Contents lists available at ScienceDirect

Neurocomputing

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

Socio-mobile landmark recognition using local features with adaptive region selection

Chunjie Zhang [a], Yifan Zhang [b,*], Xiaobin Zhu [c], Zhe Xue [a], Lei Qin [d], Qingming Huang [a,d], Qi Tian [e]

[a] School of Computer and Control Engineering, University of Chinese Academy of Sciences, 100049 Beijing, China
[b] National Laboratory of Pattern Recognition, Institute of Automation Chinese Academy of Sciences, P.O. Box 2728, Beijing, China
[c] Beijing Technology and Business University, Beijing, China
[d] Key Lab of Intell. Info. Process, Institute of Computing Technology Chinese Academy of Sciences, Beijing 100190, China
[e] Department of Computer Sciences, University of Texas at San Antonio, TX 78249, USA

ARTICLE INFO

Article history:
Received 30 September 2013
Received in revised form
23 October 2014
Accepted 25 October 2014
Available online 22 May 2015

Keywords:
Mobile
Social relationship
Adaptive region selection
Local feature
Geo-codebook
Visual recognition

ABSTRACT

With the fast development of mobile devices as well as the broadband wireless network, mobile devices are playing a more and more important role in people's daily life. Nowadays, many landmark images are captured by mobile devices. However, these images are often captured under different lightening conditions with varied poses and camera orientations. Besides, people are inherently connected by personal interests as well as various interactions. To alleviate the imaging problem with mobile devices as well as take advantage of the social information for mobile visual applications, we propose a novel socio-mobile visual recognition method using local features with adaptive region selection. We densely extract local regions and use the pixel gradients to represent each local region. Each local region is divided into 4×4 subregions to combine the spatial information. Instead of using fixed pixel numbers for each subregion, we adaptively choose the proper size of each subregion to cope with varied poses and camera orientations. The most discriminative local features are then chosen by minimizing the sparse coding loss. Besides, a geo-discriminative codebook is also generated to take advantages of images' location information. Moreover, we jointly consider the visual distances as well as user's friends' matching results to further boost the final visual recognition performance. We achieve the state-of-the-art performance on the Stanford mobile visual search dataset and the San Francisco landmark dataset. These experimental results demonstrate the effectiveness and efficiency of the proposed adaptive region selection based local features for socio-mobile landmark recognition.

1. Introduction

Recently the use of mobile devices for visual applications become more and more popular. With the fast development of hardware, mobile phones are equipped with powerful imaging devices, global positioning system (GPS) as well as broadband wireless networks. This enables new visual applications such as searching similar images [1], finding products [2] and landmark recognition [3–5] with mobile devices. Some commercial systems are also introduced such as Google Goggles [6], Nokia Point and Find [7], Amazon Snaptell [8] and Kooaba [9]. Most of these mobile visual application systems adopt the popular client/server (C/S) architecture.

Although a lot of work have been done to improve the user experience with mobile visual applications, there are still some challenges that cannot be solved very well until today. Landmark images and videos are often captured under uncontrolled conditions using various types of mobile devices with relatively low quality cameras compared with digital cameras. This means we have to design more discriminative features for better image representation. Researchers have proposed multiple types of features [3,10–15] to solve this problem. Although very effective, the performance of these simplified methods is relatively low compared with the corresponding methods implemented on the personal computers. For example, the compressed histogram of gradients (CHoG) in [13] is a simplification of histogram of gradients (HoG) for computational and transmission reasons of mobile devices. This is because of the computational limitations of mobile devices. Besides, most of these features cannot cope with large varied poses and camera orientations. However, this often happens with images captured by mobile devices. Fig. 1 shows some example images

* Corresponding author.
E-mail addresses: zhangcj@ucas.ac.cn (C. Zhang), yfzhang@nlpr.ia.ac.cn (Y. Zhang), brucezhucas@gmail.com (X. Zhu), zhxue@jdl.ac.cn (Z. Xue), qinlei@ict.ac.cn (L. Qin), qmhuang@jdl.ac.cn (Q. Huang), qitian@cs.utsa.ed (Q. Tian).

http://dx.doi.org/10.1016/j.neucom.2014.10.105
0925-2312/© 2015 Elsevier B.V. All rights reserved.

