J. Vis. Commun. Image R. 29 (2015) 114–124

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

journal homepage: www.elsevier.com/locate/jvci

TapTell: Interactive visual search for mobile task recommendation ☆

Ning Zhang [a,*], Tao Mei [b], Xian-Sheng Hua [c], Ling Guan [a], Shipeng Li [b]

[a] Ryerson Multimedia Research Laboratory, Ryerson University, 350 Victoria St., Toronto, ON M5B2K3, Canada
[b] Microsoft Research Asia, No.5 Dan Ling St., Haidian District, Beijing 100080, China
[c] Microsoft Research, One Microsoft Way, Redmond, WA 98052-6399, United States

ARTICLE INFO

Article history:
Received 3 September 2014
Accepted 9 February 2015
Available online 19 February 2015

Keywords:
Visual intent
Mobile visual search
Interactive visual search
Image retrieval
Mobile recommendation
Natural user interface
Mobile user intention
Visual vocabulary

ABSTRACT

Mobile devices are becoming ubiquitous. People use them as personal concierge to search information and make decisions. Therefore, understanding user intent and subsequently provide meaningful and personalized suggestions is important. While existing efforts have predominantly focused on understanding the intent expressed by a textual or a voice query, this paper presents a new and alternative perspective which understands user intent *visually*, i.e., via visual signal captured by the built-in camera. We call this kind of intent "visual intent" as it can be naturally expressed through a visual form. To accomplish the discovery of visual intent on the phone, we develop *TapTell*, an exemplary real application on Windows Phone seven, by taking advantages of user interaction and rich context to enable interactive visual searches and contextual recommendations. Through the *TapTell* system, a mobile user can take a photo and indicate an object-of-interest within the photo via different drawing patterns. Then, the system performs a search-based recognition using a proposed large-scale context-embedded vocabulary tree. Finally, contextually relevant entities (i.e., local businesses) are recommended to the user for completing mobile tasks (those tasks which are natural to be raised and subsequently executed when the user utilizes mobile devices). We evaluated *TapTell* in a variety of scenarios with millions of images, and compared our results to state-of-the-art algorithms for image retrieval.

1. Introduction

Mobile devices play vital roles in our daily life, from their original function of telephony, to prevalent information-sharing terminals, to hubs that accommodate tens of thousands of applications. While at a moving status, people use their phones as personal concierge in discovering what is around and deciding what to do. Therefore, mobile phone is becoming a recommendation terminal customized for individuals—capable of recommending contextually relevant entities (i.e., local businesses, e.g., a nearby restaurant or hotel). As a result, it is important to understand user intent through its multi-modal nature and the rich context available on the phone.

Existing work has predominantly focused on understanding the intent expressed by text (or the text recognized from a piece of voice). For example, previous researches attempt to estimate user's search intent by detecting meaningful entities from a textual query [1,2]. However, typing on the phone takes time and can be cumbersome for expressing user intent, especially in the situation of searching urgent matters. An alternative is to leverage speech recognition techniques to support voice as an input. For example, popular mobile search engines enable a voice-to-search mode.[1,2] Siri is one of the most successful applications that further structure a piece of speech to a set of entities.[3] However, text as an expression of user intent has two major limitations. First, it relies on a good recognition engine and works well only in a relatively quiet environment. Second, there are many cases where user intent can be naturally and conveniently expressed through the visual form rather than text or speech (such as an unknown object or text, an artwork, a shape or texture, and so on) [3]. As an alternative, we believe that image is a powerful complementary carrier to express user intents on the phone.

Since *intent* is generally defined as "a concept considered as the product of attention directed to an object or knowledge",[4] we can define *mobile visual intent* as follows:

☆ This paper has been recommended for acceptance by Prof. M.T. Sun.

* Corresponding author.
E-mail addresses: ning.zhang@ryerson.ca (N. Zhang), tmei@microsoft.com (T. Mei), xshua@microsoft.com (X.-S. Hua), lguan@ee.ryerson.ca (L. Guan), spli@microsoft.com (S. Li).

[1] http://www.discoverbing.com/mobile/.
[2] http://www.google.com/mobile/.
[3] http://www.apple.com/ios/siri/.
[4] Merriam-Webster Dictionary, 2002.

http://dx.doi.org/10.1016/j.jvcir.2015.02.007
1047-3203/

Definition 1 (*Mobile visual intent*). Mobile visual intent is defined as the intent that can be naturally expressed through any visual information captured by a mobile device and any user interaction with it. This intent represents user's curiosity of certain object and willingness to discover either what it is or what associated tasks could be practiced in a visual form.

The following scenarios demonstrate mobile visual intent and how expressed intent can be predicted and connected to mobile task recommendations. The goal is not only to list related visual results, but also to provide rich context to present useful multimedia information for mobile task suggestion. Mobile tasks are defined as those tasks which can be raised and subsequently executed naturally and conveniently when the user uses mobile devices. For instance, a user passes by an unknown landmark that draws his attention. He can take a picture of it with his smartphone, and use visual intent analysis to acquire related information of this landmark. In another scenario, assuming a restaurant across the street draws an individual's attention, he can take a picture of the restaurant façade with his mobile device, and indicate the object of interest in the captured image. By applying visual intent analysis, the information about this restaurant or its neighborhood points-of-interest are recommended.

Fig. 1 shows two corresponding scenarios. The visual intent model consists of two parts: visual recognition by search and mobile task recommendations. The first problem is to recognize what is captured (e.g., a food image), while the second is to recommend related entities (such as nearby restaurants serving the same food) based on the recognition results for social activities and tasks. This activity recommendation is a difficult task in general since visual recognition in the first step still remains challenging. However, the advanced functionalities, such as natural interaction and a set of available rich context on the mobile device, bring us opportunities to accomplish this task. For example, although one image usually contains multiple objects, a user can indicate an object or some text of interest through a natural drawing pattern, so that visual recognition can be reduced to search a similar single object. Moreover, the contextual information, such as geo-location, can be used for location-based recommendations.

Motivated by the above observations, we present in this paper our recently developed *TapTell* system, which is able to recognize visual intent through interactive search-based visual recognition, and eventually provide contextual recommendation. A natural user interaction is proposed to achieve the *Tap* action, in which three drawing patterns are investigated (i.e., circle, line, and tap). We conclude that a so-called "O" drawing pattern is the most natural interaction for users, which integrates user intelligence to select the targeted object. We propose a novel context-embedded vocabulary tree, which incorporates the context from surrounding pixels of "O" to search similar images from a large-scale image database. Through this kind of user interaction (i.e., "O" drawing pattern), standard visual recognition can be improved. The *Tell* action is accomplished by recommending relevant entities by incorporating the recognition result with the rich contextual information.

Except for proposing the framework of mining visual intent on mobile device, the contributions of this paper can be concluded as follows:

1. We investigate three different kinds of drawing patterns for specifying object-of-interest (and text) by a user study. We conclude that "O" provides the most natural and effective way to interactively formulate user's visual intent and thus reduce the ambiguous nature of user intent in the picture.
2. We propose a context-aware visual search algorithm in which a novel context-embedded vocabulary tree is designed. The algorithm is able to achieve better visual recognition performance by embedding the context information around the "O" area into a standard visual vocabulary tree [4].
3. A viable system developed on Windows phone platform is presented as a demonstration to solidly illustrate proposed method and algorithm.

The rest of the paper is organized as follows. Section 2 reviews related work. Section 3 presents the details of *TapTell*. Experiments and evaluations are given in Section 4, followed by the conclusion in Section 5.

2. Related work

Since the proposed *visual intent* is an original term, we retrospect the evolution of intent in general and walk readers through the formation of the *intent* from text, voice, and visual inputs, with both desktop-based and mobile domain-based search and recognition.

