Interactive Mobile Visual Search for Social Activities Completion Using Query Image Contextual Model

Ning Zhang[†], Tao Mei[‡], Xian-Sheng Hua[‡], Ling Guan[†], Shipeng Li[‡]

[†] *Ryerson Multimedia Research Laboratory, Ryerson University*
Toronto, Ontario, Canada, M5B 2K3
{n2zhang,lguan}@ee.ryerson.ca

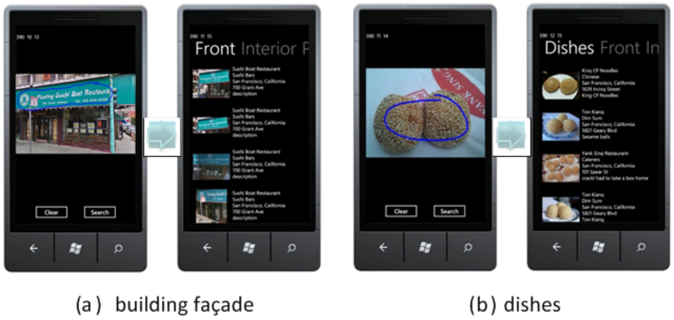
[‡] *Microsoft Research Asia, Beijing, P. R. China, 100080*
{tmei,xshua,spli}@microsoft.com

***Abstract*—Mobile devices are ubiquitous. People use their phones as a personal concierge not only discovering information but also searching for particular interest on-the-go and making decisions. This brings a new horizon for multimedia retrieval on mobile. While existing efforts have predominantly focused on understanding textual or a voice query, this paper presents a new perspective which understands visual queries captured by the built-in camera such that mobile-based social activities can be recommended for users to complete. In this work, a query image-based contextual model is proposed for visual search. A mobile user can take a photo and naturally indicate an object-of-interest within the photo via circle based gesture called "O" gesture. Both selected object-of-interest region as well as surrounding visual context in photo are used in achieving a search-based recognition by retrieving similar images based on a large-scale of visual vocabulary tree. Consequently, social activities such as visiting contextually relevant entities (i.e., local businesses) are recommended to the users based on their visual queries and GPS location. Along with the proposed method, an exemplary real application has been developed on Windows Phone 7 devices and evaluated with a wide variety of scenarios on million-scale image database. To test the performance of proposed mobile visual search model, extensive experimentation has been conducted and compared with state-of-the-art algorithms in content-based image retrieval (CBIR) domain.**

I. Introduction

Mobile devices play vital roles in our daily life, from their original function of telephony, to prevalent information-sharing terminals, and recently developed tens of thousands of applications. While on the go, people are using their phones as a personal concierge discovering what is around and deciding what to do.

Comparing with conventional text or voice queries for information search on-the-go, there are many cases that visual queries can be more naturally and conveniently expressed via mobile device camera sensors (such as an unknown object or text, an artwork, a shape or texture, and so on) [1]. However, conventional content-based image retrieval (CBIR) suffers from a semantic gap, in which one possible reason could be "a picture is worth a thousand words". This can be interpreted such that too much information provided by a single image can confuse the computer vision system for recognition.



(a) building façade

(b) dishes

Fig. 1. A snapshot of two different scenarios. A user can take a photo, specify his/her interested object, and then get the search result and recommendation for social activities completion.

Motivated by the above observation, we present in this paper our recently developed system, which is an interactive search-based visual recognition model aiming at completing social activities. A natural user interaction is proposed to circle (also called "O" gesture) the region of interest (ROI) from the query image. Thus, user intelligence is integrated to select the target object. We propose a novel context-embedded vocabulary tree, which incorporates the context from surrounding pixels of "O" in order to search similar images from a large-scale of images. Through "O" gesture based user interaction, standard visual recognition can be improved. Consequently, the system is able to recommend activities for completion such as relevant entities and interesting local business, by incorporating the recognition result with the rich contextual information. Figure 1 shows the example scenarios snapped from our realized application.