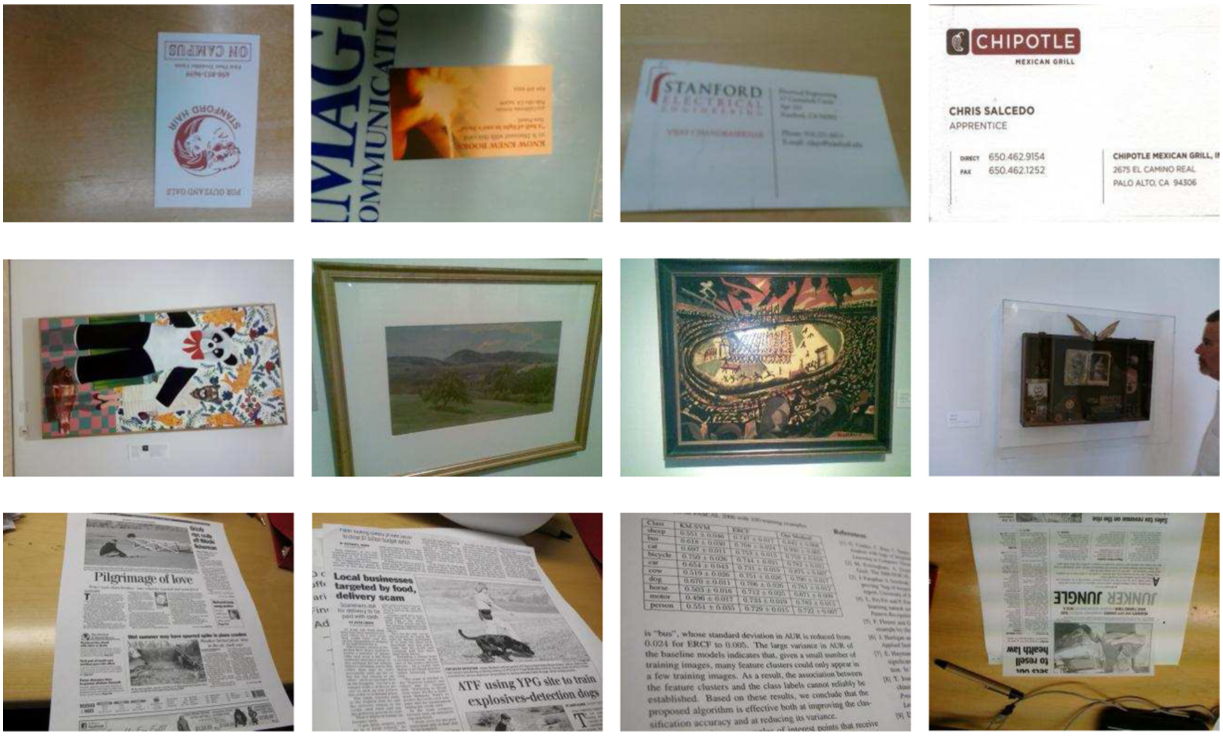


Fig. 1. Example images captured with various mobile devices of different rotations and poses. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

captured by mobile devices. Since mobile devices are powered with batteries, it is unwise to conduct too much computation on the mobile devices. All these limitations hinder the performance of mobile related visual applications.

To overcome the limitations with general visual applications mentioned above, researchers resorted to more specific applications with domain knowledge that can be used, such as mobile landmark recognition [1,3] and achieved good performances. However, how to extract proper descriptors efficiently and effectively is still a very challenging problem even with specific visual applications. Fortunately, with the fast development of the 3G/4G broadband wireless networks, the transmission speed is dramatically improved. This makes it possible to transmit images very quickly and do the feature extraction and analysis on the server side instead.

With the fast development of world wide web 2.0, online social platforms are becoming indispensable for consumers, such as Flickr [16], Youtube [17] and twitter [18]. Besides, users are inherently connected with various types of relationships. The uses of social information for analysis and marketing have become very popular. By viewing individual user as a part of groups, the social relationship modeling technique tries to model the similarities and sharings of individual user. The uses of social relationship for visual applications have also been widely explored by researchers [19–25] with single source domain as well as multiple domains. The learned knowledge can be used for advertising [26,27], event recognition [25] as well as media analysis [24] with encouraging performances. However, the social relationship modeling is task dependent. We need to make use of well chosen social relationships for specific tasks.

Mobile devices are often used as an efficient tool for social content exploration and management. Various online social platforms have become indispensable for everyone's daily life. The marriages of social and mobile access pose challenges as well as opportunities for new socio-mobile visual applications. On one hand, the ‘mobile’ facet plays emphasis on the context, personal interests as well as the interactions between individual users. On the other hand, ‘social’ emphasizes the sharing and similarity aspects of individuals. How to efficiently make use of both the social and mobile aspects of visual contents becomes a hot topic and should be carefully addressed.

In this paper, we propose a novel socio-mobile landmark recognition method using local features with adaptive region selection. Instead of extracting features on the mobile devices, we send images to the server where local features with adaptive regions are extracted. This enables us to make use of the powerful servers for efficient image representation and save the power of mobile devices. We adopt the dense local feature extraction strategy and use the pixel gradients to represent each local region. Each local region is sub-divided into spatial subregions to combine the spatial information. Instead of using fixed pixel numbers for each subregion (e.g. 4×4 pixels for each subregion and 4×4 subregions are used for the SIFT descriptor), we propose to adaptively choose the proper size of each subregion to cope with varied poses and camera orientations. The discriminative local features are then chosen to represent images with minimum encoding loss. Besides, a geo-discriminative codebook is generated to take advantages of image's location information which eventually helps to improve the visual recognition performance. Moreover, for a query image, we jointly consider the visual distances as well as the user's friends' matching results to further boost the final performance. Experimental results on two public landmark datasets prove the effectiveness and efficiency of the proposed adaptive region selection based local features for socio-mobile landmark recognition. Fig. 2 shows the flowchart of the proposed algorithm.

The rest of this paper is organized as follows. We give the related works in Section 2. In Section 3, we show the details of the proposed adaptive region selection based local feature extraction for socio-mobile landmark recognition. Experimental results on two public datasets are given in Section 4. Finally, we conclude in Section 5.

2. Related work

With the fast development of mobile devices, mobile visual applications have become more and more popular and have drawn

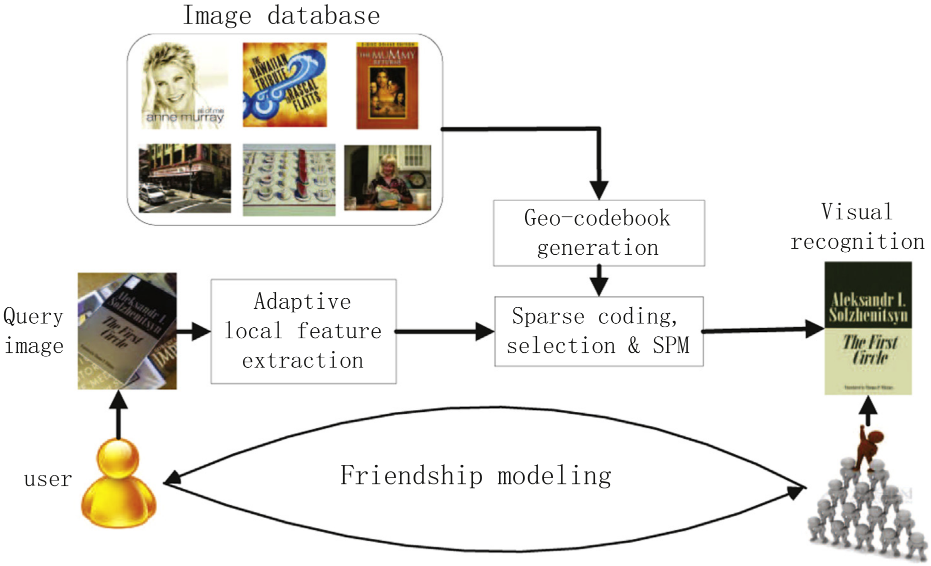


Fig. 2. Flowchart of the proposed socio-mobile landmark recognition using local features with adaptive region selection.

