investigating in this paper the way for better understanding user intent from the phone-captured picture and improving mobile visual search performance, while keeping users in the loop.

The basic premise is that, in mobile visual search, if users are willing to spend limited efforts interacting with the captured pictures to express their intent more clearly, e.g., selecting the ROI in one step, the application would receive more focused purpose and provide a better search performance. Different from many existing mobile visual search applications which do not take advantage of multitouch functionality on mobile devices, this paper presents a prototype of *interactive mobile visual search* to help users formulate their visual intent more conveniently via simple multitouch interactions. This kind of search leverages limited yet natural user interactions on the phone to achieve more effective visual search performance while maintaining satisfying user experience. The proposed *interactive mobile visual search* extends conventional two-step search ("snap → search") one step further, to "snap → interact → search," as shown in Fig. 1. The user only takes one step to interactively specify their intent.

We first conducted a focus study to gain insight into usage patterns and concerned factors on mobile visual search. Typical tasks of *Informational, Transactional*, and *Social* as well as evaluation metrics of *Efficiency, Interface*, and *Effectiveness*

Manuscript received June 28, 2012; revised January 12, 2013; accepted February 25, 2013. Date of publication June 14, 2013; date of current version October 11, 2013. This work was supported in part by the National Basic Research Program of China under Grant2012CB316304 and National Natural Science Foundation of China under Grants 61232013 and 61225009. The associate editor coordinating the review of this manuscript and approving it for publication was Dr. Nicu Sebe.

J. Sang and C. Xu are with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing 100190, China (e-mail: jtsang@nlpr.ia.ac.cn; csxu@nlpr.ia.ac.cn).

T. Mei and S. Li are with Microsoft Research Asia, Beijing 100080, China (e-mail: tmei@microsoft.com; spli@microsoft.com).

Y.-Q. Xu is with Tsinghua University, Beijing 100084, China (e-mail: yqxu@tsinghua.edu.cn).

C. Zhao is with Microsoft, Redmond, WA 98052 USA (e-mail: chzha@microsoft.com).

Color versions of one or more of the figures in this paper are available online at http://ieeexplore.ieee.org.

Digital Object Identifier 10.1109/TMM.2013.2268052

[1] [Online]. Available: http://www.leafsnap.com/

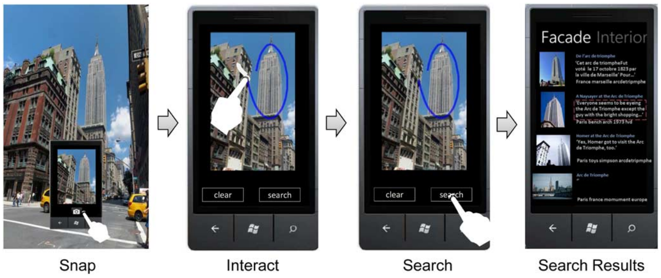


Fig. 1. Process of interactive mobile visual search: snap → interact → search. The prototype is deployed on Window Phone 7 (WP7) devices.

were summarized. The limitations of existing mobile visual searchers were discussed accordingly. Most participants in this study expressed the demands for an touchable or gesture-based interaction, as long as the search performance could be improved via very limited yet natural interactions. Based on that, we introduced four modes of interactions, *crop, line, lasso, tap*, and implemented an interactive visual search system on the multi-touch mobile devices. Usability evaluation study consisting of 40 users was then conducted following the summarized tasks and metrics. We compared the search with interactions (i.e., the proposed four modes) versus conventional non-interaction mode. Armed with the observations, we present design principles for future mobile visual searchers. In summary, the contributions are as follows.

- A focus study is carried out for the first time to investigate how users behave and expect on a mobile visual searcher. The typical tasks and evaluation metrics are summarized accordingly.
- Based on the findings from the focus study, we have identified the needs of interaction and introduced novel interaction modes for mobile visual search.
- An in-depth usability evaluation on the advantage of interaction and comparison between the proposed four interaction modes is performed.
- We give the design principles for future mobile visual search systems.

The remainder of this paper is organized as follows. Section II reviews related work. Section III presents our focus study which initiates this work, while Section IV introduces the proposed TapTell prototype. Section V presents the design of different user interactions on mobile devices. Section VI presents our user studies based on the interaction design, followed by the discussions and conclusions in Section VII and VIII, respectively.

II. Related Work

Here, we explain that there is a gap between the fact people prefer to interact with their mobile devices and the way that most mobile visual search applications work. The goal of this work is to fill the gap by introducing a more natural user interaction.

A. Mobile Visual Search

The Internet has become a necessity for daily life. More people than ever before have the need for pervasive Internet, which creates immense potential for the mobile services as an Internet access point. According to a recent industrial report, studies show that by 2015, 80% of people will access Internet from mobile devices.[2] While search continues to be the main activity on mobile Internet,[3] mobile visual search, where people capture the object in visual proximity and search it via mobile Internet, emerges as a new member of the mobile search family.

Due to its commercial value, mobile visual search has drawn extensive academic and industrial attentions. Early deployments of such commercial applications include Google Goggles,[4] Microsoft Bing Mobile,[5] Kooaba,[6] Nokia Point and Find,[7] Amazon SnapTell,[8] and so on. Most existing applications work on recognizing the picture as a whole, while neglecting to take the advantage of smartphone functionalities such as the multitouch interaction. As a result, they only achieve success in limited vertical domains, e.g., book and CD covers, artwork, and name card. People are not allowed in these applications to express their visual intent through any interaction with the picture. The only system enabling interactive features is Google Goggles, which recently supports cropbox for ROI selection. Based on our investigation of typical usage patterns on mobile visual search, we present a natural user interface, consisting of three novel interaction modes (e.g., *line, tap,* and *lasso*). These interaction modes as well as the *crop* mode of Goggles are explored through a usability evaluation.

In academia, there exists extensive research on the problems in mobile visual search. Topics of interest include the design of compact descriptors or image signatures [15], [21], the incorporation of side information like GPS and RFID [4], inverted index compression and transmission [5], and so on. However, mobile visual search is still in its early stage. State-of-the-art algorithms can work well only on highly textured objects and under controlled lighting conditions. Back to our story of leaf recognition, LeafSnap requires to center the leaf samples on a plain white background and the recognition time for each sample takes at least one minute. Recently, two inspiring works take user operation into consideration to improve mobile visual search performance. Yu *et al.* proposed an active query sensing system to analyze the failed query and guide mobile users to take a second query [23], while in [22] the authors encouraged users to formulate their search intent by puzzling exemplary images. Our work shares similar spirits with these two works in introducing user interaction into the loop. Specifically, our concern in this work is that, besides visual search technique development, we may improve the user experience and system performance by natural interface design and enabling user-centric features, e.g., interactive operations.

B. Interacting With Mobile Devices

User interface design has experienced mechanical switch, keyboard, mouse, pen to touch screen, gesture, move, and so on. Compared with using a mouse, keyboard, or pen, touch,

[2] Ericsson Connectivity Report 2011. [Online]. Available: http://www.ericsson.com/ericsson_connectivity_report_feb2011.pdf

[3] AdMob Mobile Metrics Highlights 2010. [Online]. Available: http://metrics.admob.com/May-2010-AdMob-Mobile-Metrics-Highlights.pdf

[4] [Online]. Available: http://www.google.com/mobile/goggles/.

[5] [Online]. Available: http://www.discoverbing.com/mobile/.

[6] [Online]. Available: http://www.kooaba.com/.

[7] [Online]. Available: http://www.pointandfind.nokia.com/

[8] [Online]. Available: http://www.snaptell.com/

gesture, and body moves are more natural, engaging, and convenient. More and more mobile devices, be they smartphones, PDAs, or tablets enable reactions to pressure, motion, shaking, and even moving in space, which enhance means to observe and engaging with data [26]. The new generation of consumers have been accustomed to the novel means of engaging with information, which provides significant potential of incorporating interaction features into mobile applications.