The contribution of this work is three-fold: 1) a natural user interface that enables users to interactively specify their search intents more conveniently, 2) an innovative image contextual model for retrieval which leverages pixel context and user interaction such that not only ROI of user selection but also surrounding visual information of the object-of-interested are incorporated, and 3) a contextual entity suggestion for activity completion that provides meaningful and contextually relevant recommendations.

The rest of the paper is organized as follows. Section II reviews related work. Section III presents the details of the proposed interactive visual search mechanism using image contextual model for social activities completion. Experiments and evaluations are given in Section IV, followed by conclusions in Section V.

II. Related work

Due to its potential of practicability, mobile visual recognition and search has drawn extensive attention from research community. For instance, 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 [2]. 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 [3].

On the other hand, contextual features such as location information has been adopted and integrated successfully to the mobile-based visual search. Schroth *et al.* utilized GPS information and segmented the search area from a large environment of city to several overlapping subregions to accelerate the search process with a better visual result [4]. 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, and etc. [5].Girod *et al.* investigated the mobile visual search problem from a holistic point of view with practical analysis under mobile device constraints, such as memory, computation, devices, power and bandwidth [6]. An extensive analysis with various feature extraction, feature indexing and matching, is conducted using real mobile-based Stanford Product Search system. They demonstrated a low-latency interactive visual search framework with satisfactory performance.

The aforementioned visual search methods and applications on mobile have demonstrated their merits. Alternatively, we believe that combining visual recognition techniques with personal and local information will provide contextually relevance recommendations. Hence, this work proposes a mobile visual search model to suggest potential social activities on-the-go.

In this work, the smart phone hardware of camera and touch screen are taken advantage of in order to facilitate the expressions of user's ROI from the pictures taken. The visual query along with such a ROI specification then go through an innovative contextual visual retrieval model to achieve a meaningful connection to database images and their associated rich text information. Once the visual recognition is accomplished, associated textual information of retrieved images are further analyzed to provide meaningful recommendation.

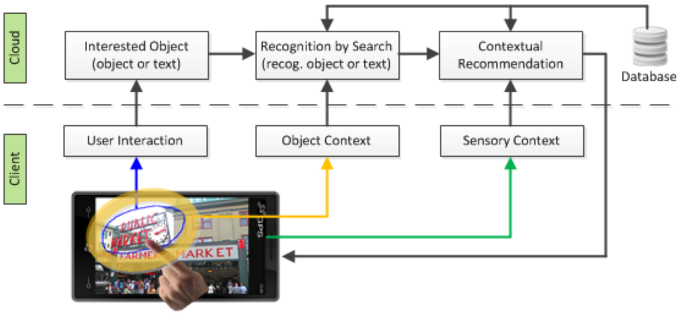

Cloud
Interested Object (object or text)
Recognition by Search (recog. object or text)
Contextual Recommendation
Database
Client
User Interaction
Object Context
Sensory Context
PUBLIC MARKET
FARMERS MARKET

Fig. 2. The proposed framework of mobile visual search and activity completion model using image contextual model, including 1) "O"-based user interaction, 2) image context model for visual search, and 3) contextual entity recommendation for social activities.

III. Interactive Visual Search Model for Activities Completion

A. Overview

Figure 2 shows the framework of our visual recognition and activity recommendation model. It can be in general divided into the client-end and cloud-end. On the client, a user's visual search intent is specified by the "O" gesture on a captured image. On the cloud, with user selected object and the image context around this object, a recognition-by-search mechanism is applied to identify user's visual intent. 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. Finally, the specified visual search results 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 the social activities.

The "O" gesture utilizes the multi-touch screen of the smart phone. Users don't need any training and can naturally engage with the interface right away. After the trace has been claimed 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 contains N pixel-wise positions (x_j, y_j). We applied the principle component analysis (PCA) to find the two principle components. The purpose of this work is to formulate a boundary of the selected region from the arbitrary "O" gesture trace. We also calculated the mean μ and covariances Σ based on $\mathbf{D}$ and non-correlated assumption along the two principle components:

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

Figure 3 shows the computation of principle components from the "O" query. Once the principle components are identified, proposed image contextual model for mobile visual search is used to identify the interested object indicated by the user.