a lot of researchers' attention [1–5]. Chandrasekhar et al. [1] released a mobile visual search dataset with camera-phone images of products, CDs, business cards, etc. A comparison with different types of local region detection and description methods was also made [1]. He et al. [2] encoded each local feature into a very small number of hash bits. A boundary feature was then used in the reranking step to describe the object shapes for efficient mobile product search with good performance. Mobile landmark recognition was also widely studied. Ji et al. [3] proposed a location discriminative vocabulary coding for landmark retrieval. The geo-information of images were used to learn the codebooks. Each image was encoded by the corresponding codebooks. To cope with the dataset bias, an online codebook adaption scheme was adopted. Gall et al. [28] also proposed an online codebook generation approach for visual tracking. Xiao et al. [4] tried to classify landmark images with 3-D modeling. Since the 3-D modeling requires a number of training images, the method proposed in [4] had to obtain enough images to get reliable results. Bimbo et al. [5] studied the problems in a camera network by exploiting distinctive visual landmark maps with pan-tilt-zoom camera. The use of different parameter settings for local feature extraction and minimizing the joint within-class and between class distances for selection was also proposed by Eitzinger et al. [29] which was further optimized by adaptively choosing the parameters [30].

Instead of transmitting images directly, researchers tried to extract compact visual features [3,10–15] from images in order to alleviate the transmission burden. Lowe [10] proposed SIFT features which is widely used for visual applications with good performance. Mikolajczyk and Schmid [11] systematically evaluated the performance of multiple types of local descriptors. Bay et al. [12] introduced the speeded up robust features (SURF) to cope with noise. Chandrasekhar et al. [13] proposed the compressed histogram of gradients descriptor (CHoG) and was widely used with mobile visual applications. This descriptor tried to compress the histogram of gradients descriptor [31] with about 20 reduction in bit rate. Jiang et al. [14] used compact codes with query-adaptive ranking for similar image search. Liu et al. [15] studied the mobile routing problem. The use of locality sensitive hashing (LSH) [32] and its variants [33,34] were also explored for mobile visual applications. Most of previous works tried to extract local features with region detection to save computational cost, such as Harris–Laplace, Hessian–Laplace and the Difference-of-Gaussian region detectors [10,35]. However, the use of dense local features has been proven more effective than region detectors for visual applications [36,37]. Besides, to make use of different color channels, Sande et al. [38] proposed to use the color SIFT descriptor which improved the performance of object and scene recognition. Although the SIFT is scale and rotation invariant, it is unable to cope with the latitude and longitude angle changes. To solve this problem, Morel and Yu [39] proposed the affine SIFT descriptor and was further extended by Kulkarni and Li [40] for image classification. Xu et al. [41] used regularized LDA for tag refinement while Wang et al. [42,43] tried to search images in one machine with nearest neighbor search.

With the fast development of social network, the use of social information for visual applications [19–27] has become popular. Lin et al. [19] proposed an innovative, scalable spectral analysis framework for internet scale monitoring of multirelational social media data, encoded in the form of tensor streams. Rota et al. [20] proposed a real time detection algorithm of social interactions in surveillance video with good performance. Yang et al. [21] mined the label correlation with relaxed visual graph embedding and applied it to web and personal image annotation. A neighbor voting scheme was used by Li et al. [22] in order to learn the social tag relevance while Bao et al. [23] mined social emotions from affective text. To make use of the social information from multiple domains, Liu et al. [24] proposed a hybrid social media network to mine the content, concept and user relationships. An efficient optimization algorithm was also proposed to solve this problem. Duan et al. [25] exploited web images for event recognition in consumer videos using a multiple source domain adaptation approach. The use of social network for advertising was also studied by researchers [26,27]. Aly et al. [26] tried to solve the targeting problem with web-scale user modeling while Singh et al. [27] rewrote the null e-commerce queries to recommend products. McAuley and Leskovec [44] studied the problem of classifying images with social-network metadata. They found that social-network metadata are very useful or even be able to outperform methods based on image content. Besides content based technique, collaborative filtering [45] is another efficient technique which is widely adopted for social recommendation. It tries to filter for information using collaboration among different users and viewpoints. Active learning [46,47] can also help to incorporate more information for visual applications.

The use of location information to assist visual applications has also become popular, especially with the popularity of the global positioning system (GPS). Nowadays, more and more mobile devices are equipped with GPS which can capture geo-tagged images. This is very effective for image analysis, especially for visual applications that depend on location information [1,3,5,48], such as landmark recognition. To represent images, a geo-related codebook was constructed with sparse coding [49]. A city-scale landmark dataset with the GPS information captured using a mobile mapping vehicle of San Francisco was introduced by Chen et al. [48]. Takacs et al. [50] also proposed to use the GPS signal to retrieve nearby images which alleviated the computation cost and also improves the final performances. The use of text detection and recognition technique was also proposed by Wang et al. [51] with encouraging performance. Zha et al. [52] used multi-camera context for group activities detection.

3. Adaptive region selection based local feature extraction for socio-mobile landmark recognition

In this section, we give the details of the proposed adaptive region selection based local feature extraction method for socio-mobile landmark recognition. We show how to extract local feature with adaptive region in Section 3.1. The generation of geo-codebook and the selection of discriminative local features are given in Section 3.2. Section 3.3 shows how to model the social relationship for landmark recognition.

3.1. Adaptive region based local feature extraction

The use of local feature has become very popular and its effectiveness has been proven by many researchers for various visual applications. The local feature based image representation goes one step beyond image pixels hence helps to alleviate the semantic gap. Basically, there are two components for local feature representation: local region selection and local feature extraction.

First, a local region should be selected either by detection algorithms or pre-defined sampling. For region detection, the Harris–Laplace, Hessian-Laplace, Difference-of-Gaussian region detectors [10,35] are used while dense sampling is widely chosen as an effective pre-defined sampling method. We choose to use the dense sampling scheme for local region selection as it has been shown more effective than region detectors. However, traditional dense local region selection methods may not be able to cope with the large pose variation and illumination changes of mobile device captured images. Besides, the resolutions of mobile captured images vary with the types of devices used. Fig. 1 shows some example images with varied poses and illumination changes captured by various mobile devices.