The powerful interaction-enabled mobile devices have been dramatically changing how we work and entertain on the move. The best-selling mobile game—Angry Birds,[9] serves as a best example of how interaction contributing to a successful mobile application. Its success demonstrates that a simple yet elegant interaction concept can maximize the creativity of mobile users. Beyond game, the importance of mobile interaction is visible in applications from work to daily lives. It has been studied and suggested that visual interfaces will improve mobile-based search in [7]. In multimedia browsing and editing, Jokela *et al.* have identified the importance of interaction design in mobile video editing [16]. The guidelines for the mobile video browsers are discussed in [14], where three types of interface, GUI, gestures and physical movement are investigated. Recently, touch screen gestures were specially designed and implemented to help disabled persons for interaction with digital devices [12], [17]. Mobile interaction also exists in education [8] and hedonic parts of life [13], [19].

In one word, interaction shifts the role of mobile user from a passive information responder to an active, real-time participant. Interestingness and engagement are very important for a winning mobile interface. For mobile visual search to succeed, we need to think beyond simply snapping and towards interfaces offering rich interactions. In this paper, we investigate the advantage of human interactions in mobile visual search scenario and introduce novel operations to interact with mobile devices.

III. Focus Study

To build attractive mobile applications, it is important to know user needs and understand what are key characteristics defining a great mobile application. As a starting point of our work, we conducted a focus study on usage patterns and concerned factors of mobile visual search. The objective is twofold. First, the study aims to gain understanding on typical users needs/tasks for mobile visual search and features/metrics contributing to a successful mobile visual searcher. This reveals information helpful in designing follow-up structuring questionnaires. Second, the study aims to provide some hypotheses according to the tasks and metrics summarized, which can be further tested quantitatively by the usability evaluation study.

A. Participants and Procedure

Eight participants (three female and five male, 20–30 years old) were recruited for the focus study, who were selected based on two criteria: the potential interest for mobile visual search and their previous experiences on mobile visual searchers. The selected participants are all graduate or undergraduate students, with educational backgrounds ranging from computer science, software engineering, to industrial design.

The focus study was loosely structured and conducted in a group discussion which was guided by a moderator. The moderator tried to ask all questions, while thinking aloud was encouraged for the participants and serendipity was allowed. The structure of the focus study questions is as follows.

- *Behavior-oriented discussion*. The discussion started with open-end questions regarding the participants' mobile search habits. For example: How do you explore the unknown but interesting things on the move? What are the differences between mobile search and desktop search? What are the differences between mobile visual search and mobile text/voice search?
- *Task-oriented discussion*. This discussion was used to investigate the typical needs and expectation of users on mobile visual search. These questions have more specific purposes. For example: in what kind of scenarios will you benefit from mobile visual search? What functions do you especially expect the mobile visual searcher to help?
- *Evaluation discussion*. This was related to the metrics of defining a good mobile visual search application. The example questions are: How do you rank an mobile visual searcher? What feature do you care about most?
- *Open discussion*. We encouraged participants to reflect on previous discussions and comments. For example: based on the summarized tasks and metrics, what are the limitations of mobile visual searchers you used? Do you have any suggestion on how to improve the system performance?

B. Results

1) Usage Patterns: We first investigated the common behaviors of the users when they want to obtain information or explore interested things on the move. Five of the eight participants considered textual keyword-based mobile search as their first choice, while the other three preferred to call to friends or ask from surrounding persons for help.

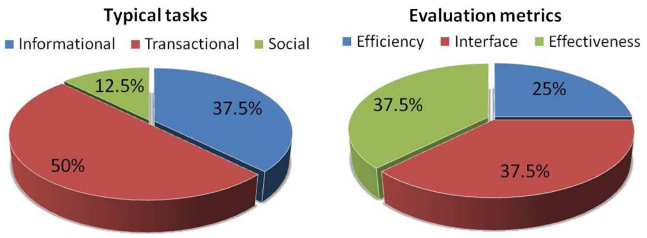
The physical constraint of mobile devices makes user search behaviors quite different from that on desktop PCs. Based on the participants' discussions, searching on mobile devices is more focused while searching on PCs is more exploratory. This is consistent with the conclusion of previous mobile usage pattern study [6].

Regarding the different mobile search methods, the participants argued that each method has its own strength. The selection depends on the tasks and contexts. For example, text-based mobile search is more suitable for non-figurative concepts, e.g., checking weather forecast; while to find out indescribable things or tangible things, e.g., finding a song or a picture stuck in your head,[10] they would prefer voice- and image-based mobile search. However, two participants expressed their concerns about the upload discharge cost for voice- and image-based search. In this session, the participants also showed their confidence about the great potential of mobile visual search-based applications, which was detailed in the following discussions.

2) Tasks and Metrics: The usage pattern investigation above helps participants move smoothly to the tasks and metrics evaluation. For example, discussion on the difference of mobile

[9] [Online]. Available: http://www.rovio.com/index.php?page=angry-birds

[10] [Online]. Available: http://www.shazam.com/



(a) (b)

Fig. 2. Summarized typical tasks and evaluation metrics from focus study. The percentages indicates the proportion of what participants need or care most.

versus desktop search provides insight into the typical tasks for mobile visual search, where highly exploratory tasks like finding lowest-cost airfare will be first excluded. For the evaluation metrics, the limited input means and small screen raise the need for a natural interface, while the mobility nature calls for the need for low discharge cost and power consumption.

Broder came up with a trichotomy of desktop search purposes: navigational, informational, and transactional [2]. Mobile search differs significantly from desktop search, not just because of the devices but because people's different needs on the go. Based on the discussion, we modified the previous trichotomy and categorized the typical tasks on mobile visual search into: *Informational, Transactional*, and *Social*.

- **Informational:** this is to acquire detailed information about the unknown or interesting things, e.g., identifying logo, leaf and artwork, what this landmark is and what story is behind, etc.
- **Transactional:** this is to obtain information to help make decision and conduct actions, e.g., where to serve the cuisine, where to buy the book, CD and movie, comparing shopping, and checking rental information.
- **Social:** this is to communicate with or get to know somebody, e.g., the persons in the same place or people who also like this stuff.[11]

The proportion of the three types of needs is shown in Fig. 2(a). It can be seen that *Transactional* is the most preferred tasks on mobile visual search (50%), which may be due to its on-the-go characteristic—the users have more focused goals and prefer to perform mobile visual search to help make decision. *Social* is a special need evolving with location-based services. Its application scenario is not clearly defined yet. However, during the discussion, one participant put forward a very interesting idea for the social needs (if not violating the privacy issue).

"*I'm too shy to directly strike up a conversation with a lovely girl encountered on the street. With just a snap of the girl, the application will provide with her Facebook homepage so I could get to know her gradually.*"

The evaluation metrics to rate an mobile visual search facility also derived from the focus study, which include *Efficiency*, *Interface*, and *Effectiveness*:

- **Efficiency:** it describes the extent to which efforts are used to produce a specific outcome. The efforts include discharge cost, power consumption and response time.
- **Interface:** a good interface for mobile visual searcher should be natural and intuitive, ease of operation, flexible, attractive and helpful to clarify user intent.
- **Effectiveness:** this is related to search performance (e.g., accuracy), task-dependent search results, as well as personalized search and recommendation results.

For the metric of effectiveness, task-dependence means the mobile visual searcher should be aware of the task context, e.g., history and travel information are expected if the captured photo is a landmark, while the address of shoe store and price information are preferred if the photo is a pair of shoes. To understand the personalization metric, one comment from participant is quoted:

"*Different persons expect different results for the same picture*, e.g., *for a cuisine picture, some want price, some need the recipe, while some like to know the nutrition included*."

The proportion of the evaluation metrics is shown in Fig. 2(b). It is shown that that interface was emphasized by the participants equally to effectiveness (37.5%). This does not surprise us, since, in "Experience is King" era, interface design has been the key factor for mobile applications [11].

3) Hypothesis: The issues most participants complained about include irrelevant returned results and the rigid requirements for capturing environment and skills, e.g., "*The performance is highly sensitive to the capture angle and light condition, which is inconvenient to be used on the move.*" Some useful advice was provided, such as conducting image preprocessing before uploading to server,[12] targeting high precision with low recall and enabling ROI selection after capturing. One of the participants suggested the following.