For desktop user intent mining, an early study on web search taxonomy is introduced by Broder [5]. In this work, the most searched items belong to an "informational" category, in which it

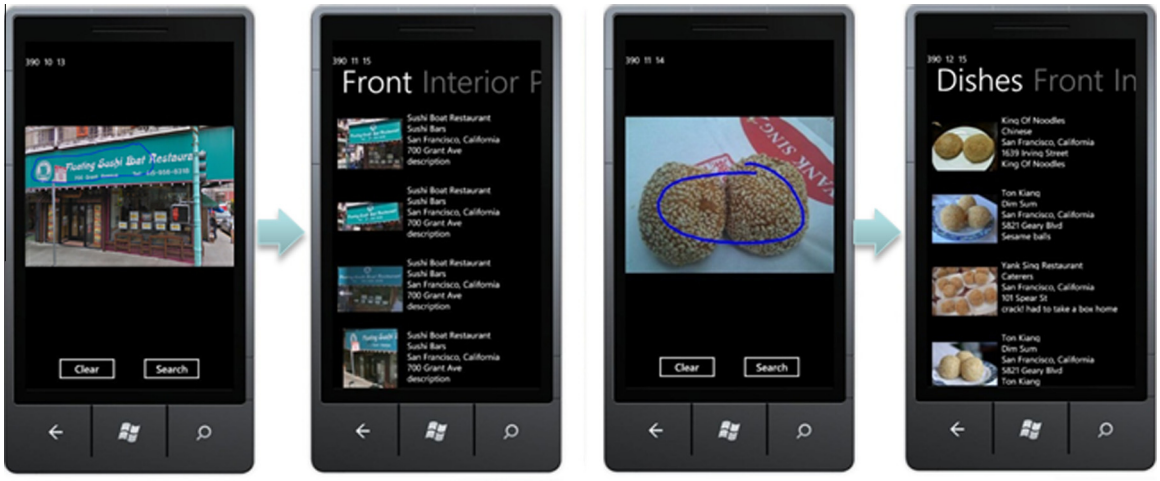

Fig. 1. A snapshot of *TapTell* with two different scenarios. A user can take a photo, specify his/her object-of-interest or text via different drawing patterns (e.g. tap, circle, or line), and then get the search and recommendation results through *TapTell*.

sought related information to answer certain questions in a user's mind. A later work from Rose and Levinson further categorized the informational class to five sub-categories, where the *locate* of a product or service occupies a large percentage [6]. On the other hand, compared to general web searches, intents derived from mobile information have strong correlation with the surroundings and the *status quo*. Church and Smyth conducted a diary study of user behavior of mobile-based text search and summarized a quite different categorization from its general web search counterpart [7]. Besides the informational category at 58.3%, a new geographical category which is highly location dependent takes a share of 31.1% of total search traffic. From a topic perspective, *local services* and *travel* & *commuting* are the most popular ones out of 17 total topics, with 24.2% and 20.2% entries respectively. It can be concluded that the characteristics of the mobility play an important role for intent discovery and understanding on hand-held devices [8].

2.1. Mobile visual search in industry

Due to its potential for practicality, mobile visual search is one of the research areas drawing extensive attention from both industry and academia. Table 1 summarizes an incomplete but the most popular visual search applications from industry. Different from those industrial applications, the proposed system is innovative in terms of an interactive drawing-pattern-based visual search system to help users specify their visual intent, with a consequent recommendation based on the visual search results and contextual information. In this perspective, our system leverages the visual search results to formulate a second query to accomplish task completion on mobile devices, which is significantly different from existing applications.

2.2. Mobile visual search in academia

In academia, quite a few research efforts have been put into developing compact and efficient descriptors, which can be achieved on the mobile end. Chandrasekhar et al. developed a low bit-rate compressed histogram of gradients (CHoG) feature which has a great compressibility [9]. Tsai et al. investigated in an efficient lossy compression to code location information for mobile-based image retrieval. The performance is also comparable with its counterpart in lossless compression [10].

Table 1
A summary of mobile visual search applications in industry.

Application	Features	Techniques	Company
Goggles[a]	Product, barcode, cover, landmark, name card, artwork	Visual search, OCR	Google
Bing Vision[b]	Cover, art, text, barcode	Visual search, OCR	Microsoft
Flow[c]	Cover (CD/DVD/book/video-games), barcode	Visual search	Amazon A9 Laboratory
Kooaba[d]	Logos, cover, landmarks	Visual search	Smart visuals
Lookthatup[e]	Paintings, posters, labels	Visual search	LTU technologies
WordLens[f]	Real-time English/Spanish translation	OCR, AR	QuestVisual

[a] http://www.google.com/mobile/goggles/.
[b] http://www.discoverbing.com/mobile.
[c] http://www.a9.com/-/company/flow.jsp.
[d] http://www.kooaba.com.
[e] http://www.lookthatup.com.
[f] http://www.questvisual.com.

On the other hand, contextual features such as location information have been adopted and integrated successfully into mobile-based visual searches. Schroth et al. utilized GPS information and segmented searching area from a large environment of city to several overlapping subregions to accelerate the search process with a better visual result [11]. Duan and Gao proposed a side discriminative vocabulary coding scheme, extending the location information from conventional GPS to indoor access points, as well as surrounding signs such as the shelf tag of a bookstore, scene context, etc.[12]. Ji et al. proposed a so-called location discriminative vocabulary coding (LDVC) by incorporating region information into visual descriptor coding, and consequently compact the feature transmission for mobile-based landmark search [13].

Additionally, other researchers targeted practical applications and provided promising solutions. Takacs et al. proposed a loxel-based visual feature to describe region-related outdoor object features [14]. Chen and Tsai proposed methods on using image processing techniques to find book spines in order to index book inventories based on bookshelf images [15]. Girod et al. investigated mobile visual search from a holistic point of view with practical analysis under mobile device constraints of memory, computation, devices, power and bandwidth [16]. An extensive analysis using various feature extraction, indexing and matching techniques is conducted using real mobile-based Stanford Product Search system. They demonstrated a low-latency interactive visual search with satisfactory performance.

3. *TapTell* system

3.1. Overview

Fig. 2 shows the framework of *TapTell*. In general, it can be divided into the client-end and cloud-end. We also outline the process into an "intent space", which abstracts the system into three stages: intent expression, intent prediction, and task recommendation. On the client-end, a user's visual intent is specified by the his/her interactive drawing pattern (e.g., "O" drawing pattern) on a captured image. On the cloud-end, with user selected object and the image context around this object, a recognition-by-search mechanism is applied to identify user's visual intent.

For example, the intent can be checking details of a landmark, a restaurant façade, recognizing and translating a menu text, getting the ingredients of a dish, and so on. We have designed a novel context-embedded vocabulary tree to incorporate the "O" context (the surrounding pixels of the "O" region) in a standard visual search process. At last, the specified intents are mapped to the associated metadata by leveraging the sensory context (e.g., GPS-location), which are used to recommend related entities to the user for completing mobile tasks. From the perspective of intent space, intent expression recognizes the object specified by the user-mobile interaction. Intent prediction formulates intent expression and incorporates image context. Finally, a task recommendation is achieved by taking both the predicted intent as well as sensory context.