B. Image Contextual Model for Mobile Visual Search

The visual search method is based on a recognition scheme using the vocabulary tree proposed by Nister *et al.* [7]. This

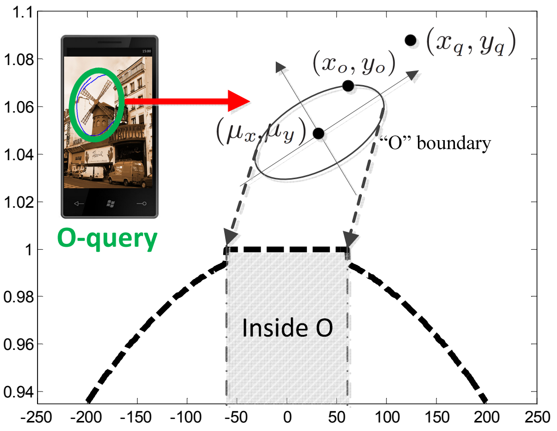


Fig. 3. The illustration of user indicated "O" query, and the computation of principle components of the query. (μ_x, μ_y) is the center of "O" query, (x_o, y_o) is a pixel on the "O" boundary, and (x_q, y_q) is a query pixel.

method provides a fast and scalable mechanism and is suitable for large-scale and expansible database. Moreover, when we adapt this method in the mobile domain, the "O" gesture fits naturally to provide a focused object selection for better recognition. Different from using the entire image as the visual query in [7], we have the user-indicated ROI from the "O" gesture (called "O" query) in the proposed system. We design a novel context-aware visual search approach in which a context-embedded vocabulary tree (CVT) is built to take the surrounding pixels around the "O" query into consideration.

In the CVT, each leaf node corresponds to a visual word i, associated with an inverted file (with the list of images containing this visual word i). Note that we only need to consider the images d in the database with the same visual words 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

$$\begin{aligned} s(q,d) &= \| \mathbf{q} - \mathbf{d} \|_2^2 \\ &= \Big(\sum_{i|d_i=0} |q_i|^2 + \sum_{i|q_i=0} |d_i|^2 + \sum_{i|q_i\neq 0, d_i\neq 0} |q_i - d_i|^2 \Big) \end{aligned} \tag{2}$$

where $\mathbf{q}$ and $\mathbf{d}$ denote the feature vectors of the query q and image d in the database, which are represented by the vector of tf-idf elements q_i and d_i (i denotes the i-th visual word in the vocabulary tree), respectively. q_i and d_i are the *term frequency-inverse document frequency* (tf-idf) value for the i-th visual word in the query and the image, respectively, which can be given by

$$\begin{aligned} q_i &= tf_{q_i} \cdot idf_i, \qquad (3) \\ d_i &= tf_{d_i} \cdot idf_i. \qquad (4) \end{aligned}$$

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 visual word i (i.e., the images whose descriptors are classified into the leaf node i). tf_{q_i} and tf_{d_i} are the numbers of i-th visual words in q_i and d_i, respectively.

One possible way to adapt this standard image search to our context-aware search is to simply use the "O" query as the query input, without considering the surrounding pixels around "O". This is equivalent to using a "binary" weights of the *term frequency* tf_{q_i}: the weight is 1 inside "O" and 0 outside "O". As we have mentioned 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* $tf^o_{q_i}$ for the "O" query. This is to provide a "soft" weighting of the *term frequency* by incorporating the image context outside the "O" query, which was neglected in the standard binary scheme [7]. When quantizing descriptors in the proposed CVT, the $tf^o_{q_i}$ for the "O" query for a particular visual word i is formulated as:

$$tf^o_{q_i} = \begin{cases} tf_{q_i}, & \text{if } q_i \ \in \text{O} \\ tf_{q_i} \cdot \min\big\{1, \frac{\Re(x_q,y_q)}{\Re(x_o,y_o)}\big\}, & \text{if } q_i \ \notin \text{O} \end{cases} \tag{5}$$