To alleviate these problems, instead of fixing or simply resizing the size of local regions, we propose to adaptively select local regions in this paper. Usually, local region is divided into 4×4 sub-regions with equal size, such as SIFT and HoG. This scheme is very effective because it combines the spatial information of image pixels within the local region. We follow this successful sub-region setup but choose sub-regions adaptively. This is achieved by changing the size of each sub local region. Take the 16×16 pixels local region setup for example, each sub-region is of 4×4 pixels. For each horizontal and vertical sub-region, we adaptively choose the number of pixels from 4 to 2 with one pixel interval. Fig. 3 shows a simplified adaptive local region selection of 2×2 sub-regions with only vertical pixel number changes. The extension to the 4×4 sub-regions is straightforward. Instead of extracting one local feature for each local region, we are able to extract multiple local features to describe one local region with this adaptive selection technique. The HoG features of these local regions are then extracted and the most representative local region is chosen by minimizing the reconstruction error described in Section 3.2. Instead of using one feature to represent each local region, we can extract multiple features using the adaptive local region selection strategy. This adaptively selection strategy simulates the pose and scale variations of images captured by mobile devices hence is able to represent mobile captured better. Note that multiple scale local features can also be extracted to cope with the scale variations. In this way, we are able to alleviate the problems mentioned above and generate proper local features for further processing.

After a local region is selected, a proper description scheme should be used to represent this local region. Various descriptors have been proposed which can be broadly divided into gradient based (e.g. scale-invariant feature transform (SIFT), histogram of gradients (HoG)) and statistics based (moment and its variances [53,54]). The gradient based methods are able to encode more information, hence are more discriminative than statistics based methods. We follow this successful setup and use the histogram of orientation gradients [31] to describe these adaptive chosen local regions. Note that other types of local region descriptors (e.g. moments) can also be used. However, directly computing the HoG descriptor for each adaptive local region is computational impossible because of the heavy computational

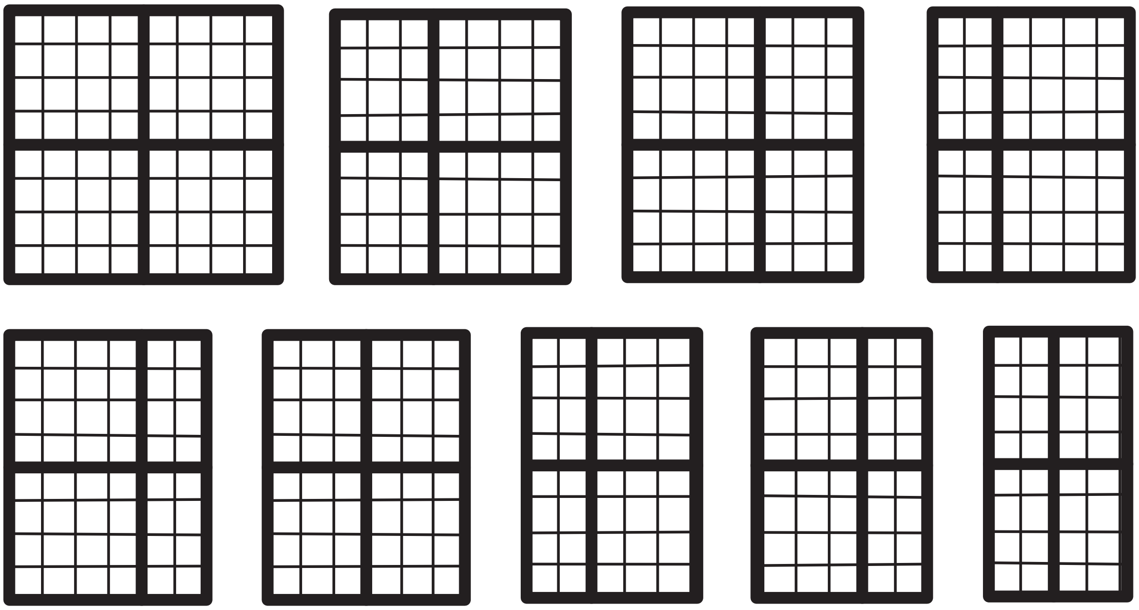


Fig. 3. A simplified adaptive local region selection with 2×2 sub-regions.

resource required. Since for each sub-region, we adaptively choose the number of pixels from 4 to 2 with one pixel interval, this means we need to extract 3 HoG features (44, 33 and 22 pixels respectively) for each sub-region. Besides, since each local region is divided into 4×4 sub-regions, and for each sub-region, 3 HoG features are calculated. Hence, this results in approximately $3^{16=4\times 4}$ HoG features' calculation to describe this local region. To speed up the computation, we choose to calculate the histogram of orientation gradients for each sub-region instead. We then concatenate the HoG of the 4×4 sub-regions together to form the final HoG descriptor for this local region. In this way, we are able to reduce the computational cost from approximately 3^{16} to 3^2 times instead. Besides, this sub-region decomposition also enable parallel computation which can be carried out on a distributed server very efficiently. We will give the computational complexity analysis in the Experiments Section.

3.2. Geo-codebook construction and discriminative local feature selection

The use of location information to assist visual recognition has been widely used by researchers [3,15,32,48]. Of the various location information used, the GPS information is the most widely used as more and more devices are embedded with GPS receivers. This helps to improve the performance and reduce the computational cost for location constrained visual applications (e.g. landmark recognition), especially when the image database is very large. Fig. 4 shows the GPS coordinate map of the San Francisco landmark dataset [48]. However, the GPS information may be inexact in urban environments because a clear view of at least four satellites is needed [55]. We also plot the captured locations of query images in Fig. 4. We can see from Fig. 4 that only use the GPS alone is not enough for reliable visual recognition [48]. It would be more effective if we combine the visual information with the GPS information for joint recognition.

Motivated by this, we first partition the database images with their GPS information using k-means clustering algorithm. Each cluster has images captured within this cluster, hence are more likely to be related. In this paper, we use a relatively large K. Note that different may correspond to the same image cluster. We learn the codebook for each cluster separately. In this way, we are able to combine the location information with visual codebook to construct a geo-codebook. However, the hard partition of k-means clustering is not able to cope with images that lie on the boundary of the partitions very well. To alleviate this problem, we first expand the size of each cluster to 1.2 times of its original size. Images that lie within this expanded area are then used for codebook construction. We use the sparse coding scheme for the geo-codebook construction because of its good performance [37].