In essence, the design principles of this proposed TapTell system include:

- **Natural User Interaction (NUI)**. The system is able to provide users an interface for natural interaction. Users can select a region-of-interest(ROI) to highlight the important visual content to express their intention in an intuitive manner, without cumbersome and awkward operation.
- **Robustness**. The system is able to accommodate users' operation error, which is likely to occur during the interaction. Hence, the system should incorporate the not-so-accurate ROI

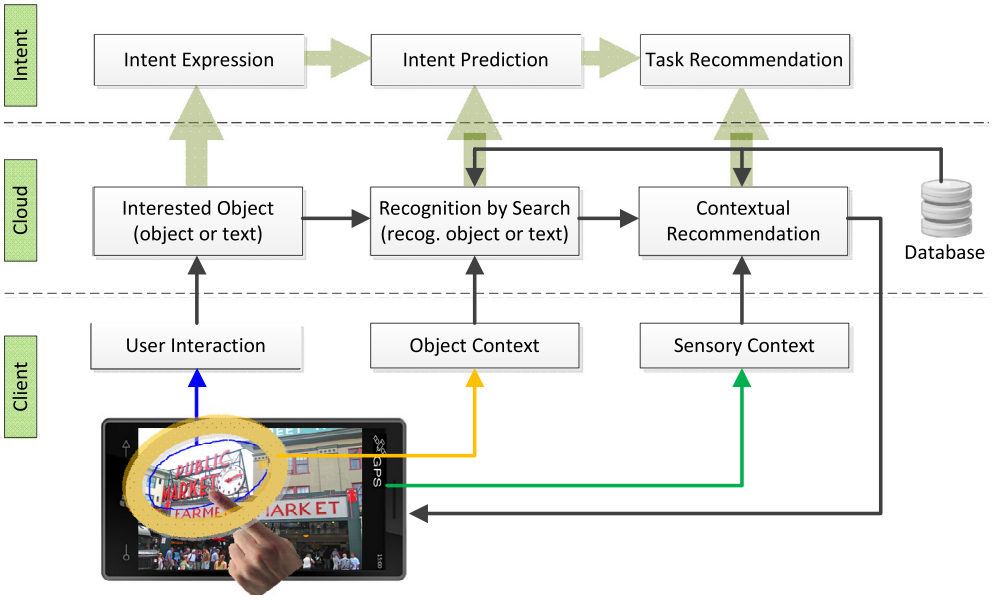

Fig. 2. The framework of *TapTell*, including (1) "O"-based user interaction, (2) image recognition by context-aware visual search, and (3) contextual entity recommendation for mobile tasks which is based on sensory context.

contour selected by the users, and pass their visual content preference to the search engine smoothly.

- **Accuracy**. The system is able to map users' visual preference and reflect it in the visual search process, so that a satisfactory visual retrieval result is obtained.
- **Efficiency**. The system is able to provide adequate visual search result and consequent task recommendation for users in real-time without much latency.

Section 3.2 introduces how the "O" drawing pattern works and how to convert from a user's particular drawing pattern to a visual intent that is computable for the system. Section 3.3 presents the visual recognition by search, with the help of both image context as well as sensory GPS information. Section 3.4 describes the recommendation method, using text metadata associated with visual features to achieve a better re-ranking.

3.2. Drawing patterns for specifying visual intent

It has been studied and suggested that visual interface will improve mobile search experiences [17]. In this section, we have performed a user study to identify the most natural and efficient drawing pattern for specifying the visual intent on mobile devices. By taking advantage of natural interaction on smart-phones, we defined three drawing patterns for specifying visual intents on captured photos as follows:

- **Tap.** A user can "tap" on the pre-determined image segments, in which a captured image is automatically segmented on-the-fly developed by Deng et al. [18]. Then, the tapped segments indicated by user's drawing pattern will be connected as the ROI. The ROI will be further used as the visual query, as shown in Fig. 3(a).
- **Line.** A user can draw straight "lines" to form a rectangular bounding box. The region in the box will be used as the visual query, as shown in Fig. 3(b).
- **O (*circle*).** A user can naturally outline an object of irregular shape. The "O" drawing pattern can be also called the *circle* or *lasso*. Note that an "O" is not limited to a circle, but any arbitrary shape, as shown in Fig. 3(c).

We performed a user study following the principles of focus group in the field of human–computer interaction [19]. In this study, 20 participants (11 female and 9 male, 20–34 years old) were invited. The background of the participants range from programmer, buyer, editor, accountant, secretarial, human resources; as well as undergraduates and graduates in computer science, animation, design, geology and social sciences. After being introduced to the basic functions of *TapTell* and getting familiar with the system, they were asked to perform 3–5 subtasks using different drawing patterns. A subtask is defined as following: a participant is asked to pick one landmark or cuisine dish and takes a picture of it. The subtask ends either the participant is satisfied with the results, or disappointed by the system performance and decided to give up. After the usability evaluation, participants were requested to fill a questionnaire, designed with seven-point Likert scale. They were asked to specify their level of agreement at various aspects of user experiences with three aforementioned drawing patterns, including: speed, user-friendliness, ease of operation, flexibility, novelty, clear search intent, relevance of search result, and overall rate. From this study, "O" drawing pattern obtains the highest overall rating score of 6.1 for the 1–7 rate scale, while "line" and "tapping" drawing patterns receive 5.6 and 3.5, respectively. Their comments on "tapping" and "line" are: (1) tapping is sometimes too sensitive and image segmentation is not always satisfying, and (2) the "line" is not convenient for selecting an arbitrary object.

Therefore, we adopt "O" drawing pattern to mine visual intent in *TapTell*. The "O" drawing pattern utilizes touch screen of the smart-phone. Users do not need any training and can naturally engage with the mobile interface immediately. After the trace (the blue thin line in Fig. 2) has been drawn on the image, sampling points along the trace-line are collected as $\{\mathbf{D}|(x_j,y_j)\in\mathbf{D}\}_{j=1}^{N}$, which contain N pixel-wise positions (x_j,y_j). We applied principle component analysis (PCA) to find two principle components (which form the elliptical ring depicted by thick orange line in Fig. 2). The purpose of this part is to formulate a boundary of the selected region from an arbitrary "O" drawing pattern trace. We also calculated mean μ and covariances Σ based on $\mathbf{D}$ and non-correlated assumption along the two principle components:

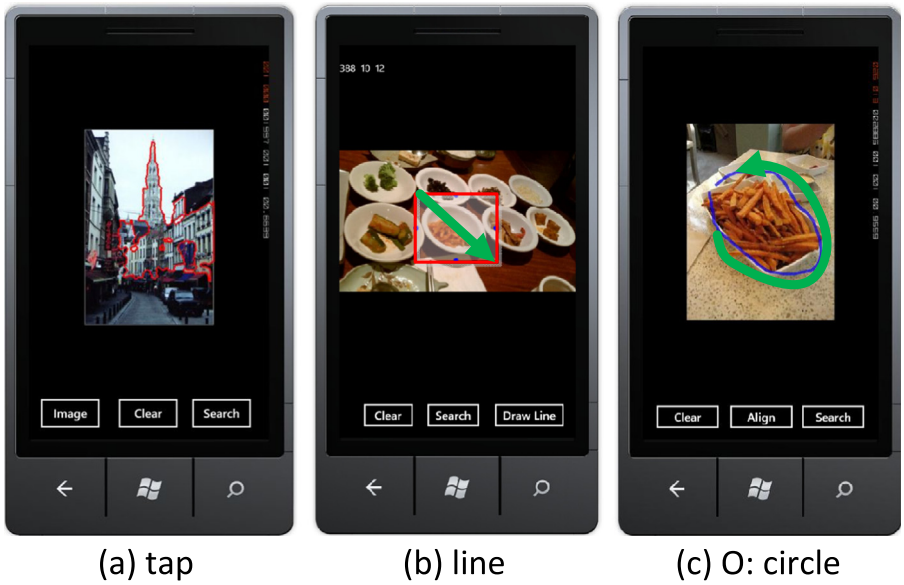

Fig. 3. Different drawing patterns for specifying user intent in *TapTell*: (a) "tap"—selection of image segments, (b) "line"—rectangular box, and (c) "O"—circle or lasso.

$$\mu = [\mu_x, \mu_y] \qquad \Sigma = \begin{vmatrix} \sigma_x^2 & 0 \\ 0 & \sigma_y^2 \end{vmatrix} \tag{1}$$

Fig. 4 shows the computation of principle components from the "O" query. Once the principle components are identified, the recognition by search method is used to identify the object-of-interest indicated by the user.