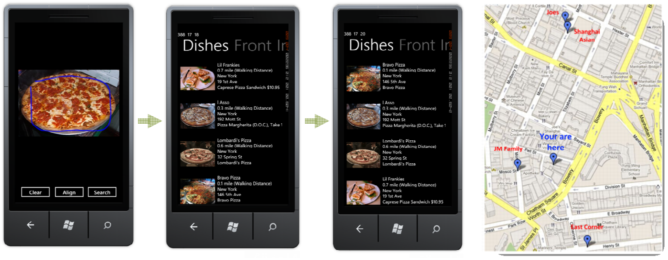
where $\Re(x_o, y_o)$ and $\Re(x_q, y_q)$ denote the Gaussian distance of the pixel (x_o, y_o) and (x_q, y_q) with respect to the center of "O" query (μ_x, μ_y). Figure 3 shows the definition of these pixels in the query image q. The Gaussian distance $\Re(x, y)$ for an arbitrary pixel (x, y) is given by

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

The "soft" weighting is a piece-wised 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 q is outside the "O" query, the modulating term is $\min\big\{1, \frac{\Re(x_q,y_q)}{\Re(x_o,y_o)}\big\}$, which is to guarantee that the soft weighting is always 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 the boundary position (x_o, y_o). The mean values μ_x and μ_y are calculated from "O" gesture sample data, while α and β are tunable parameters to control the standard deviation for the bivariate normal distribution. A is the amplitude value that could be set as $A = 1.0$. α and β reflect the importance of the horizontal and vertical axis (or directions) when employing the PCA technique. Empirically, we set $\alpha = 5$, and $\beta = 1$, which indicates that the horizontal axis is usually more important than the vertical, as users prefer to take pictures using their phone cameras horizontally.

C. Social Activities Completion

After the visual search and obtaining the best candidate result from recognition, we utilize the powerful and rich metadata as a better feature to search again and rerank the results using rich context to achieve the final activity completion. To be specific, we want to achieve a recommendation based on the text retrieval result of the description associated with the top searched image result. The Okapi BM25 ranking function is used to compute a similarity score based on the text similarity



O-query

Top-1 visual search result, "Bleecker Street Pizza at 69 7th Ave S. New York"

Nearby restaurants serving "Pizza" via text keyword search

Fig. 4. This is a result of recommendation list, which is also being visualized in a map to help users to picture the distances between query and the results, so that a better social activity completion can be accomplished.

[8]. We extracted the keywords $Q_t = \{q_{t_1}, q_{t_2}, ..., q_{t_n}\}$ by projecting the text query to a quantized dictionary. Then we compute the relevance score of query Q_t and database image descriptions D_t. Detailed score computation techniques can be referred in the reference [8]. As the last step, we rerank the searched results based on the GPS distance of the user's current location. The approach we adopted in this work is based on an effective region selection using the quadkey representation in geolocation hashing [9]. A quadratic based hashing code is formed for each GPS enabled image. Associated hashing codes are used for calculating hamming distance and mapped to the real ground distance. The final GPS result is ranked to provide the location-based recommendation result. Figure 4 demonstrates a sample result of the recommendation list.

IV. Experiments

A. Data and Settings

The client-end application is developed on a Windows Phone 7 HD7 model with 1GHz processor, 512MB ROM, GPS sensor and 5 megapixel color camera. At the cloud, a total of one million visualwords were created from 100 million local descriptors (SIFT descriptor with difference of Gaussian detection in this experiment) points, with a hierarchal tree structure used, consisting of 6 level of branches, each of which has 10 sub-branches or nodes. In constructing the vocabulary tree, each visualword takes up to 168 bytes storage, where 128 bytes are for SIFT local feature storage, and 4 bytes for 10 subordinate children nodes connection. In total, 170 megabytes 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-based crawling tool. 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 is a crawled commercial local services data, mainly focusing on the restaurant domain. A total of 332,922 images associated with 16,819 restaurant entities from 12 US cities were crawled with associated metadata.