Formally, for images of each cluster, let $X=[x_1,x_2,\dots,x_N]\in\mathbb{R}^{D\times N}$ be the set of local image features where N is the number of local features. $x_i\in\mathbb{R}^{D\times 1}, i=1,2,\dots,N$, D is the dimension of local features and M is the codebook size. $B\in\mathbb{R}^{M\times D}$ is the codebook to be learned. The sparse coding tries to solve the following optimization as

$$\min_{\alpha_i,B}\|x_i^T-\alpha_i B\|^2+\lambda\|\alpha_i\|_1 \tag{1}$$

where α_i is the sparse coding parameters for local feature $x_i, i=1,2,\dots,N$, λ is the parameter which controls the sparsity of α_i. This problem can be solved by alternatively searching for the optimal codebook B or coding parameter α_i while keeping coding parameter α_i or codebook B fixed. When we fix the codebook B and learn the optimal α_i, problem 1 can be rewrote as

$$\min_{\alpha_i}\|x_i^T-\alpha_i B\|^2+\lambda\|\alpha_i\|_1 \tag{2}$$

This can be solved using the feature-sign-search algorithm [36] or the Lasso algorithm. We use the feature-sign-search algorithm [36] for computational time reason. When the coding parameter α_i is fixed, problem 1 can be rewrote as

$$\min_B\|x_i^T-\alpha_i B\|^2 \tag{3}$$

This can be optimized using the Lagrange dual [49]. Moreover, locality constraint [56] with 5 neighbors is also used to help to improve the performance of sparse coding.

After the geo-codebook was learnt, we can encode the adaptive region based local features with the codebook B fixed. For landmark images, the geo-codebook is very effective. For general recognition task, the location information is less important. However, we believe that with the increase of mobile captured images, the geo-discriminative codebook can also help to recognize general images (including but limited to landmark images) by increasingly training the codebook with these images. The direct use of all the extracted HoG features with adaptive region is computational expensive and

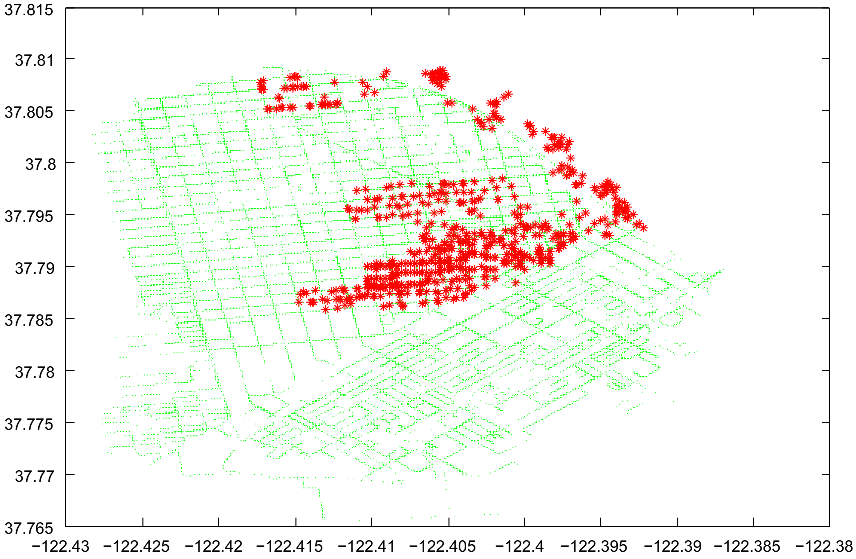


Fig. 4. The GPS coordinate map of the San Francisco landmark dataset [48] (green) and the captured location of query images (red). The horizontal axis and the vertical axis represent the longitude and latitude coordinates respectively. The negative value of the horizontal axis means the west longitude while the positive value of the vertical axis represents the north latitude. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

unnecessary. This is because we extract multiple HoG features by varying the pixel number of each sub-region to cope with the varied pose and illumination changes. There is information redundancy with there extracted features. Besides, for one particular image, the pose variation and illumination changes is fixed after the image is taken. It is necessary to choose the most discriminative local feature with adaptive region for better image representation. Moreover, local feature selection also helps to get ride of some noise local features and boost the robustness of the proposed algorithm.

Motivated by this, we propose to choose the discriminative local feature by minimizing the reconstruction error of sparse coding. This is because minimizing the reconstruction error helps to preserve more information that is lost during the local feature encoding process. Many works [14,37,40,56,57] have shown the effectiveness of this strategy. Formally, let P be the number of adaptive local features, $x_p, p = 1, 2, \ldots, P$, α_p are the corresponding sparse coding parameters. The reconstruction error e_p is defined as $e_p = \|x_p - \alpha_p B\|^2$. The discriminative sparse coding parameter for each local region is then chosen as

$$\alpha_p = \underset{\alpha_p}{\operatorname{argmin}} \|x_p^T - \alpha_p B\|^2 \tag{4}$$

After choosing the discriminative adaptive local feature for each local region, we can represent images with the coding parameters. We use the max pooling strategy proposed by Yang et al. [37] as this is very effective with the sparse coding scheme and is biologically reasonable [58]. Let $h \in \mathbb{R}^{M \times 1}$ be the final image representation, $\alpha_{p,l}$ be the l-th element of α_p. The l-th element of h can then be calculated as

$$h_l = \max(\alpha_{1,l}, \alpha_{2,l}, \ldots, \alpha_{Q,l}) \tag{5}$$

where Q is the number of local regions. To combine the spatial information of local features, the spatial pyramid matching (SPM) with three pyramids (1×1, 2×2, 4×4) is also used to boost the final performance.

3.3. Social relationship modeling for landmark recognition

We can do landmark recognition using the max pooled image representation by measuring the similarities between query image and each database image directly. If the most similar image and the query image belong to the same class, it can be viewed as a successful recognition. The recognition fails if the two images are of different classes. However, mobile users are inherently connected with various social relationships. The modeling of social relationships has been proved very effective and efficient for various applications [22,26,44]. For a particular mobile user, his/her friends may have searched similar images. It would be more effective if we can combine the social relationship with visual features for better visual recognition.