"*A metaphor is to think of searching pictures for slides on Google Image. Taking photos is analogous to the available pictures online. It's difficult to find a best picture to fit your slides, while you may pick a good one and perform simple editing offline. For me, such an editing feature is desirable on mobile visual search.*"

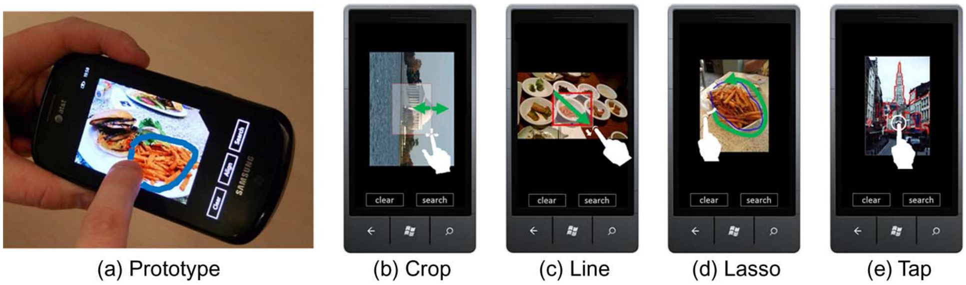
This quote brings out one of the most important motivations to design interactive operations for visual query formulation (i.e., ROI selection), namely helping express user intent more clearly and reducing the requirements for high-quality photograph capture is a viable solution to make up for the limitations of the current visual search techniques. Seven out of eight participants expressed the demands for an interactive feature to query adjustment for helping improve the performance. The only participant who declined this feature was not sure whether interaction could lead to better search results. Inspired from the discussions and comments, we came up with the following hypothesis.

Hypothesis: incorporating user interactions in the loop can improve user experience and the performance of mobile visual search.

Besides the focus study observation, similar hypothesis can also be seduced from the technical point of view. Existing mobile visual search techniques suffer from cluttered background around the interesting region in the photos. While user intent is complex and background substraction techniques are not always satisfactory, appropriate interaction design to enable users

[11] The task of *Social* can be extended to the concept "The Internet of Things", i.e., connecting people by the captured things. The popular mobile applications Color and Instragram actually all build on similar motivations.

[12] For example, deblur, image restoration, and enhancement.



(a) Prototype (b) Crop (c) Line (d) Lasso (e) Tap

Fig. 3. Prototype and different interaction modes for formulating visual intent in the prototype. (a) Prototype of TapTell. (b)–(e) Interaction modes.

with ROI selection will separate the interesting region from the background. This hypothesis serves as our motivation of following work to introduce interaction to the mobile visual search systems.

IV. Prototype

The basic premise behind interaction-enabled interface is that, by allowing users to select their ROI in the picture, the mobile visual search system can understand user intent more clearly and significantly improve search performance. We developed a prototype system, named TapTell, on WP7 devices to investigate the advantage of multitouch human interactions on mobile devices.[13] The prototype including the proposed interaction modes are shown in Fig. 3. The use of TapTell is very easy: 1) users first take a picture; 2) the captured picture will be displayed on the screen, where users can use one of the four interaction modes to select their ROI; and 3) users can further reselect the ROI by clicking the "clear" button or search similar pictures by clicking the "search" button. Therefore, by taking pictures and interacting with the mobile devices, users are able to express their intent by gesture and resort to visual search techniques to connect the captured picture with related information in the physical world.

In the prototype, we develop a scalable image indexing and searching algorithm on the cloud, which is based on a visual vocabulary tree (VT) [24]. The VT is constructed by performing hierarchical K-means clustering on a set of training feature descriptors representative of the database [18]. In our implementation, a total of 50 000 visual words are extracted from 10 million sampled dense scale-invariant feature transform (SIFT) descriptors, which are then used to build a vocabulary tree of six levels of branches and 10 nodes/subbranches for each branch. The storage for the vocabulary tree in cache is approximately 1.7 MB with 168 bytes for each visual word. The VT index scheme provides a fast and scalable mechanism suitable for large-scale and expansible databases. Besides the VT, we also incorporate the image context around user-specified ROI into the indexing scheme, by introducing a novel $tf-idf$ weighting method [25]. Our prototype can handle a large database with tens of million images and the searching time on the cloud is around tens of milliseconds on average.

The database in the prototype consists of two parts: One is from Flickr, which includes a total of 700 000 images from 200 popular landmarks in ten countries, with each image associated with its metadata (title, description, tag, and summarized user comments); The other part is a collection of commercial local businesses from Yelp,[14] which includes 350 000 user-uploaded images (e.g., food and menus) associated with 16 819 restaurants in 12 U.S. cities. The search result interface (see Fig. 1) shows the related information (i.e., title, description, location, and so on) on the right of each searched image. The participants can access the visual search performance through the returned search results.

V. Interaction Design

Observations of the evaluation metrics concerning with interface are valuable for designing new interactive operations. A great mobile visual searcher is supposed to be natural, flexible, and attractive, with simple operation and helpful to express user intent. Equipped with the state-of-the-art multitouch techniques, we introduced four modes of gesture-based interactions [Fig. 3(b)–(e)], among which *crop* is the interaction mode supported by existing mobile visual searchers, while *line, lasso*, and *tap* are newly defined. We will introduce each interaction mode in the following.

Crop: Google Goggles introduced *crop* for ROI selection. A built-in crop tool is enabled by highlighting a block located in the center of the photos. Anchor one finger on the side and drag it till the cropbox contains all the interested regions. By cropping, the unnecessary stuff is taken out and only the desirable regions are uploaded to server for analysis. For comparison convenience, we implemented a similar cropbox in TapTell on the WP7 devices [see Fig. 3(b)].

Line: the implementation of *crop* requires the selected region is located in the center of the picture. Accordingly to the summarized interface design principle, flexibility is limited for this case. To improve the flexibility and make the operation more natural, a novel interaction mode called *line* is developed [see Fig. 3(c)]. Simple as the name suggests, the users are allowed to draw a line across the interested region anywhere in the photo, and the rectangular bounding box whose left-top corner is the starting point and right-bottom corner is the ending point of the line is highlighted and selected. In case of wrong or unsatisfied selection, just draw another line and a new rectangular will replace the former one. *Line* provides a natural, real-world feel of interaction, which is a simulation of pen for freeform inputs.

[13] The WP7 device has 1-GHz processor, 512-MB RAM, 5-MB pixel digital camera, and 4-in WVGA Super AMOLED touch screen. A demo video is available at http://www.youtube.com/watch?v=SDIIsmESGEQ.

[14] [Online]. Available: http://www.yelp.com.

Lasso: both *crop* and *line* follow a rigid square selection scheme. Though operated easily, the fixed rectangular inevitably contains redundant space, which will confuse the recognition algorithm. To address this, *lasso* is designed for an arbitrary shape selection [see Fig. 3(d)]. The operation is equally simple: draw a continuous curve around the interesting object or region you want to search for and re-draw if not satisfied. Once you've got your ROI outlined, click "search" button and check the results. Note the curve is not necessarily closed up. So long as it is a continuous drawing, the boundary of the selected region will be obtained by analyzing the sample points from the trace.[15] The design of *lasso* is inspired by the Lasso selection tool for interactive Bing search on iPad.[16] The new feature was impressive and has received lots of positive responses since its release, where users find the Lasso operation very natural and engaging.

Tap: each input device has its strengths and weaknesses. While keyboard is best for text input, mouse for precise pointing, touch-based gesture is best for object manipulation and freeform expression. However, you cannot depend on gesture for precise selection. Consider *lasso* for example, although it supports an arbitrary shape selection, it's tedious and nearly impossible to draw along the accurate boundary of the interested object. Moreover, the existing touchscreen technologies do not support such a precise response. An interaction mode called *tap* is thereby designed. The photographs are automatically segmented once captured, and the boundaries of the segments are presented to users for selection [see Fig. 3(e)]. Users choose segments by just tapping on them. The selected segments will highlight their boundaries for visualization. In case of unsatisfied selection, users could press the "clear" button or retap the selected segments to cancel the current choice. If the other three interaction modes are analogous to user inputting queries, *tap* is analogous to query suggestion, where image processing algorithms rise to help.