B. Evaluation Metrics

We use both mean average precision (MAP) and normalized discounted cumulative gain (NDCG) for the evaluation. The average precision (AP) formula is presented as

$$AP_n = \frac{1}{min(n, P)} \sum_{k=1}^{min(n,S)} \frac{P_k}{k} \times I_k \tag{7}$$

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 the 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.

We adopt Normalized Discounted Cumulative Gain (*NDCG*) as the performance metric. Given a query q, the *NDCG* at the depth d in the ranked list is defined by:

$$NDCG_d = Z_d \sum\nolimits_{j=1}^{d} \frac{2^{r^j} - 1}{\log(1 + j)} \tag{8}$$

where r^j is the rating of the j-th pair, Z_d is a normalization constant and is chosen so that the $NDCG@d$ of a perfect ranking is 1.

C. Evaluation of image contextual model for mobile visual search

We investigated on the image contextual information and its effectiveness on the visual search technique, using soft weighting scheme. For the bivariate based function $\Re(x, y)$, we fix the amplitude A to 1 and tune two parameters α and β to modulate the standard deviation. Furthermore, GPS context is also taken into account so that meaningful results can be provided to users. In this setting, Figure 5 demonstrates that soft weighting approaches with 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 been drop to 2% and less than 1% for MAP and NDCG respectively.

The significance of this image contextual information with soft weighting scheme allows a robust user behavior and is seamlessly glued with the "O" gesture, which is spontaneous and natural. Unfortunately, "O" is also inevitably lack of accuracy due to the device limitation in outlining the boundary. However, soft weighting alleviates this deficient of correctness in object selection and provides a robust approach to accommodate the behavioral errors.

We also implemented the state-of-the-art contextual image retrieval model (CIRM), and compared the performance with our proposed mobile visual search model. The CIRM has demonstrated a promising result in desk-top based CBIR by applying a rectangular bounding box in emphasizing the ROI, which is easy to be manipulated using mouse at a desktop environment [10]. The weighting scheme in their work is to use two logistic functions joining at the directional (either X or Y) center of the bounding box. Then, the term frequency

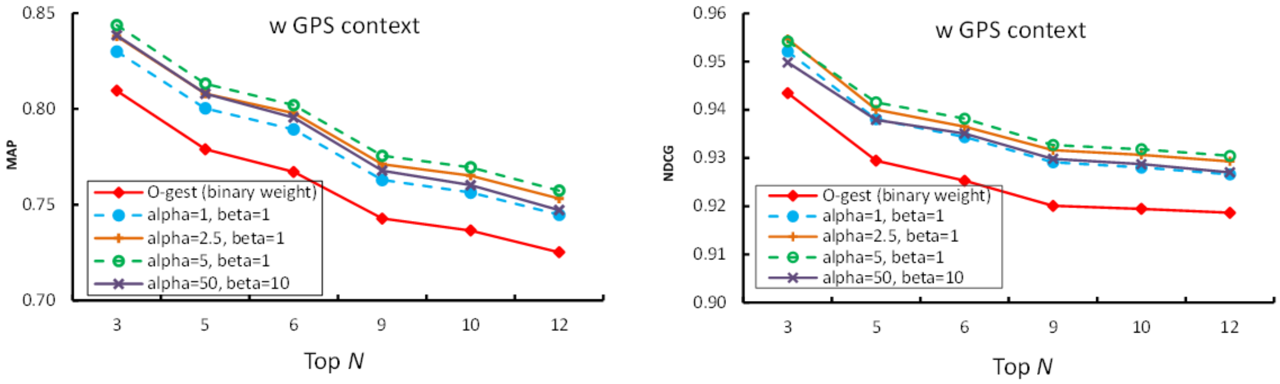


Fig. 5. Image contextual-based recognition by various parameter α and β, with GPS context.