Let h denote the representation of one user captured image for visual recognition, J is the number of this user's friends. The j-th friend queried with image whose representation is h^{f_j}. Our aim is to find the matching image with the query image by not only comparing the visual similarity but also considering the friendship information. Fig. 5 shows the flowchart of social relationship modeling for visual recognition algorithm. Let h_r denote the r-th dataset image, $r = 1, 2, \ldots, R$, R is the total number of images in the dataset. This can be achieved by minimizing the following objective function as

$$r = \underset{r}{\operatorname{argmin}}\ \Phi(h, h_r) + \Psi(h, h_r, h^{f_j}) \tag{6}$$

The first term considers the visual distance of query image and images in the dataset. Usually, the visual distance is calculated as

$$\Phi(h, h_r) = \|h - h_r\|^2 \tag{7}$$

While the second term considers the recognition results of the user's friends. h^{f_j} denotes the image representation of the image queried by the j-th friend. This can be measured by

$$\Psi(h, h_r, h^{f_j}) = \sum_{j=1}^{J} w_j \|h^{f_j} - h_r\|^2 \tag{8}$$

where w_j is the weighting parameter which controls the relatedness between the query image h and h^{f_j}. This can be calculated as

$$w_j = \exp^{-\|h - h^{f_j}\|^2/\sigma^2} \tag{9}$$

where σ^2 is the scaling parameter. Usually, the median value of $\|h - h^{f_j}\|^2, j = 1, 2, \ldots, J$, is used as the scaling parameter. Note that we only consider the case that one user only searches with one query image. However, the extension to multiple queries is straightforward by replacing Eq. (8) with

$$\Psi(h, h_r, h^{f_j}) = \sum_{j=1}^{J} \sum_{q=1}^{Q_j} w_{j,q} \|h^{f_{j,q}} - h_r\|^2 \tag{10}$$

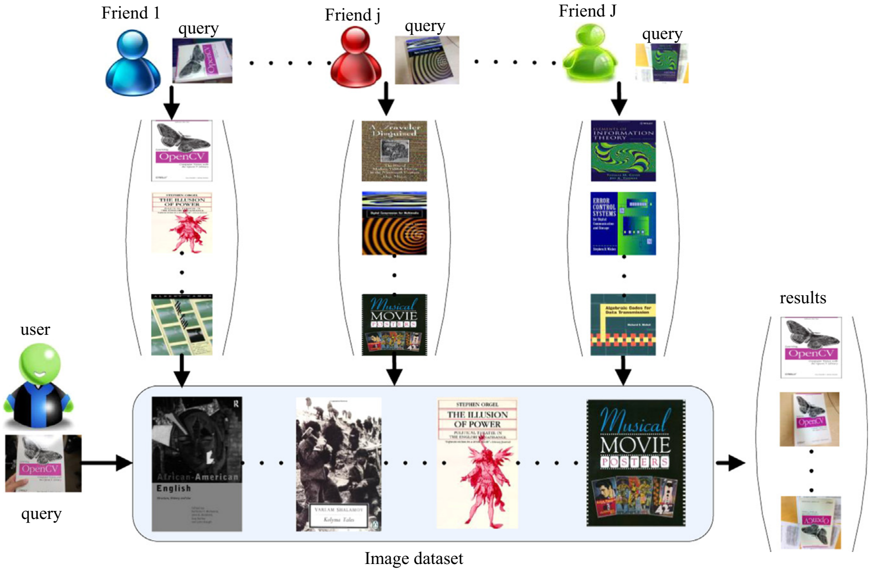


Fig. 5. Flowchart of the social relationship modeling for visual recognition algorithm.

where Q_j is the number of query images ($h^{f_{j,q}}$) the user j submitted and the weighting parameter $w_{j,q}$ can be defined as Eq. (9) accordingly. For location based visual recognition, such as landmark recognition, the GPS information can be used together with this social relationship modeling scheme to speed up the computation and improve the recognition performances.

4. Experiments

To evaluate the performance of the proposed method, we conduct experiments on two public datasets: the Stanford mobile visual search dataset [1] and the San Francisco landmark dataset [48]. All the multi-scale extracted local features are normalized with L_2 norm.

4.1. Stanford mobile visual search dataset

The Stanford mobile visual search dataset (SMVS) [1] contains 3300 query images for 1200 distinct classes of eight categories (*books, business cards, cd covers, dvd covers, landmarks, museum paintings, print* and *video frames*). These query images are captured with different camera phones as well as digital cameras (Apple: iPone 4, Palm: Pre, Nokia: N95, N97, N900, E 63, N5800 and N86, Motorola: Droid, Canon: G11 and LG: LG300) under widely varying lighting conditions over several days to several months. We view all the mobile device users as friends to generate the social relationship. Reference images with varied resolutions are also given. We resize the images to 1200 × 800 pixels to cope with the resolution caused by the use of different mobile devices. For product images (*CD, DVD* and *books*), the reference images are obtained from product websites. For landmarks, [1] used the data collected by Navteq's vehicle-mounted cameras. The reference images for video clips are extracted with key frames. For text prints, reference images are collected from [56] and the Internet. A high quality scan scheme is used to get the reference images for business cards. For museum paintings, the data are collected from the Cantor Art Center at Stanford University. Compared with other public datasets, the images of the SMVS dataset have rigid objects with varying lighting conditions, perspective distortion, realistic reference data and query images captured from heterogeneous camera phones. Fig. 6 shows some example images of the Stanford mobile visual search dataset. For the landmarks class, we use the GPS information provided by [13,48] as the location information. For the other seven classes, we view images of the same objects as belonging to the same location. The images captured with mobile phones are used as query images and the reference images are regarded as the database images to be matched. We try to find the images of the same class from the database and use the successful recognition percentage of all query images for performance evaluation. Spatial pyramid matching (SPM) [36] with three pyramids ($2^L \times 2^L, L = 0, 1, 2$) is used to combine the spatial information of local features. This technique gradually divides each image into finer image blocks (1 × 1, 2 × 2, 4 × 4). For each image block, we empirically use 1024 dimensional vector for representation. This 1024 dimensional vector for each sub-image block is concatenated to form a 21 × 1024 dimensional vector as the final image representation.