We compared the usability and user experience on the TapTell system with the four interaction modes. The user study settings and results are detailed in the following section.

VI. Usability Evaluation Study

The interaction concepts described above have been evaluated in a controlled experiment from July to August in 2011. We invited 40 users in the user study. It takes each user about 1.5 hours to finish the study. The user study has two goals: 1) to validate the hypothesis proposed in the focus study that interaction is necessary and useful for a mobile visual searcher and 2) to compare among different interaction modes and find which one is more natural and effective. The users were asked to perform different tasks with each interaction mode and fill in the questionnaires as the quantitative feedback on the usability. We recorded user actions in log files, which were jointly utilized for evaluation from an objective view.

A. User Study Settings

We recruited 40 subjects (22 females and 18 males, all 20–34 years old) to conduct the usability evaluation study. The backgrounds of the participants range from programmer, buyer, editor, accountant, secretarial, human resources staffer to undergraduates and graduates in computer science, animation, design, geology, and social sciences. The criterion for selecting subjects was that they should be frequent mobile users and very familiar with mobile search.[17] We organized the participants into two groups (20 per group) with balanced gender ratio for each. Participants of group A (control group) were asked to perform the study on the designed mobile visual searcher without interaction function, while participants of group B (experimental group) on the same mobile visual searcher enabled with the four interaction modes.

Based on the summarized typical needs for mobile visual search, one informational task and one transactional task were defined.

- **Task 1**: "*Landmark recognition*." For the informational task, we simulated a tour scenario. The colorful pictures of 40 popular landmarks from four countries (U.S., Italy, France, and China) were posted on the wall around the testing room. The participants (who have visited the tested country) were assumed traveling to these landmarks, and the task is to perform landmark recognition using the provided mobile visual search system.
- **Task 2:** "*Cuisine, on the go*." For the transactional task, a food service scenario was set up. Each participant was provided a color-printed booklet with 40 cuisines including salad, soup, burgers, meat, dessert, and drinks. They were assumed to be passing by a cuisine billboard on the street, but can neither speak the language nor know the cuisine. The task is to take a photograph of the cuisine and resorting to mobile visual search for providing information about the nearest restaurant serving this cuisine.

For participants in group A, five to eight subtasks were required for each task. For participants in group B, three to five subtasks were required for each interaction modes inside each task. The order in which the interaction modes were selected is random and decided by participants themselves. We defined a subtask as this: the participants starts one subtask by picking one landmark or cuisine and taking a picture, and ends when satisfied with the results or disappointed with the results and decided to give up trying.[18] The diagram of user study design is illustrated in Fig. 4. Therefore, it is likely that participants performed several round of "snap-interact-search" process inside one subtask. Each subtask corresponds to one log file. Besides the search times, snap, search timestamp, and interaction count of each try, the result of the subtask as well as the User id, selected Interaction mode and Task id were also recorded. The time interval between the first snap timestamp and the last

[15] We have developed a new visual search technique based on the "lasso" region using a visual vocabulary tree-based approach. Please refer to [25] for the implementation details.

[16] Lasso for Bing search on iPad: http://www.bingforipad.com/blog/2011/07/05/bing-for-ipad-searching-without-search-box/

[17] Users of smartphones with touchscreen are encouraged.

[18] An ending interface with decision buttons of "accept" and "refuse" following the results interface is used to end subtasks and submit log data. The participants were allowed to conduct another trial for the subtask if the current search results are not good, till tired of trying or getting satisfied results. The maximum times for trying depends on individual patience and varies with participants and tasks. During the user study, we gave no hint to the users about how to determine one result should be "accept" or not, since users' subjectivity guarantees the experimental results' objectivity.

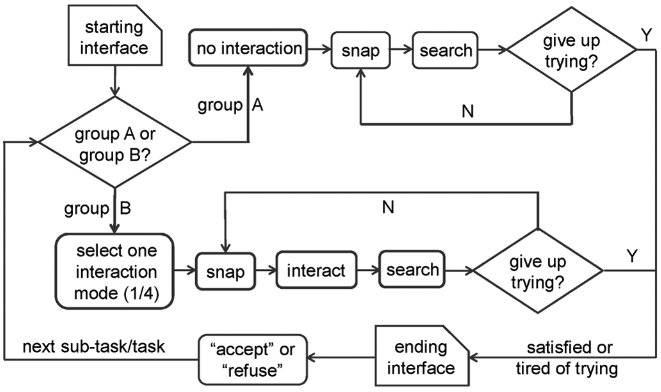


Fig. 4. Diagram of user study.

search timestamp is the search time user costs for one subtask. The structure of the log file is shown in Fig. 5.

When finishing the two tasks, the participants were asked to fill in three questionnaires: the responses to the summarized tasks and metrics, the standard usability scale (SUS) questionnaire [3], and the user experience survey based on the summarized evaluation metrics. To further validate the quality of the summarized tasks and investigate the participants' preferences, for each of the three tasks and metrics, the participants were first asked to rate their level of needs and care from 1 to 7, where 1 is not needed or cared and 7 is most needed. SUS is a standard questionnaire giving a global view of subjective assessments of effectiveness, efficiency, and satisfaction. It coordinates well with other subjective measures and has been used for a variety of research projects and industrial evaluations. SUS yields a single score representing a composite measure of the overall usability of the system being studied. It has a range of 0 to 100, the higher the better. We utilized a seven-point Likert scale and the participants were asked to specify their level of agreement to each questions [20]. Focusing on the mobile visual search application, according to the summarized evaluation metrics from focus study, we designed a user experience covering the factors of (1 *Efficiency*: *speed*, (2 *Interaction*: *user-friendly, ease of operation, flexibility, novelty, clear user intent*, and (3 *Effectiveness*: *relevance of search results*. Since the main goal of the work is to investigate the advantage of interaction and find the most usable interaction mode, the metrics concerning with *Interface* are emphasized and detailed in five subfactors. The participants were asked to rate the metrics with a scale of 1 to 7, where 1 is the worst and 7 is the best.

Note that we do not set up a hybrid mode by combining "Interaction" with "No Interaction". The reason is twofold: 1) since one of the goal is to examine which interaction mode contributes best, we need to combine "No Interaction" with each of the four interaction modes, which makes user study lengthy and diffuse and 2) it is not easy to set up a hybrid mode, e.g., how to switch between "Interaction" and "No Interaction." In addition, the mixed design will confuse the participants. However, we emphasize that in real interaction implementation, a carefully designed hybrid mode will combine the advantages of both "Interaction" and "No Interaction" and achieve best balance between effectiveness and efficiency.

B. Results

The statistical evaluation results include three parts: the responses for the summarized tasks and metrics, the subjective measures of SUS and user experience questionnaire, and the objective measure of the log analysis.

1) Summarized Tasks and Metrics: In the focus study, three types of typical tasks and evaluation metrics had been concluded from eight participants' discussions. In the usability evaluation study, the 40 participants rated the level of preferences over them, which is summarized in Figs. 6 and 7. From the average rates in Fig. 6(a), we can see that: 1) for the typical tasks, *Informational* and *Transactional* are considered equally useful, while *Social* is not so needed and 2) the participants attached emphasis on all three concerned factors, while *Effectiveness* and *Interface* are considered a little more important. It is shown in Fig. 6(b) that the participants from the usability study showed a similar preference for the summarized tasks and metrics with the focus study discussions. The percentages of participants who care the most are consistent with those of Fig. 2. To investigate the affection of backgrounds, we also examined the preferred tasks and metrics regarding different genders and occupations. The statistics for female and male participants are shown in Fig. 7. It is shown that female participants have higher "Transactional" need and value more on "Interface" metric, while some male participants emphasize on the "Social" need and generally care more about "Effectiveness" than about "Interface" metric. Regarding different occupations, we found that the participants with science-related backgrounds emphasize more on the "Informational" need and "Effectiveness" metric, while participants with arts-related backgrounds emphasize more on the "Transactional" need and "Interface" metric.