3.3. Visual intent recognition on the cloud

The proposed visual intent recognition method is based on a recognition scheme using the vocabulary tree proposed by Nister et al. [4]. This method provides a fast and scalable search mechanism and is suitable for large-scale and expansible databases because of its hierarchical tree-structured indexing. When we adapt this method in the mobile domain, the "O" drawing pattern fits naturally to provide a focused object selection for better recognition. Different from using the entire image as the visual query in [4], we have the user-indicated ROI from the "O" drawing pattern (called "O-query") in the *TapTell* system. We design a novel context-aware visual search approach in which a context-embedded

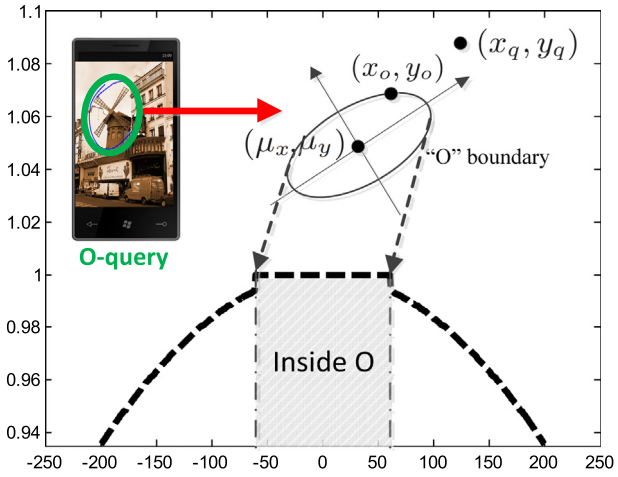

Fig. 4. The illustration of user indicated "O" query, and the computation of principle components of the query. (μ_x, μ_y) is the center of "O" query, (x_o, y_o) is a pixel on the "O" boundary, and (x_q, y_q) is a query pixel.

vocabulary tree (CVT) is built to take the surrounding pixels around the O-query into consideration. On the other hand, image context and spatial verification is not a novel idea. Representative works such as Philbin et al. [20] and Chum et al. [21], have discussed how to use geometric transformation as hypothesis to verify ranked images with reference to the query image ROIs. Comparing with above mentioned approaches, the proposed CVT method incorporates spatial analysis within the visual recognition using visual intents, instead of treating the spacial verification as a re-ranking process after the content-based image retrieval. As a summary, the CVT is able to reduce the following ambiguities:

- Issuing O-query only in image-based search engines may lead to too many similar results. The surrounding pixels provide a useful context to differentiate those results.
- The O-query may not have (near) duplicates or exist in the image database. Issuing only O-query may not lead to any search results. The surrounding pixels then can help in providing a context to search for the images with similar backgrounds.
- Hierarchically built K-means clustering for codebook generation makes the retrieval process efficient, wherein each queried local feature only goes through one particular branch at the highest level and its sub-branches instead of going through the entire codebook.

The proposed CVT-based visual search approach encodes different weights of term frequencies inside and outside the O-query. We will describe the proposed visual search approach in Section 3.3.1. We also propose a location-based context filter in Section 3.3.2 for re-ranking visual search results based on user's current location (derived from the GPS-enabled images taken by the phone camera).

For off-line image indexing, at the feature extraction stage, we adopted the same approach originated by Lowe [22]. We apply a Difference of Gaussian (DoG) based keypoints detection, and then extract a sparse-based 128-dimensional local scale-invariant feature transform (SIFT) descriptors. At the indexing stage, since our target database is large-scale, we utilize the hierarchical K-means to quantize the local descriptors and build the CVT. Then, we index the large-scale images using the built CVT and the inverted file mechanism, which is to be introduced in the following.

3.3.1. Context-aware visual search

In on-line image searches, given a query image, we can quantize the descriptor vectors of the image in a similar way to the indexing procedure, and accumulate scores for the images in the database with a so called *term frequency–inverse document frequency* (tf–idf) scheme [4]. This tf–idf method is an effective entropy weighting for indexing a scalable database. In the vocabulary tree, each leaf node corresponds to a visualword i, associated with an inverted file (with the list of images containing this visualword i). Note that we only need to consider images d in the database with the same visualwords as the query image q. This significantly reduces the amount of images to be compared with respect to q. The similarity between an image d and the query q is given by

$$s(q,d) = \|\mathbf{q}-\mathbf{d}\|_2^2 = \left(\sum_{i|d_i=0}|q_i|^2 + \sum_{i|q_i=0}|d_i|^2 + \sum_{i|q_i\neq 0, d_i\neq 0}|q_i-d_i|^2\right) \quad (2)$$

where $\mathbf{q}$ and $\mathbf{d}$ denote the tf–idf feature vectors of the query q and image d in the database, which are consisted of individual elements q_i and d_i (i denotes the i-th visualword in the vocabulary tree), respectively. q_i and d_i are the tf–idf value for the i-th visualword in the query and the image, respectively. Mathematical interpretations are given by

$$q_i = tf_{i_q}\cdot idf_i \quad (3)$$

$$d_i = tf_{i_d}\cdot idf_i \quad (4)$$

In the above equation, the *inverted document frequency* idf_i is formulated as $ln(N/N_i)$, where N is the total number of images in the database, and N_i is number of images with the visualword i (i.e., the images whose descriptors are classified into the leaf node i).

The main difference of the proposed CVT approach from the literature work [4], resides in *term frequency*. The merit of this approach is to incorporate the user interactive ROI result, and improves the *term frequency*. Therefore, the emphasis of the natural user interaction and ROI selection is smoothly delivered to the visual search process. Consequently, a better visual retrieval result is achieved. The *term frequency* representations tf_{i_q} and tf_{i_d} are computed as the accumulated counts of the visualword i in the query q and the database image d, respectively. One simple means for the *term frequency* computation is to use the O-query as the initial query without considering the pixels surrounding the "O". This process is equivalent to using "binary" weights of the *term frequency* tf_{i_q}: the weight is 1 inside "O", and 0 outside "O". A more descriptive and accurate computation is to incorporate the context information (i.e., the surrounding pixels around the O-query) in the vocabulary tree. We design a new representation of the *term frequency* $tf^o_{i_q}$ for the O-query. A "soft" weighting scheme is proposed to modulate the *term frequency* by incorporating the image context outside the O-query, which was neglected in the simple binary scheme. When quantizing descriptors in the proposed CVT, the $tf^o_{i_q}$ of the O-query for a particular query visualword i_q is formulated as:

$$tf^o_{i_q} = \begin{cases} tf_{i_q} & \text{if } i_q \in \mathrm{O} \\ tf_{i_q}\cdot \min\left\{1, \frac{\Re(x_q,y_q)}{\Re(x_o,y_o)}\right\} & \text{if } i_q \notin \mathrm{O} \end{cases} \quad (5)$$

where $\Re(x_o,y_o)$ and $\Re(x_q,y_q)$ denote the Gaussian distances of the pixel (x_o,y_o) and (x_q,y_q) with respect to the center of O-query (μ_x,μ_y). Fig. 4 shows the definition of these pixels in the query image q. The Gaussian distance $\Re(x,y)$ for an arbitrary pixel (x,y) is given by

$$\Re(x,y) = A\cdot \exp\left\{-\frac{1}{2}\left[\frac{(x-\mu_x)^2}{\alpha\sigma_x^2} + \frac{(y-\mu_y)^2}{\beta\sigma_y^2}\right]\right\} \quad (6)$$

The "soft" weighting scheme shown in Eq. (5), is a piece-wise, bivariate-based multivariate distribution outside the O-query, and a constant 1 inside the O-query. The position (x_o,y_o) is the boundary of the O-query contour where the weight 1 ends. In the case that a visualword i_q is outside the O-query, the modulating term is $\min\left\{1, \frac{\Re(x_q,y_q)}{\Re(x_o,y_o)}\right\}$, such that the soft weighting is guaranteed to be less than 1. The term $\frac{\Re(x_q,y_q)}{\Re(x_o,y_o)}$ is the ratio of which the query point (x_q,y_q) should be weighted with respect to its closest boundary position (x_o,y_o). Mean values μ_x and μ_y are calculated from "O" drawing pattern sample data, while α and β are tunable parameters to control the standard deviation for the bivariate normal distribution. Fig. 4 also illustrates this "soft" weighting schemes in the CVT when a projection view along one principle axis is sliced and presented. Parameter A is the amplitude value controlling the highest possible weighting scale. Parameters α and β reflect the importance of the horizontal and vertical axis (or directions) when employing the PCA technique. Empirically, we set α with higher value than β to indicate that the horizontal axis is usually more important than the vertical one. This is because most pictures are taken by the phone camera horizontally.