tf_q is formulated as:

$$\begin{aligned} tf_q \propto \quad & min(\frac{1}{1+exp(\delta_X(x_l-x_i))}, \frac{1}{1+exp(\delta_X(x_i-x_r))}) \\ * \quad & min(\frac{1}{1+exp(\delta_Y(y_t-y_i))}, \frac{1}{1+exp(\delta_Y(y_i-y_b))}) \end{aligned} \quad (9)$$

where x_l, x_i, x_r represent x pixel values of left boundary, detected feature point, and right boundary along the x-axis direction, respectively. Similarly, y_t, y_i, y_b are the y pixel values of top boundary, detected feature point, 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 to value 0, the further x_i from the bounding box; while ideally close to value 1 when feature point is inside the bounding box. δ_X and δ_Y are two tunable parameters for finding the best performance of bounding box. Detail algorithm and explanation can be found in reference [10].

Figure 6 depicts the MAP and NDCG measurements using the GPS context information, by comparing the proposed Gaussian-based soft weighting contextual method with the CIRM model as well as the CBIR benchmark using the whole query image without contextual model. Again, the proposed method outperformed both the CIRM and the CBIR approaches. However, the best performance of the CIRM model at $dX = 0.0001$ and $dY = 0.0001$ is closed to the performance of our proposed contextual model at $\alpha = 5$ and $\beta = 1$.

D. *Evaluation of mobile activities completion*

For the mobile activity completion, our approach is to use the visual photo taken by users as the starting point. We target at providing a recommendation list based on the text search which associated with the recognized object and GPS context reranking. We take the approach on firstly identifying the object and matching it to the dataset. We then use the matched result metadata as the text query to do a text-based search. The final result is then reranked 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 random picked 306 photos and manually labeled and categorized them into 30 featured themes of food dishes, such as a beef, soup or a burger. We built a 300 words by extracting the most frequently used words in the image description.

TABLE I
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

In order to create a real scenario, we printed out the dishes in a menu style with both texts and images. We then took the pictures of the dishes as the visual query and tried to find the duplicated/near-duplicated images from the dataset. We assumed that the best match of the visual recognition result is user's true search intent. Such an intent was then carried by the associated metadata, which were quantized using the prepared 300-word dictionary. The quantized words were then searched with a returned rank list based on the text similarity. The final step is to rerank the result list using GPS distance.

Table I presents the MAP results with the initial visual query and newly formatted description query after visual recognition, with a comparison of visual-based search result. The Table demonstrates that the performance of the description-based search is much better than the visual-based search. This is reasonable in a sense that text is a better description than the visual content. However, the merit of the visual input is its role in filling the niche when an individual doesn't have the language tool to express him/herself articulately. This can be demonstrated at the initial visual search (@0), where the result is at a high precision rate of 96.08%. Such a high accuracy provides a solid foundation to use the associated metadata for the second stage description-based search. Once the visual query is mined accurately, the role of search query is then passed from visual content to text metadata for a better search result.