The participants responded that the summarized tasks and metrics basically covered their intent on a mobile visual searcher. However, some really interesting applications and suggestions were proposed. For example, one participant mentioned his need of solving mathematical equations as follows.

"*While it is difficult to issue an equation query using the search box, taking a photo of the equation and resort to visual analysis for help maybe a cool idea.*"

We note that the summary and observations in this work only serve as a guideline for the design of mobile visual search, not expected to fit all the increasing new situations or requirements. Besides the new application scenario, the participants also commented on the demands of supporting more categories and targeting at various user groups. Combining the discussions in the focus study, we concluded with the following observations with tasks and evaluation metrics.

Observation 1: compared with desktop search, mobile visual search has a significant characteristic of helping users make decision. The *Transactional* interests are as important as the *Informational* interests. *Social* represents a novel application scenario combining mobile visual search with mobile SNS, which has a great potential in the future. Besides, tasks related to visible stuffs in the physical world are encouraged, regarding to the mechanics of mobile visual search.

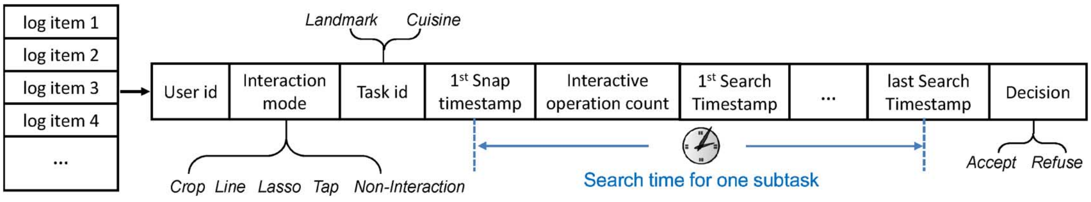
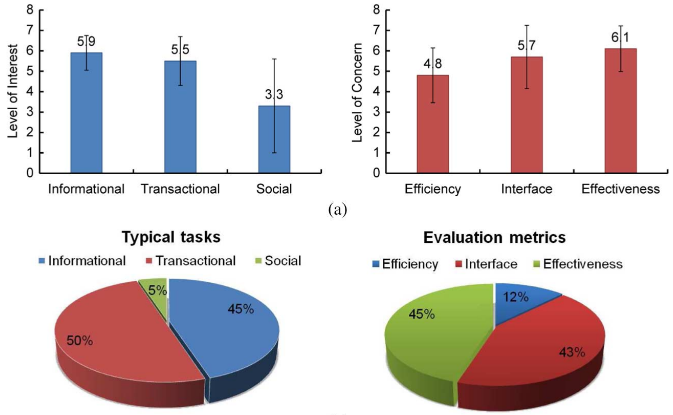


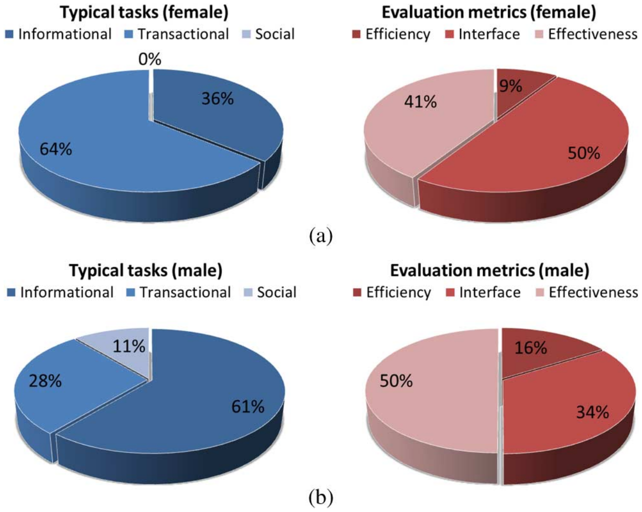
Fig. 5. Structure of the log file.



(a)

(b)

Fig. 6. Preferences over the summarized tasks and evaluation metrics for the formal user study. (a) Average rate. (b) Percentage of participants need or care most.



(a)

(b)

Fig. 7. Preferences over the summarized tasks and evaluation metrics for the formal user study. (a) Female participants. (b) Male participants.

Observation 2: for the mobile visual search users, *Interface* design is equally important as *Effectiveness*. Due to the mobility, *Efficiency* is also one factor users care about. In addition, on condition of current techniques, results of high precision with low recall are preferred.

2) Subjective Evaluation: **SUS Results**. Fig. 8 shows the average SUS scores and the standard derivatives (SD) for group A (*No Interaction*) and the four interaction modes of group B. It is seen that three out of four interaction modes (i.e., *crop, line, lasso*) obtained a higher SUS score than *No Interaction*. ANOVA is a widely used technique for the statistical significance test [1].[19] ANOVA test revealed that the results of SUS survey are significant and confident (see Table I). The modes of *line* (SUS $= 73.75$, SD $= 15.49$) and *lasso* (SUS $= 78.17$, SD $= 17.13$) were perceived as the most usable. Observed from the detailed responses, we find that: 1) the reason lowering the SUS score of *No Interaction* is that many participants "do not find very confident using the system" and 2) for the *tap* mode, some participants "think it cumbersome to use."

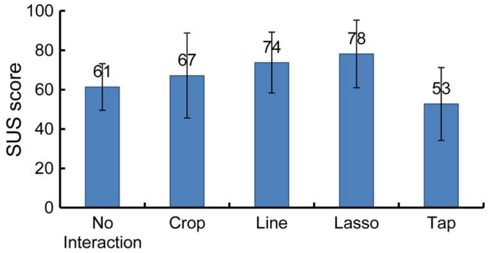


Fig. 8. Average SUS scores for different settings.

TABLE I
ANOVA RESULTS FOR SUBJECTIVE EVALUATION

		F	df	Sig.
SUS test		6.69	4	< 0.001
User experience	Speed	22.21	4	< 0.001
	user-friendly	3.28	4	< 0.05
	operation	1.9	4	> 0.05
	flexibility	11.2	4	< 0.001
	novelty	4.93	4	< 0.01
	clear intent	7.68	4	< 0.001
	result relevance	7.74	4	< 0.001
	overall Rate	17.2	4	< 0.001

User Experience Results: the statistics of user experience questionnaire and its ANOVA test are shown in Fig. 9 and Table I. Except for metric of "Ease of Operation" $(\mathrm{Sig.} > 0.05)$, other metrics show statistically significant results. Similar with the SUS results, except *tap*, the other three interaction modes are all superior to *No Interaction* mode. *No Interaction* received comparable rates only on the metrics of "Speed" and "User-friendly." Its overall rate of 3.76 is much lower than that of the other four modes (4.79). Out of the eight metrics, *lasso* is considered to be the most attractive in seven, while *tap* has the highest score on "Novelty." *Lasso* obtains a high average score of 5.95. Preceded by *lasso, line* is the second-best interaction mode, with an average score of 5.49 and overall rate of 5.65.

[19] The value of F denotes the significance of the result, the larger F is, more significant the result is. df is short for degree of freedom. $Sig.$ denotes the probability we should deny the hypothesis, the smaller $Sig.$ is, more confident the result is.

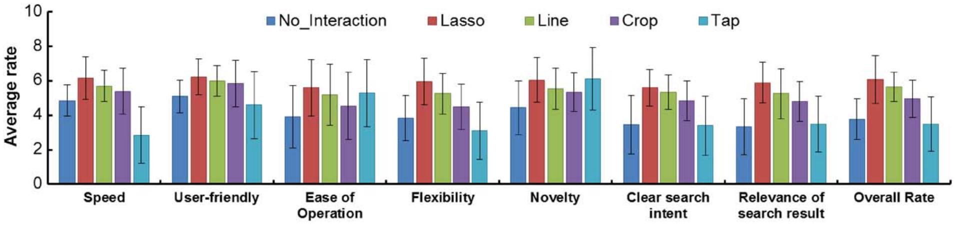


Fig. 9. Average rates for the user experience survey.