3.3.2. Location-based filtering

Context information collected by mobile sensors plays an important role to help to identify users' visual intents. Similar with the inverted file index method, each piece of image context information is indexed with the image itself during the off-line database construction. In our system. a GPS filter-based process is used to remove the non-correlated images after the initial visual search. This is because GPS as an important context filter can be used to efficiently explore users' true intents by precisely knowing their locations. This process is formulated as:

$$S_L(q,d) = s(q,d)\cdot\phi(\mathbf{q},\ \mathbf{d})$$
$$where\quad \phi(\mathbf{q},\ \mathbf{d}) = \begin{cases} 1, & \text{if } dist_{quadkey}(\mathbf{q},\mathbf{d}) \in \mathrm{Q} \\ 0, & \text{if } dist_{quadkey}(\mathbf{q},\mathbf{d}) \notin \mathrm{Q} \end{cases} \quad (7)$$

The visual similarity term $s(q,d)$ is modulated by a location-based filter $\phi(\mathbf{q},\ \mathbf{d})$. This filter is based on the GPS effective region Q, which describes the geographical distance between the query and the database images. We defined $dist_{quadkey}(\mathbf{q},\mathbf{d})$ as the quadkey distance between the query $\mathbf{q}$ and the database image $\mathbf{d}$. The definition of the quadkey distance, $dist_{quadkey}(\mathbf{q},\mathbf{d})$, is adopted from the Bing Maps Tile System.[5] It converts the GPS coordinates to a hashing-based representation for fast search and retrieval. We encode the GPS signal to a quadruple tiles code, which consists 23 digits number with the ground resolution of possible 0.02 m accuracy. The formulation of this distance is computed by the Quadkeys representation. GPS context from mobile sensor is collected first. The standard WGS-84 is encoded to the quadkey representation. To determine the physical distance between two captured images, a hamming distance is computed between those two quadkeys. Then, the ground distance is calculated by summing up the associated granular scales from Bing Maps lookup table. In the illustration of how to calculate $dist_{quadkey}(\mathbf{q},\mathbf{d})$ shown in Fig. 6, pictures of the same landmark (the Brussels town hall) with both the front and the back façades are taken at different location $\mathbf{q}$ and $\mathbf{d}$. These two photos have different WGS-84 information, which have 10 out of 15 quadkey digits identical after Bing Maps projection. In other words, the hamming distance between these two codes is 5, which is calculated

[5] http://www.msdn.microsoft.com/en-us/library/bb259689.aspx.

using tables to approximate a ground distance of about 305 m. Hence, the effective ground distance (305 m) is the quadkey distance.

3.4. Recommendation

Recently, Jain and Sinha proposed to re-examine the fundamental issue between content and context and why researchers should utilize both of them to bridge the semantic gap [23]. Guy et al. suggest that while machine learning and human computer interactions play key roles in recommendations, personalization and context-awareness are also crucial in establishing an efficient recommendation system [24]. We agree with their arguments that it is necessary to connect data and users. We also believe that smart-phones provide perfect platforms for such data-users connection, from human computer interaction, to visual search, and finally, to the recommendation. In our work, after the visual intent expression and identification, we utilize rich metadata as a better feature to search. We also use powerful context to re-rank metadata-based search result for the final task completion [8]. To be specific, we adopt the metadata associated with the top image search result as our textual query. Then, we obtain the mobile activity recommendations based on the text retrieval results. The Okapi BM25 ranking function is used to compute a ranking score based on text similarity [25]. We extract the keywords $Q_t = \{q_{t_1}, q_{t_2}, \ldots, q_{t_n}\}$ by projecting the text query to a quantized text dictionary. Then, we compute the relevance score of query Q_t and database image descriptions D_t. Detailed score computation techniques can be referred to in reference [25]. In the last step, we re-rank the search results based on the GPS distance of the user's current location. Fig. 5 demonstrates a sample result of the recommendation list and location-based re-ranking.

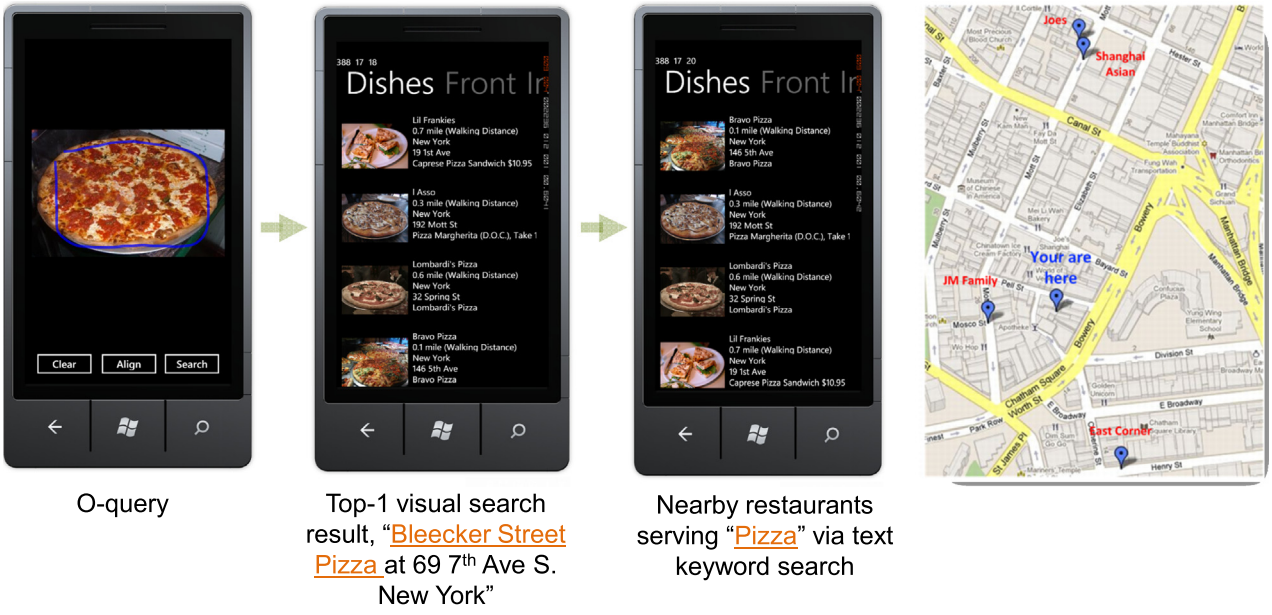

Fig. 5. A visual result of recommendation list, which is visualized in a map to help users to picture the distances between the query and the results.

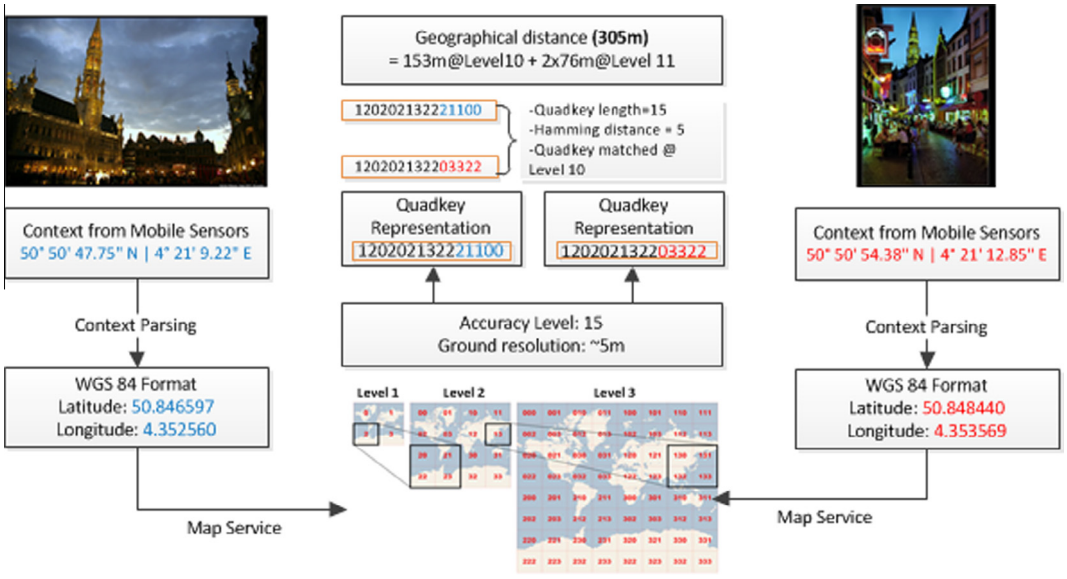

Fig. 6. Quadkeys quantization and hashing from GPS, and images ground distance estimation using Microsoft Bing Map service.