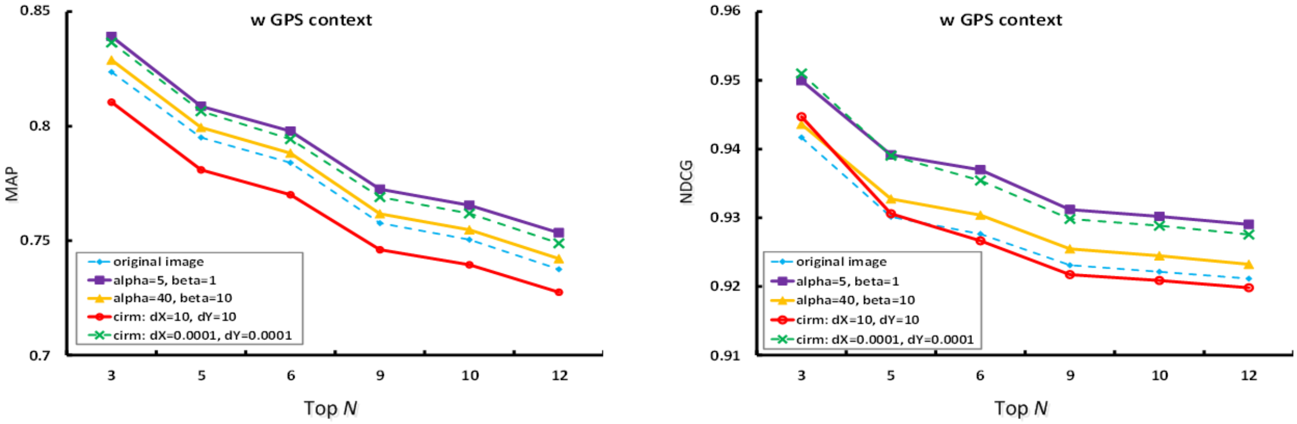


Fig. 6. Comparison of image contextual-based recognition model by various parameter α and β, with the conventional CBIR (original), as well as CIRM approach with parameter dX and dY, with GPS context.

E. Video Demonstration

We also have uploaded a video demo to showcase the system which we codenamed TapTell. The video speed is set to x1.7 than the original footage to make this video demo more compact and agreeable to watch [1].

V. Conclusion

This paper presents an interactive visual search using query image context to complete activities such as local businesses recommendation. We demonstrated a natural and agreeable user-mobile interaction "O" behavior. We also proposed a soft weighting scheme for using both "O" object and its surrounding context information from the query image to achieve the mobile visual search. We evaluated various weighting scheme with GPS conditions and verified that image context outside the "O" region plays a constructive role in improving the search performance. We also compared our approach with the state-of-the-arts algorithms, and it is demonstrated that the proposed method outperformed both the conventional CBIR and the CIRM approaches. Once the context information is associated with the visual query, more reliable contextual text and GPS features are taken advantaged of in searching and reranking and ultimately recommending interesting and related activities.

For the future work, we plan to further investigate other recommendation schemes, including both social and personal status of the mobile users. In addition, we want to utilize other local feature such as low bit-rate descriptors as well as dense-based sampling SIFT points in order to achieve a better object recognition performance.

References

[1] J. Smith, "Clicking on Things," *IEEE MultiMedia*, vol. 17, no. 4, pp. 2–3, 2010.

[2] V. Chandrasekhar *et al.*, "CHoG: Compressed histogram of gradients A low bit-rate feature descriptor," pp. 2504–2511.

[3] S. Tsai *et al.*, "Location coding for mobile image retrieval," in *Proceedings of the 5th International ICST Mobile Multimedia Communications Conference*, 2009, pp. 1–7.

[4] G. Schroth *et al.*, "Mobile visual location recognition," *Signal Processing Magazine*, vol. 28, no. 4, pp. 77–89, 2011.

[5] L.-Y. Duan and W. Gao, "Side Discriminative Mobile Visual Search," in *2nd Workshop on Mobile Visual Search*, 2011.

[6] B. Girod *et al.*, "Mobile visual search," *Signal Processing Magazine*, vol. 28, no. 4, pp. 61–76, 2011.

[7] D. Nister and H. Stewenius, "Scalable recognition with a vocabulary tree," in *Proc. IEEE CVPR*, 2006, pp. 2161–2168.

[8] S. Robertson and H. Zaragoza, "The probabilistic relevance framework: BM25 and beyond," *Information Retrieval*, vol. 3, no. 4, pp. 333–389, 2009.

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

[10] L. Yang, B. Geng, A. Hanjalic, and X. Hua, "Contextual image retrieval model," in *Proceedings of the ACM International Conference on Image and Video Retrieval*. ACM, 2010, pp. 406–413.

[1] **http://www.viddler.com/v/c56ad66c**