To evaluate the effectiveness of the proposed algorithm (Abbr. ALF+location+SR), we compare its performance with several other methods [13,59,60]: the fast Hessian interest point detector along with Compressed Histogram of Gradients descriptor [13], the Hessian-affine interest point detector and SIFT feature [59], Difference-of-Gaussian interest point detector and the SIFT feature [60]. To show the effectiveness of location information and social relationship for visual recognition, we also give the per-class performances of using adaptive local features with discriminative selection (Abbr. ALF) and adaptive local features with location information (Abbr. ALF+Location). We use the percentage of images that match as the quantitative performance measurement. Table 1 shows the per class performances of these methods.

We can see from Table 1 that the proposed method outperforms other methods [13,59,60] for all the eight classes which proves the effectiveness of the proposed method. We can have some conclusions from Table 1. First, the ALF outperforms [60] by 3.1 percent. This is because the use of adaptive local regions for local feature extraction helps to cope with the large pose variation and illumination changes of mobile device captured images. Besides, the selection of discriminative local features also helps to choose the suitable local feature to describe each local region, hence helps to get better image representation which eventually improves the recognition performances. Second, the use of location information helps to boost the performance by 0.5 percent over ALF, especially for the landmark class. This class is relatively more difficult than the other seven classes because the query images are captured several months after the database is gathered. Only using visual features is not able to cope with the appearance changes with time very well while the use of location information helps to alleviate this problem. Third, the combination of social relationship for visual recognition can further improve the performance by 0.8 percent over ALF. One query image is often unable to represent the objects to be matched well. However, the combination of different queries captured by multiple users helps to alleviate this problem by collaboratively matching the most suitable images. Moreover, the performances of interest point based local region selection methods [13,59,60] are relatively unstable compared with dense selection. This is because interest point detectors relay on a selection scheme which may not robust enough for mobile visual applications, especially for mobile images with clustered background and heavy occlusions. On analysis of the per-class performance, we can have similar conclusions as [1] that the outdoor classes (e.g. landmark) are relatively more difficult than indoor classes (e.g. books, DVDs). This is because indoor images are captured under relatively controlled conditions while outdoor images are not.

4.2. San Francisco landmark dataset

The San Francisco landmark dataset [48] is collected by first using a mobile mapping vehicle composes of 360° sensor (Velodyne HDL-64E), panoramic camera (Ladybug 3), high-definition cameras (Prosilica), Global Positioning System (GPS), Inertial Measurement Unit (IMU) and Distance Measurement Instrument (DMI). The Velodyne LIDAR sensor has 64 lasers mounted on upper and lower blocks of 32 lasers each with the Ladybug 3 covers about 80 percent of a full sphere. Six high quality 1600 × 1200 Sony CCD sensors capturing 8MP panoramas at 15 frames/second. The GPS and IMU are used to geo-reference these sensors. A dataset for San Francisco which has about 150k panoramic images of 4-meter intervals is then converted to approximately 1.7 million perspective images (1.06 million perspective central images (PCIs) and 638k perspective frontal images (PFIs)). To alleviate the shadowing problem, histogram equalization is found by [43] as a very effective way to improve the recognition accuracy. We resize the images to 640 × 480 pixels. For the query images, most users take photos in portrait and landscape mode. These images are then gravity-aligned using mobile devices' motion sensor readings. In total, a set of 803 query photos of landmarks in San Francisco are taken with several different camera phones by several people after a few months of the dataset image collection. The query images are cluttered with vehicles, pedestrians and seasonal vegetation. The photometric and geometric distortions are often very large compared with other well-designed datasets. The GPS coordinates are also provided for geo-related applications. 596 query images have actual GPS information while the rest 207 query images' GPS information is obtained by a Gaussian model. To construct the