4. Experiments

4.1. Data and settings

The client-end application is developed on a Windows Phone 7 HD7 model with 1 GHz processor, 512 MB ROM, GPS sensor and 5 megapixel color camera. In the cloud, a total of one million visual-words is built from 100 million sampled local descriptors (SIFT in this experiment). A hierarchal tree structure consisting of six levels of branches is used, where each superior branch has 10 sub-branches or nodes. In constructing the vocabulary tree, each visualword takes up to 168 bytes storage, where 128 bytes are for the clustering vector (same size as SIFT), and 4 bytes for ten subordinate children nodes connection. In total, 170 megabytes of storage is used for the vocabulary tree in cache.

The dataset consists of two parts. One is from Flickr, which includes a total of two million images, with 41,614 landmarks equipped with reliable GPS contextual information. With a further manual labeling effort, 5,981 images were identified as the groundtruth such that the landmark object façade or the outside appearance can be traced from the image. The second part of the database is a crawled commercial local services data, mainly focusing on the restaurant domain. In this part, a total of 332,922 images associated with 16,819 restaurant entities from 12 US cities were crawled with associated metadata.

4.2. Evaluation metrics

We use mean average precision (MAP) for the evaluation, where MAP is the mean value of average precisions (APs). The average precision (AP) formula is presented as

$$AP@n = \frac{1}{min(n,P)} \sum_{k=1}^{min(n,S)} \frac{P_k}{k} \times I_k \tag{8}$$

The number of top ranks is represented as n. The size of the dataset is denoted as S, and P is the total number of positive samples. At index k, P_k is the number of positive results in the top n returns, and I_k is described as the result of the k_{th} position.

4.3. Objective evaluations

4.3.1. Evaluation of context-embedded visual recognition

We investigated image contextual information and its effectiveness in recognition by search technique, using the soft weighting scheme. For the bivariate-based function $\Re(x,y)$, we fixed the amplitude A to 1 and tuned two parameters α and β to modulate the standard deviation. We conducted two sets of experimentation without and with GPS context shown in Fig. 7. In general, using the soft weighting scheme improves search performance, compared to the binary weighting method. Specifically, without using GPS information, $\alpha = 50$ and $\beta = 10$ provide the best performance for the MAP measurement. The results of this parameter choice outperforms the binary weight method by 12%.

Similarly, after incorporating the GPS context, the soft weighting method again outperformed the binary one, but in a much higher precision range. This does not surprise us since geolocation is an important feature for differentiating objects and their recognition, and eventually associated visual intent. Different from its counterpart in the non-GPS scenario, parameter $\alpha = 5$ and $\beta = 1$ outperforms other parameter choices, as well as the baseline binary weighting scheme. The margin difference from the soft weighting and the binary case has dropped to 2% for MAP. This result demonstrates the importance of the GPS context.

It can be observed that parameter α is higher than parameter β for the best performance in Fig. 7. The reason is due to the fact that most images are taken horizontally. Therefore, information is appreciated more and weighted higher by α horizontally than its counterpart β vertically. Similar patterns can also be observed in the following evaluations.

The significance of this image contextual information with soft weighting scheme allows robust user behavior and is seamlessly glued with the "O" drawing pattern, which is spontaneous and natural. The shortcoming of the"O" is that it inevitably suffers from lack of accuracy due to device limitations in outlining the boundary, compared to other drawing patterns, such as segmentation or line-based rectangular shape. However, soft weighting alleviates this deficiency of correctness in object selection and provides a robust method to accommodate behavioral errors when drawing the outlines of the ROI.

4.3.2. Evaluation and comparison with contextual image retrieval model (CIRM)

We also implemented a state-of-the-art contextual image retrieval model (CIRM) and compared its performance to our proposed context-embedded visual recognition [26]. The CIRM has demonstrated a promising result in desktop-based CBIR by applying a rectangular bounding box in highlighting the emphasized region, which can be achieved using mouse control at a desktop platform. The weighting scheme in CIRM model is to use two logistic functions joined at the directional (either X or Y) center of the bounding box. Then, the term frequency tf_q is formulated as:

$$tf_q \propto min\left(\frac{1}{1+exp(\delta_X(x_l - x_i))}, \frac{1}{1+exp(\delta_X(x_i - x_r))}\right) * min\left(\frac{1}{1+exp(\delta_Y(y_t - y_i))}, \frac{1}{1+exp(\delta_Y(y_i - y_b))}\right) \tag{9}$$

where x_l, x_i, x_r represent x pixel values of the left boundary, detected feature point, and the right boundary along the x-axis direction, respectively. Similarly, y_t, y_i, y_b are the y pixel values of the top

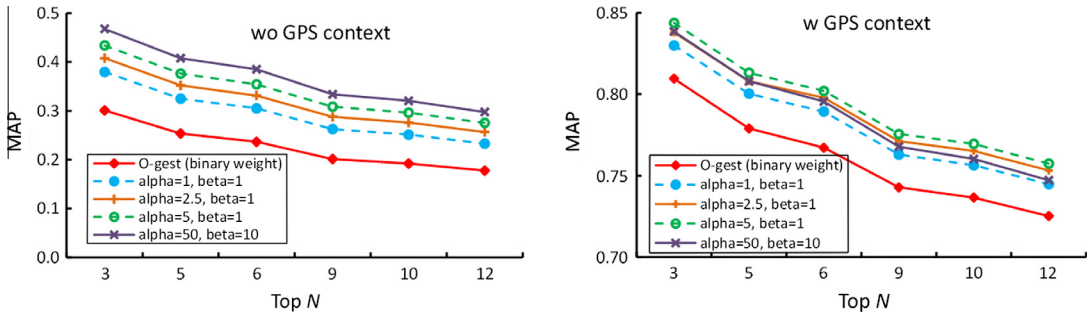

Fig. 7. MAP measurements of proposed context-embedded visual recognition using various parameter α and β, compared with the binary weight scheme, without and with GPS information.

boundary, detected feature point, and the bottom boundary along the y-axis, respectively. The geometric relations $x_l < x_i < x_r$ and $y_t < y_i < y_b$ hold for this bounding box, such that the tf_q should be approaching the value 0, the further x_i from the bounding box; while ideally close to value 1 when the feature point is inside the bounding box. δ_X and δ_Y are two tunable parameters for finding the best performance of the bounding box. Detailed explanation of the algorithm can be found in reference [26].

Fig. 8 shows MAP measurements, by comparing the proposed Gaussian-based contextual method with the CIRM model, as well as the CBIR method using the original image. Without using the GPS context, it appears that the proposed method with parameters $\alpha = 40$ and $\beta = 10$ outperformed both CIRM in its best result with parameter $dX = 0.0001$ and $dy = 0.0001$, and the CBIR result of the original image without using contextual model. The reason of a small discrepancy between the original image query and the proposed method is due to an adopted large-scale dataset. As a result, the retrieval performance is not much different in this case. In the circumstance of using the GPS context re-ranking. Again, the proposed method outperformed the CIRM method and the CBIR algorithms. However, the best performance of the CIRM model at $dX = 0.0001$ and $dY = 0.0001$ is close to the performance of our proposed contextual model at $\alpha = 5$ and $\beta = 1$. This result can be explained, such that, by adopting the GPS filtering, the margin of various methods is reduced.