Tap has a poor user experience concerning "Speed", "Flexibility", etc. However, for the metrics of "Ease of Operation" and "Novelty", *tap* has a high evaluation score, which shows the potential of this interaction mode.

Sampled participants' comments are quoted in the following to illustrate the subjective results.

- "*It's boring repeatedly taking photos of the same object. I need to modify my capturing distance and angle all along.*"[20]
- "*Crop is similar to Line, except that Crop restricts the cropbox in the center of the photo, which is not very convenient.*"
- "*Tap is too restrictive. I can only select one of the segments the system suggests.*"
- "*While Line and Lasso are natural by solely using fingers, Crop can help locate the ROI by initializing a box.*"
- "*Tap is interesting, but the segmentation result is confusing. What's more, the small segment is difficult to target for selection.*"

The participants expressed similar complaints on *No Interaction* with the focus study. The four interaction modes each have advantages and disadvantages. For *tap*, the interaction concept is considered appealing, while it is the poor segmentation results that exacerbates the user experience. It is out of the scope of this paper to discuss how to create this segments, since we focus on the interaction design. However, we are working towards improving the *tap* implementations by combining with other interaction mode in the future.

Combining with the user questionnaire statistics and related comments, we have the following two observations.

Observation 3: a mobile visual searcher enabled with interaction is more useful than that without interaction. Interaction contributes to a more easy and focused selection, which leads to better matching results.

Observation 4: a good interaction mode should allow users to locate their ROI easily, precisely and quickly. Among the introduced four types of gesture-based interactive operations, *lasso* is considered to be the best considering efficiency, interface, and effectiveness.

3) Log Data Analysis: Besides the subjective user survey questionnaires, we also assessed the usability and user experience by analyzing the log data. Corresponding to the three factors of *Efficiency, Interface,* and *Effectiveness*, the log data analysis also includes three parts: search times and search time cost, interactive operation count, and "accept" ratio and normalized discounted cumulative gain (NDCG).

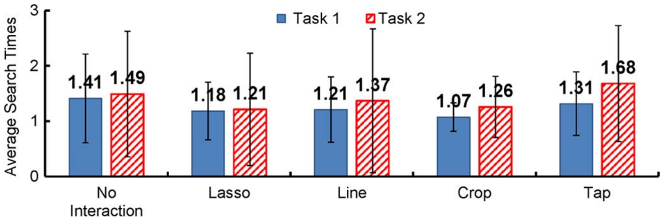


Fig. 10. Average search times for one subtask.

TABLE II
ANOVA RESULTS FOR LOG DATA ANALYSIS

	Task	F	df	Sig.
Search times	1	0.97	4	> 0.05
	2	2.31	4	> 0.05
Time cost	1	3.75	4	< 0.01
	2	6.93	4	< 0.001
Operation count	1	6.79	3	< 0.001
	2	4.97	3	< 0.001

Efficiency: search times and search time cost: search times, i.e., how many times the participants performed the "snap → (interact →) search" process before they find the useful information, reflects the ability that the application helps clear user search intent.[21] Fig. 10 shows the average search times statistic. It is shown that interaction enabled settings basically need equal or less search times than *No Interaction*. However, the unsatisfactory segmentation results from *tap* sometimes confuse users, which result in the higher search times of *tap* mode. ANOVA test shows that the reduction of search times is not very significant (see Table II).

Search time cost measures the time used for the participants to finish one subtask, which is very important for mobile applications. The results are illustrated in Fig. 11 and Table II. Actually, interactions sacrifices users' time and operation complexity for better search results. Therefore, the interaction modes have no apparent advantages regarding the search time cost. Subtasks finished by using *crop* and *lasso* cost the least time, while once again, for the *tap* mode, since the segmentation algorithm takes 1–8 s to process one photograph, the overall search timecost becomes higher accordingly.

Interface: interactive operation count: the count of interactive operations measures the operation complexity and inter-

[20] This comment is quoted from one participant from group A.

[21] To calculate search times, search time cost, and interactive operation count, we only consider the "accepted" subtasks.

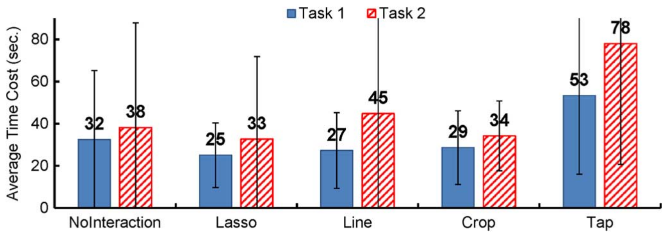


Fig. 11. Average search time cost for one sub-task.

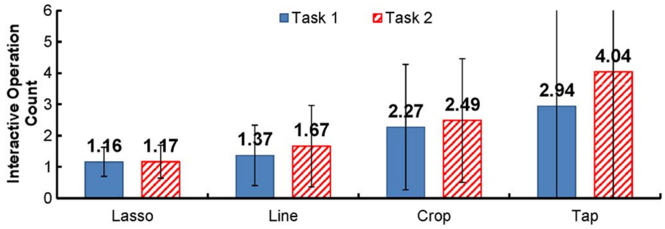


Fig. 12. Interactive operation count for one trial.

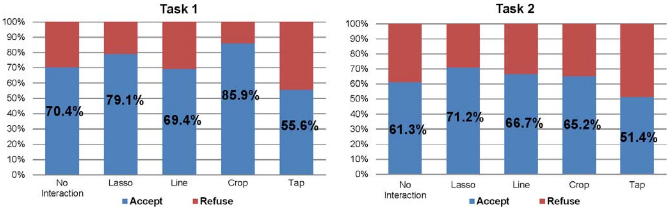


Fig. 13. “Accept” ratio of different settings.

face user-friendlyness. From ANOVA test in Table II, the operation count varies significantly among the four interaction modes. Fig. 12 shows that most of the participants with *lasso* just need one interactive operation (the average count is 1.16 and 1.17 for the two tasks), which is less than the others. Using *crop*, the participants have to anchor and drag the cropbox several times to modify the selection. The interactive operation needed for *crop* (2.27 and 2.49) is nearly two times of those for *lasso*. For the high interactive operation count of *tap*, one reason is that most of the photos were over-segmented and not convenient for selection, the other is that participants found it hard to tap one segment to include all they were interested.

Effectiveness: “accept” ratio and NDCG: since the users are allowed to select “accept” or “refuse” to end a subtask, the “accept” ratio is an important indication for the overall performance. Fig. 13 shows the “accept” ratio of different settings for the two tasks. The results are consistent with the user subjective survey data that the interaction modes are generally better than *No Interaction*, and *lasso* performs best among the four interaction modes. The rectangular selection of *crop* and *line* can not conform to the arbitrary shape of objects and interested regions. In addition, it is interesting to find that the “Landmark recognition” task has a consistent higher “accept” ratio than the “Cuisine, on the go” task does. This may be because that the appearances of even the same cuisine change from restaurant to restaurant.

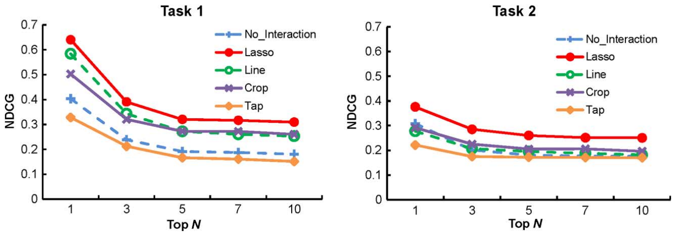


Fig. 14. NDCG evaluation of different settings.

To objectively access the effectiveness of different settings, we also use information retrieval metric of NDCG to evaluate the relevance of the search results. Given a query q, the NDCG at depth p in the ranked list is defined by

$$\mathrm{NDCG}@p = Z_p \sum_{i=1}^{p} \frac{2^{\mathrm{rel}_i} - 1}{\log_2(1+i)} \tag{1}$$

where rel_i is the graded relevance of result at position i, and Z_p is a normalization constant and chosen so that $\mathrm{NDCG}@p$ of a perfect ranking is 1. The captured photographs, the selected ROI, and the top 100 matching results for each of the 928 searches were stored with the log data. We hired three graduates to label the relevance of the top ten results to each captured photographs with a scale of 1–5, where 1 is not relevant and 5 is very relevant. The final relevance value rel_i is averaged over the three labels. The average results are shown in Fig. 14. It accords with the “accept” ratio and demonstrates the importance of interaction. For task 2, the top 1 of *No Interaction* achieves a competitive NDCG score of 0.31. This is because the sampled cuisine pictures were taken from the indexed dataset, and it is relatively easy to find the duplicate images.

By analyzing the log data, we come up with further observations as follows.

Observation 5: with suitable interaction design, at the expense of limited operation cost, the mobile visual searcher is expected to obtain better matching results and user experience. Combining with the significant ANOVA results of subjective evaluation, we validate the Hypothesis summarized from the focus study.

Observation 6: among the four introduced interaction modes, *lasso* has the least interactive operation count and best performance. The consistence with the highest subjective evaluation for *lasso* demonstrates the users’ preferences over simple yet effective interactive operations.

VII. Discussions

Based on the above analysis and observations, this work has several key takeaways for designing future mobile visual search applications.

Mobile visual search calls for evolution: interaction with human in the loop serves as one of the solutions: apparently, the visual search techniques can not keep up with the increasing needs on mobile visual search. Evolution beyond technological development is desirable. The user study in this paper shows that suitable interaction design contributes to better mobile visual search performance due to the clarified understanding of user’s search intent. The presented interactive mobile visual search prototype represented one of the first works in this new paradigm.

Balance between natural user experience and effective visual search performance: *lasso* is the winner: following the users' usage patterns and mobile devices' support of flexible input means, touch or gesture-based interactive operations are preferred. Mobile users can only afford simple interactions when mobility, the less time and operation cost, the better. Most importantly, users expect improved performance by interaction. A good interaction concept should help users modify their ROI selection and lead to a better search result. If no improved performance is promised, interaction will be considered burdensome instead of helpful. We have shown that *lasso* achieved the best balance among the four modes.

Mobile users crave easy-task-completion beyond mobile visual search: there are huge business potentials in mobile visual search market: The mobile visual search has significant applications concerning with transaction. Mobile users are likely to use it to help make decision. The design of the search result details as well as the user interface are expected to match the users' needs on the go. It is helpful to link the search results with local business in the physical world. People need more information beyond mere searched similar pictures, e.g., recommendation of local businesses, so that they can explore surroundings and complete related tasks very conveniently [9].

VIII. Conclusion and Future Works

The aim of this paper was to adopt a broad investigation on mobile visual search, from the typical needs, evaluation metrics to interaction design and interaction modes comparison. *Transactional* was considered as the one preferred need on a mobile visual search, while *Interface* was given equal, if not more, importance with *Effectiveness*. The subjective user survey and the objective log analysis results demonstrate the advantage of interaction and show that interaction mode of *lasso* obtains the best user experience and performance.

In this paper, we intend to utilize the multitouch feature of mobile devices to design different gesture-based interactions. While *line* and *crop* are basically the same, *lasso* enables users with arbitrary selection, *tap* is analogous to segment suggestion and distinguishes itself from the others. Though *tap* gets poor performance due to the implementation issue, the participants have shown major interests to it and approved its novelty and convenience. Moreover, the freedom selection of *lasso* is new to mobile visual search and obtained satisfied performance. We will work on a novel interactive operation combining the idea of *tap* and *lasso*, i.e., users are allowed to conduct a *lasso* alike operation to connect the suggested multiple segments. This will integrates the advantage of *tap*'s automatical suggestion with precise boundary and *lasso*'s naturalness and flexibility.

Understanding user intent on the go is not trivial for practical mobile applications. We strive to make clear the user intent by interactive operations of ROI selection in this work. However, a complex scenario cannot be identified by simply selecting a salient area, for example, if one user takes a picture and boxes out the area of food, does he/she want to perform an *Informational* task or a *Transactional* task? In this case, incorporating context information, e.g., time, location, weather, and even a user's mood will be helpful, which is also one of our ongoing works.

Another interesting direction is to employ image-processing techniques for query image classification. Different vertical search modules will be called and corresponding results information will be provided based on the query category. In addition, although "snap" is the usual input option for a mobile visual searcher, it should not be the only one. It is interesting to enable alternative input options and implement mobile visual search as plug-ins to other mobile applications, e.g., mobile web browser, video player, and image browser.

References

[1] R. A. Bailey, *Design of Comparative Experiments*. Cambridge, U.K.: Cambridge Univ., 2009.

[2] A. Z. Broder, "A taxonomy of web search," *SIGIR Forum*, vol. 36, no. 2, pp. 3–10, 2002.

[3] J. Brooke, *SUS: A "Quick and Dirty" Usability Scale*. London, U.K.: Taylor and Francis, 1996.

[4] D. M. Chen, G. Baatz, K. Köser, S. S. Tsai, R. Vedantham, T. Pylvänäinen, K. Roimela, X. Chen, J. Bach, M. Pollefeys, B. Girod, and R. Grzeszczuk, "City-scale landmark identification on mobile devices," in *Proc. CVPR*, 2011, pp. 737–744.

[5] D. M. Chen, S. S. Tsai, V. Chandrasekhar, G. Takacs, R. Vedantham, R. Grzeszczuk, and B. Girod, "Inverted index compression for scalable image matching," in *Proc. DCC*, 2010, p. 525.

[6] K. Church, B. Smyth, K. Bradley, and P. Cotter, "A large scale study of European mobile search behaviour," in *Proc. Mobile HCI*, 2008, pp. 13–22.

[7] K. Church, B. Smyth, and N. Oliver, "Visual interfaces for improved mobile search," in *Proc. Workshop Visual Interfaces Social and Semantic Web*, 2009, pp. 1–10.

[8] E. del Carmen Valderrama Bahamóndez, C. Winkler, and A. Schmidt, "Utilizing multimedia capabilities of mobile phones to support teaching in schools in rural panama," in *Proc. CHI*, 2011, pp. 935–944.

[9] A. K. Dey, J. Hightower, E. de Lara, and N. Davies, "Location-based services," *IEEE Pervasive Computing*, vol. 9, no. 1, pp. 11–12, Jan. 2010.

[10] B. Girod, V. Chandrasekhar, R. Grzeszczuk, and Y. A. Reznik, "Mobile visual search: Architectures, technologies, and the emerging MPEG standard," *IEEE MultiMedia*, vol. 18, no. 3, pp. 86–94, 2011.

[11] J. Gong and P. Tarasewich, "Guidelines for handheld mobile device interface design," in *Proc. DSI Annual Meeting*, 2004, pp. 3751–3756.

[12] T. J. V. Guerreiro, H. Nicolau, J. A. Jorge, and D. Gonçalves, "Assessing mobile touch interfaces for tetraplegics," in *Proc. Mobile HCI*, 2010, pp. 31–34.

[13] F. Heinrichs, J. Schning, and D. Schreiber, "The hybrid shopping list: Bridging the gap between physical and digital shopping lists," in *Proc. Human-Comput. Interaction With Mobile Devices and Services*, 2011, pp. 251–254.

[14] J. Huber, J. Steimle, and M. Mühlhäuser, "Toward more efficient user interfaces for mobile video browsing: An in-depth exploration of the design space," in *Proc. ACM Multimedia*, 2010, pp. 341–350.

[15] R. Ji, L.-Y. Duan, J. Chen, H. Yao, T. Huang, and W. Gao, "Learning compact visual descriptor for low bit rate mobile landmark search," in *Proc. IJCAI*, 2011, pp. 2456–2463.

[16] T. Jokela, J. Lehikoinen, and H. Korhonen, "Mobile multimedia presentation editor: Enabling creation of audio-visual stories on mobile devices," in *Proc. SIGCHI*, 2008, pp. 63–72.