4.3.3. Evaluation of mobile recommendations

For the recommendations, our method is to use the visual photo taken by users as the starting point, and to provide recommendation lists based on text searches associated with the recognized object. First, we identify the object and match it to the database. Then, we use the matched metadata as a text query to do a text-based search. The final result is then re-ranked by the relevant GPS distance from the query's image location to the ranked list image locations.

The evaluation was conducted exclusively on a vertical domain of food cuisines. We randomly picked 306 photos and manually labeled and categorized them into 30 featured themes of food dishes, such as beef, soup, burger, etc. We built a 300 word text dictionary by extracting the most frequently used words in the image description.

In order to produce a real restaurant scenario, we printed out dishes in a menu style with both texts and images. We took pictures of the dishes as the visual query and attempted to find the duplicated/near-duplicated images from the dataset. We assumed that the best match of the visual recognition result would be user intent. Such intent was carried by the associated metadata, which were quantized using the prepared 300-word dictionary. The quantized words were searched with a ranked list based on the text similarity. The final step was to re-rank the result list using GPS distance.

Table 2 presents the MAP result with the initial visual query and newly formatted text description query after visual recognition. The Table demonstrates that the performance of the text description-based search is much better than the visual-based search. The "Visual-based" MAP value only represents one-step visual search result. The "Description-based" MAP value is a two-step process. The first search uses the visual query. After obtaining the top result (@0) and its associated text description, a text-based search is implemented using the text descriptors obtained. This also explains why there is no value @0 for the "Description-based"

Table 2
MAP evaluation of the visual-based and description-based performance.

MAP	@0	@1	@2	@3	@4
Visual-based	96.08	53.06	37.61	29.60	24.59
Description-based	n/a	75.65	72.66	70.78	65.93

Table 3
A summary of the captured average time cost (TC) in seconds, based on surveyed subjects.

	Number of trials	Total TC	Interaction TC	Search TC	EPH TC	FtS+PI-TC	
						FtS	PI
TT-Task1	1.28	45.07	31.04	14.03	4.17	11.06	10.86
GG-Task1	1.43	39.52	26.71	12.81	4.13	15.78	
TT-Task2	1.10	36.60	24.44	12.15	4.51	8.61	9.76
GG-Task2	1.20	31.32	20.68	10.64	4.00	13.10	

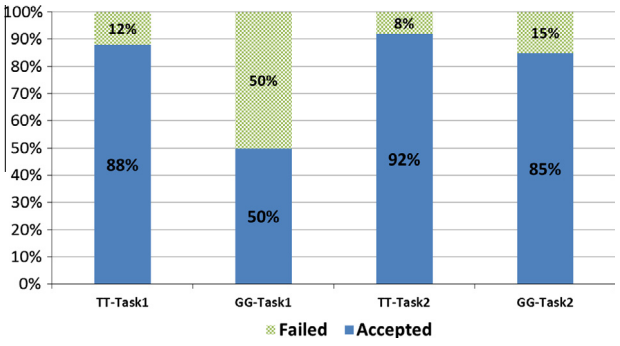

Fig. 9. Accepted rate of participant trails in Task 1 and Task 2.

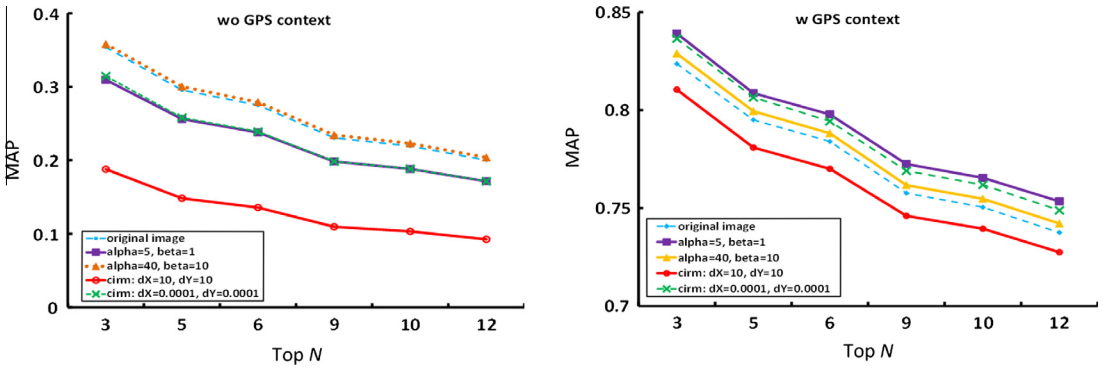

Fig. 8. MAP measurements of proposed context-embedded visual recognition by various parameter α and β, compared with the conventional CBIR (original) algorithm, as well as CIRM algorithm with parameter dX and dY, without and with GPS information.

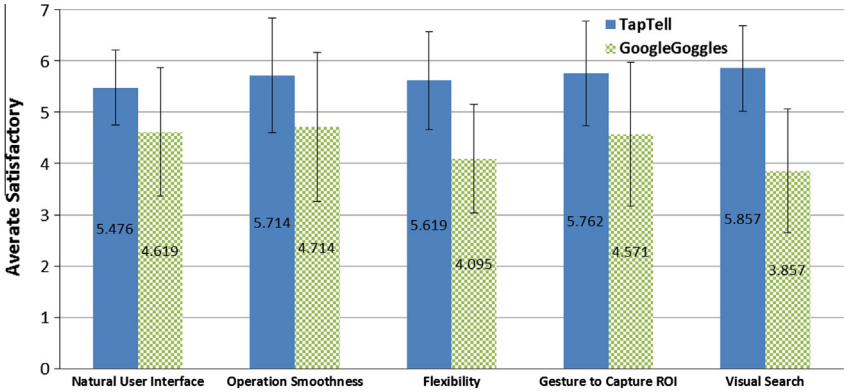

Fig. 10. Average satisfactory level of the proposed TapTell and Google Goggles systems.

MAP. Starting from @1 and so on, both "Visual-based" and "Description-based" MAPs are available. This result is reasonable in the sense that text is a better description than visual content once the ROI is identified and linked with precise textual metadata. However, the merit of the visual input is its role in filling the niche when an individual does not have the language tools to express him/herself articulately. We demonstrate that during the initial visual search (@0), the visual-search result is at a high precision rate of 96.08%. Such accuracy provides a solid foundation to utilize associated metadata as a description-based query during the second stage search. In summary, once the visual query is mined accurately, the role of the search query is then shifted from visual content to text metadata for a better result (see Table 3).

4.4. User study

We conducted a subjective evaluation on user experience with the *TapTell* system. A total of 20 people participated the survey, 10 male and 10 female. The background of the participants ranges from programmer, editor, secretarial, human resources, to undergrad and graduate students in computer science and social science. All 20 participants have various experiences in using smart-phones with touch screen. We simulated two scenarios as tasks. **Task 1**: a food service scenario, participants are presented a menu including various cuisine dishes. The assumption is that a user walked by a restaurants and took pictures of illustrated dishes from the menu. **Task 2**: an informational seeking tourism scenario, poster-size fine prints of building façade/landmarks were used to simulate the situations that a tourist passed by a building which caught her attention. Each participant was asked to perform 20 rounds of "snap-interact-search" process. A task was ended either with a satisfactory result or terminated with a given-up trial after 3 times attempts. We compared our TapTell (TT) application with the latest Google Goggles (GG) application (ver. 1.9). Fig. 9 depicts accepted/failed rate of TT and GG. A total of 315 trials are accepted out of 400 rounds, performed by 20 participants, who were on both TT and GG applications, executes two tasks on each application, and accomplished five trials per task.

For the accepted search trials, we further broke down the total time cost (Total-TC) to two stages of interaction-TC, search-TC. The interaction-TC is the time spent on the user interaction. It can be further decomposed to entering-the-photo-mode-TC (EPH-TC), Frame-the-shot-TC (FtS-TC), and perform-interaction-TC(PI-TS). Since the FtS and TF are performed together fo the GG case, we then combined these two together. The average number of trials to obtain an accepted result is also listed. By averaging both tasks, the Total-TC for the TapTell is 5.42 s more than Google Goggles, in which about 4.05 s is in interaction-TC, and 1.37 s is in Search-TC. It is expected that GG will perform better than TT because the former is a released commercial product while the latter is a proof-of-concept application without much code optimization. The average TC for TT is about 15.4% more than the GG, while about 17.2% in the interaction, and 11.9% in search.

Besides the time measurement, a survey was also conducted by each user after their participation. They were asked about the usefulness of and satisfaction with the proposed TT and the benchmark GG systems, in terms of user-interface friendly and natural level, operation smoothness, flexibility in forming the visual query, drawing pattern performance (TT O-drawing pattern vs. GG cropping) to select out the ROIs during the interaction, and the visual search result. The survey is designed based on a seven-point Likert scale, ranked from 1 to 7, where 1 is the least and 7 is the most. Fig. 10 summarizes an average survey result with standard deviation bar errors. It can be observed that the user experience for TapTell is better than Google Goggles in all five categories, the average score of the TapTell and the Google Goggles are 5.69 and 4.37 of the Linkert scale, respectively.

4.5. Video demonstration

We also have uploaded a video demo to showcase the *TapTell* system. The video speed is set to x1.7 more than the original footage to make this video demo more compact and agreeable to watch.[6]

5. Conclusion

We have proposed a contextual-based mobile visual intent model in this paper. As an implementation of the application of this kind, *TapTell* achieves mobile visual recognition and recommendation. Meaningful tasks are recommended to mobile users by leveraging multimedia content (as input) and rich contextual information. We have investigated on different drawing patterns from tapping the segments, to drawing lines of rectangle, and to "O"-circle via the touch screen, and demonstrated that the "O" behavior is the most natural and agreeable user-mobile interaction. Along with the BoW model, we proposed a new context-embedded vocabulary tree for soft weighting using both "O" object and its surrounding image context to achieve better mobile visual search performance. We verified that image context outside the "O" region plays a constructive role in improving the recognition. It is also demonstrated that the proposed method outperforms both conventional CBIR using original image query and the state-of-the-art CIRM algorithm. User study is conducted to compare

[6] http://youtu.be/6dA45vrdfTk.

the proposed TapTell with the state-of-the-art Google Goggles system. TapTell has a comparable time duration for the total, user interaction and search time costs, while achieves better scores based on the post-experiment user survey.

TapTell opens a new perspective to connect multimedia data with user intent through the following innovations: (1) a natural user interface that enables users to interactively specify their intents more conveniently, (2) a better visual search approach that leverages pixel context and user interaction, and (3) a contextual recommendation system that predicts relevant mobile tasks and facilitates activity suggestion. In our future work, we will further investigate other recommendation schemes, including both social and personal status of the mobile user. We will also experiment to use additional testing approach such as validation mechanism to consolidate learned system parameters α and β. Furthermore, we will try to utilize other local features such as dense-based sampling SIFT points in order to achieve a better object recognition performance.

References

[1] J. Polifroni, I. Kiss, M. Adler, Bootstrapping named entity extraction for the creation of mobile services, in: Proceedings of LREC, 2010, pp. 1515–1520.
[2] X. Yin, S. Shah, Building taxonomy of web search intents for name entity queries, in: Proceedings of WWW, 2010, pp. 1001–1010.
[3] J. Smith, Clicking on things, IEEE MultiMedia 17 (2010) 2–3.
[4] D. Nister, H. Stewenius, Scalable recognition with a vocabulary tree, in: Proc. IEEE CVPR, 2006, pp. 2161–2168.
[5] A. Broder, A taxonomy of web search, in: Proc. of ACM SIGIR, 2002, pp. 3–10.
[6] D. Rose, D. Levinson, Understanding user goals in web search, in: Proc. of the ACM WWW, 2004, pp. 13–19.
[7] K. Church, B. Smyth, Understanding the intent behind mobile information needs, in: Proc. of ACM International Conference on Intelligent User Interfaces, 2009, pp. 247–256.
[8] J. Zhuang, T. Mei, S.C.H. Choi, Y.-Q. Xu, S. Li, When recommendation meets mobile: Contextual and personalized recommendation on the go, in: Proc. of ACM International Conference on Ubiquitous Computing, Beijing, China, 2011, pp. 153–162.
[9] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, B. Girod, CHoG: compressed histogram of gradients a low bit-rate feature descriptor, in: Proc. of IEEE CVPR, 2009, pp. 2504–2511.
[10] S. Tsai, D. Chen, G. Takacs, V. Chandrasekhar, J. Singh, B. Girod, Location coding for mobile image retrieval, in: Proceedings of the 5th International ICST Mobile Multimedia Communications Conference, 2009, pp. 1–7.
[11] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, E. Steinbach, Mobile visual location recognition, Signal Process. Mag. 28 (2011) 77–89.
[12] L.-Y. Duan, W. Gao, Side discriminative mobile visual search, in: 2nd Workshop on Mobile Visual Search, 2011.
[13] R. Ji, L. Duan, J. Chen, H. Yao, J. Yuan, Y. Rui, W. Gao, Location discriminative vocabulary coding for mobile landmark search, Int. J. Comput. Vis. 96 (2012) 290–314.
[14] G. Takacs, V. Chandrasekhar, N. Gelfand, Y. Xiong, W. Chen, T. Bismpigiannis, R. Grzeszczuk, K. Pulli, B. Girod, Outdoors augmented reality on mobile phone using loxel-based visual feature organization, in: Proc. of ACM MIR, 2008, pp. 427–434.
[15] D. Chen, S. Tsai, B. Girod, C. Hsu, K. Kim, J. Singh, Building book inventories using smartphones, in: Proc. ACM Multimedia, 2010, pp. 651–654.
[16] B. Girod, V. Chandrasekhar, D. Chen, N. Cheung, R. Grzeszczuk, Y. Reznik, G. Takacs, S. Tsai, R. Vedantham, Mobile visual search, Signal Process. Mag. 28 (2011) 61–76.
[17] K. Church, B. Smyth, N. Oliver, Visual interfaces for improved mobile search, in: Workshop on Visual Interfaces to the Social and the Semantic Web, 2009.
[18] Y. Deng, B. Manjunath, Unsupervised segmentation of color-texture regions in images and video, IEEE Trans. Pattern Anal. Machine Intell. 23 (2001) 800–810.
[19] A. Dix, J. Finlay, G. Abowd, R. B e ale, Human Computer Interaction, third ed., Prentice Hall, 2004.
[20] J. Philbin, O. Chum, M. Isard, J. Sivic, A. Zisserman, Object retrieval with large vocabularies and fast spatial matching, in: Proc. of IEEE CVPR, 2007, pp. 1–8.
[21] O. Chum, J. Philbin, J. Sivic, M. Isard, A. Zisserman, Total recall: automatic query expansion with a generative feature model for object retrieval, in: Proc. of IEEE ICCV, 2007, pp. 1–8.
[22] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2004) 91–110.
[23] R. Jain, P. Sinha, Content without context is meaningless, in: Proc. ACM Multimedia, 2010, pp. 1259–1268.
[24] I. Guy, A. Jaimes, P. Agulló, P. Moore, P. Nandy, C. Nastar, H. Schinzel, Will recommenders kill search?: recommender systems-an industry perspective, in: Proc. of ACM Conference on Recommender Systems, 2010, pp. 7–12.
[25] S. Robertson, H. Zaragoza, The probabilistic relevance framework: BM25 and beyond, Inform. Retrieval 3 (2009) 333–389.
[26] L. Yang, B. Geng, Y. Cai, A. Hanjalic, X. Hua, Object retrieval using visual query context, IEEE Trans. Multimedia 13 (2011) 1295–1307.