[17] S. K. Kane, J. O. Wobbrock, and R. E. Ladner, "Usable gestures for blind people: Understanding preference and performance," in *Proc. SIGCHI*, 2011, pp. 413–422.

[18] D. Nistér and H. Stewénius, "Scalable recognition with a vocabulary tree," in *Proc. CVPR*, 2006, pp. 2161–2168.

[19] M. G. Petersen, A. B. Lynggaard, P. G. Krogh, and I. W. Winther, "Tactics for homing in mobile life: A fieldwalk study of extremely mobile people," in *Proc. Mobile HCI*, 2010, pp. 265–274.

[20] L. Rensis, "A technique for the measurement of attitudes," *Archives Psychol.*, vol. 140, pp. 1–55, 1932.

[21] S. S. Tsai, D. M. Chen, V. Chandrasekhar, G. Takacs, N.-M. Cheung, R. Vedantham, R. Grzeszczuk, and B. Girod, "Mobile product recognition," in *Proc. ACM Multimedia*, 2010, pp. 1587–1590.

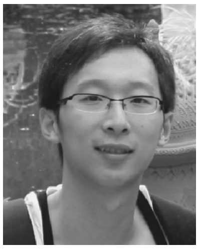
[22] Y. Wang, T. Mei, J. Wang, H. Li, and S. Li, "Jigsaw: Interactive mobile visual search with multimodal queries," in *Proc. ACM Multimedia*, 2011, pp. 73–82.

[23] F. X. Yu, R. Ji, and S.-F. Chang, "Active query sensing for mobile location search," in *Proc. ACM Multimedia*, 2011, pp. 3–12.

[24] N. Zhang, T. Mei, X.-S. Hua, L. Guan, and S. Li, "Tap-to-Search: Interactive and contextual visual search on mobile devices," in *Proc. IEEE Int. Workshop Multimedia Signal Process.*, 2011, pp. 1–5.

[25] N. Zhang, T. Mei, X. Hua, L. Guan, and S. Li, "Interactive mobile visual search for social activities completion using query image contextual model," in *Proc. MMSP*, 2012, pp. 238–243.

[26] J. Zimmerman, J. Forlizzi, and S. Evenson, "Research through design as a method for interaction design research in human-computer-interface," in *Proc. SIGCHI*, 2007, pp. 493–502.



Jitao Sang received the B.E. degree from the SouthEast University, Nanjing, China, in 2007, and the Ph.D. degree from the Institute of Automation, Chinese Academy of Sciences, Beijing, China, in 2012.

He is an Assistant Professor with the National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China. His research interests include multimedia content analysis, social media mining and social network analysis. He has authored and coauthored several refereed research papers in these areas.

Dr. Sang coauthored the Best Student Paper at Internet Multimedia Modeling 2013 and was the Best Paper Candidate at ACM Multimedia 2012. He has served as a special session organizer at MMM 2013 and publication chair at ACM ICIMCS 2013. He was awarded the Special Prize of President Scholarship by Chinese Academy of Sciences.

Tao Mei (M'07–SM'11) received the B.E. degree in automation and Ph.D. degree in pattern recognition and intelligent systems from the University of Science and Technology of China, Hefei, China, in 2001 and 2006, respectively.

He is currently a Researcher with Microsoft Research Asia, Beijing, China. He has authored or coauthored over 140 papers in journals and conferences and has authored eight book chapters. He holds five U.S. patents and more than 30 pending. His current research interests include multimedia information retrieval and computer vision.

Dr. Mei is a Senior Member of the Association for Computing Machinery (ACM). He .was the recipient of several paper awards from prestigious multimedia conferences, including the Best Paper Award and the Best Demonstration Award at ACM Multimedia 2007, the Best Poster Paper Award at IEEE MMSP 2008, the Best Paper Award at ACM Multimedia 2009, the Top 10% Paper Award at IEEE MMSP 2012, the Best Paper Award at ACM ICIMCS 2012, the Best Student Paper Award at IEEE VCIP 2012, and the Best Paper Finalist at ACM Multimedia 2012. He is an associate editor of *Neurocomputing* and the *Journal of Multimedia*, a Guest Editor of the IEEE TRANSACTIONS ON MULTIMEDIA, the *IEEE Multimedia Magazine*, the ACM/Springer *Multimedia Systems*, and the *Journal of Visual Communication and Image Representation*. He is the Program Co-Chair of MMM 2013, and the General Co-Chair of ACM ICIMCS 2013.

Ying-Qing Xu (SM'08) received the B.Sc. degree in mathematics from Jilin University, Changchun, China, and the Ph.D. degree in computer graphics from the Academia Sinica, Beijing, China.

He is currently a Professor with the Department of Information Art and Design, Tsinghua University, Beijing, China, where he has worked since October 2011. He was a Researcher with Microsoft Research Asia from 1999–2011. He has coauthored over 70 papers and holds over 20 U.S. patents. His research interests are in computer graphics, computer vision, human computer interaction, and multimedia.

Chen Zhao received the B.S. degree in psychology from Peking University, Beijing, China, and the Ph.D. degree in human factors from the Institute of Psychology, Chinese Academic Sciences, Beijing, China.

She is a User Experience Researcher with the Office Design division of Microsoft, Redmond, WA, USA. Before she joined Microsoft, she was a Lead Researcher with the Human Computer Interaction group, Microsoft Research Asia (MSRA). Her main research interest is cross-culture design and enterprise social computing. She was the Manager of the User Experience team at IBM China Research Lab before she joined MSRA in May 2008. She started and built the first UX team and usability lab at IBM China.

Dr. Zhao has been an avid participant in the ACM SIGCHI community. She is a founder of ACM SIGCHI China Chapter and served as the vice chair. She was a co-chair for CSCW 2011, an associate chair for CSCW 2012 and 2010 and CHI 2011 and 2012, proceeding cochair for the International Conference of Intercultural Collaboration, known as IWIC before) 2010 and poster chair for IWIC2009.

Changsheng Xu (M'97–SM'99) is a Professor with the National Lab of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, Beijing, China, and Executive Director of the China-Singapore Institute of Digital Media. His research interests include multimedia content analysis/indexing/retrieval, pattern recognition and computer vision. He has 30 patents either granted or pending and has authored and coauthored over 200 refereed research papers in these areas.

Dr. Xu is ACM Distinguished Scientist. He is an associate editor of the IEEE TRANSACTIONS ON MULTIMEDIA and *ACM Transactions on Multimedia Computing, Communications and Applications*. He served as a program chair of ACM Multimedia 2009. He has served as an associate editor, guest editor, general chair, program chair, area/track chair, special session organizer, session chair and TPC member for over 20 IEEE and ACM prestigious multimedia journals, conferences and workshops.

Shipeng Li (F'11) received the B.S. and M.S. degrees from the University of Science and Technology of China, Hefei, China, in 1988 and 1991, respectively, and the Ph.D. degree from Lehigh University, Bethlehem, PA, USA, in 1996, all in electrical engineering.

He joined Microsoft Research Asia (MSRA), Beijing, China, in May 1999. He is now a Principal Researcher and Research Manager of the Media Computing group. He also serves as the Research Area Manager coordinating the multimedia research activities at MSRA. From October 1996 to May 1999, he was with the Multimedia Technology Laboratory, Sarnoff Corporation (formerly David Sarnoff Research Center and RCA Laboratories) as a Member of the Technical Staff. He has been actively involved in research and development in broad multimedia areas. He has made several major contributions adopted by MPEG-4 and H.264 standards. He invented and developed the world first cost-effective high-quality legacy HDTV decoder in 1998. He started P2P streaming research at MSRA as early as in August 2000. He led the building of the first working scalable video streaming prototype across the Pacific Ocean in 2001. He has been an advocate of scalable coding format and is instrumental in the SVC extension of H.264/AVC standard. He first proposed the 694; Media 2.0 concepts that outlined the new directions of next generation internet media research (2006). He has authored and coauthored more than 200 journal and conference papers and holds over 90 U.S. patents in image/video processing, compression and communications, digital television, multimedia, and wireless communication.