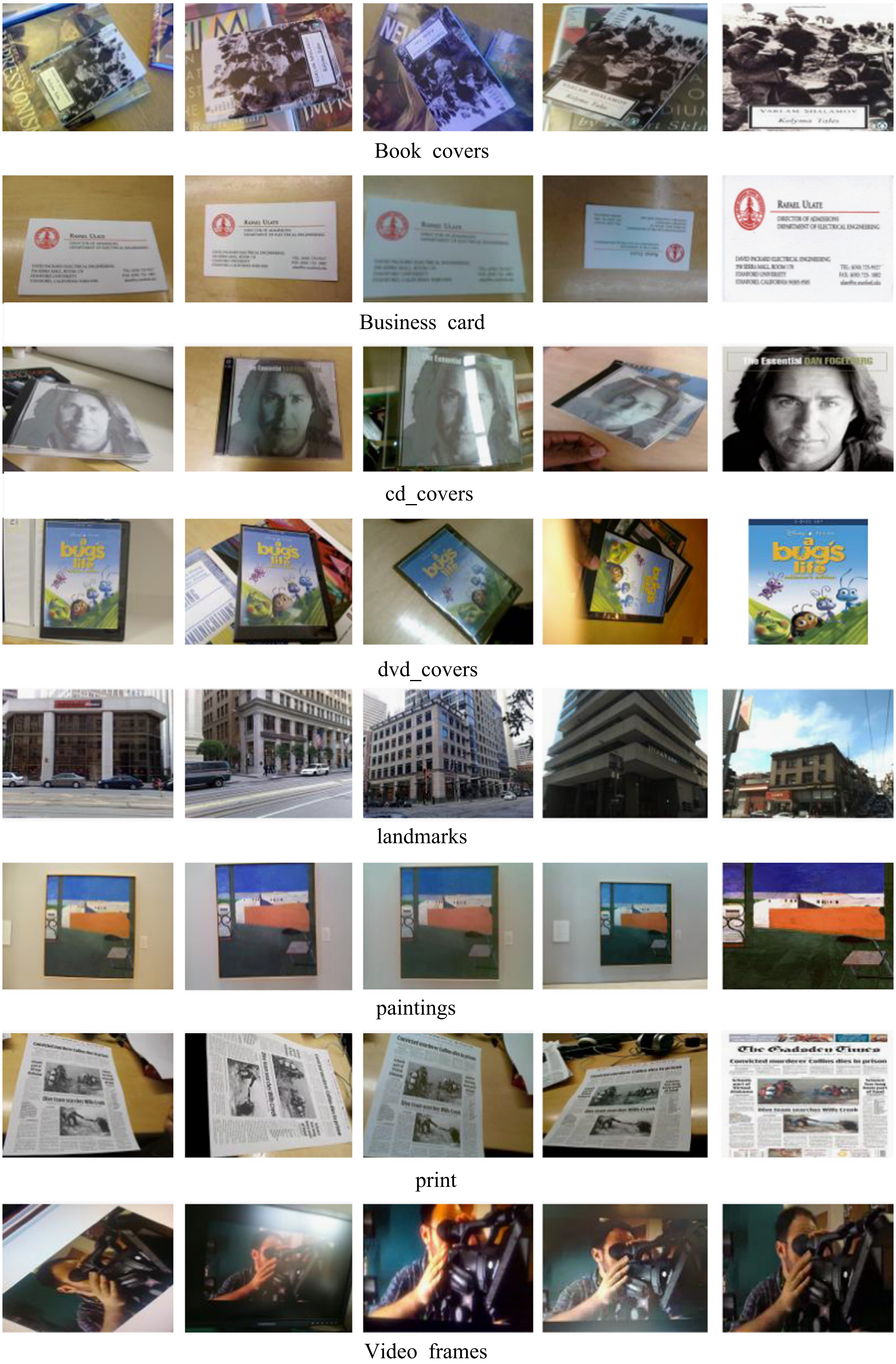


Fig. 6. Example images of the Stanford mobile visual search dataset.

social relationship network, we follow the setup as on the SMVS dataset and view the each mobile device user as friends. Since the 803 query images are taken by several people with various camera phones, we view the people with query images taken on the same location as friends. This is achieved by clustering the GPS locations of query images and views each cluster as one location. The friend

Table 1
Image match accuracy of different methods and the proposed adaptive region based local feature with social relationship modeling method for each category on the Stanford mobile visual search dataset. FH+CHoG [13]: fast Hessian detector with compact Histogram of Gradients descriptor; HA+SIFT [59]: Hessian Affine detector with SIFT descriptor; DoG+SIFT [60]: Difference-of-Gaussian detector with the SIFT descriptor; ALF: adaptive local feature; ALF+Location: adaptive local feature with location information; ALF+Location+SR: adaptive local feature with location information and social relationship.

Class	[13]	[59]	[60]	ALF	ALF+Location	ALF+Location+SR
DVDs	90.0	99.3	99.8	99.9	**100**	**100**
Videos	96.8	99.5	98.4	99.6	**99.8**	**99.8**
Books	97.2	96.0	98.2	98.9	99.3	**99.5**
CDs	84.1	94.9	93.6	95.5	96.1	**96.2**
Prints	69.9	70.5	89.5	91.7	92.2	**92.5**
Cards	72.5	83.7	87.8	89.4	89.5	**89.9**
Art	79.8	63.5	68.4	83.3	83.3	**83.6**
Landmarks	49.3	68.1	68.7	70.9	73.4	**74.3**
Average	79.9	84.4	88.1	91.2	91.7	**92.0**

relationship can then be assigned accordingly. Fig. 7 shows some example images of the San Francisco landmark dataset. We set the codebook to 1 million, as [48] did. Hierarchal sparse coding is used to speed up computation. Note that this dataset is much more difficult to recognize than the Stanford mobile visual search dataset not only because this dataset is larger but also because of the large inter and intra class variations.

For fair comparison, we follow the same experimental setup as Chen et al. [48] did. We first match the PCI and PFI images separately to get an initial result and then merge the results to get the final match result. Since the GPS information with about 300 m is proven very efficient for landmark recognition, we also adopt this location information setup. However, only using the GPS information is not enough for efficient recognition because there are 230 different buildings on average within 300 m radius [48]. We choose the DoG interest point detector and SIFT local feature descriptor scheme [48] for comparison as it achieves good performance on the San Francisco landmark dataset. The number of correct buildings over the total 803 query images with the number of top N candidates is used to measure the performance of the proposed method.

Fig. 8 shows the performance comparison on the 1.06M perspective central images with [48]. We can see from Fig. 8 that the use of GPS information to limit the search area can substantially improve the performance. This is in accordance with [48]. Besides, the use of adaptively selected local features outperforms the simple SIFT feature based scheme. The buildings have few discriminative visual features and are also with many repetitive structures and simple projective or affine transformations that cannot easily captured their 3D geometries. The performance of ALF over DoG+SIFT shows the usefulness of adaptively selecting local regions and discriminatively choosing local features. Moreover, social relationship modeling also helps to improve the precision over top N results. This is because one query image can hardly character the target landmark very well while the joint consideration of friends' queries of the same or nearby buildings can boost the recognition accuracy.

To show the roles of the proposed adaptive local features (ALF), the use of GPS information as well as the social relationships (SR) on the perspective central images of the San Francisco landmark dataset, we also give the performances of using adaptive local features, GPS and social relationships in Table 2 as well as the performances of combining the three roles with the top N (1, 4, 10, 25 and 40) retrieved images. The performances of using DoG+SIFT and DoG+SIFT+GPS [48] as well as the 3D method in [4] are also provided. Note that when only GPS or SR information is used, the HoG is used for local region representation. We do not use the adaptive local features in order to show the influences of using either GPS or SR information alone. We can see from Table 2 that the use of adaptive local features, GPS information and social relationship can help to improve the performances. Besides, the improvement of using the GPS information is larger than using the adaptive local features or social relationships alone. This is because the location of landmarks plays an important role for recognition which is also proven by [48]. The method of [4] performs not as well as [48] and the proposed method. We believe this is because the 3D point clouds reconstruction needs a number of images (5000 images per landmark is used in [4]) of the same landmark to reconstruct the 3D information. This strategy works well for famous landmarks of which large quantity of images can be collected. Besides, the training images of [4] are cleaner than that of the San Francisco landmark dataset. Finally, we are able to achieve the best performance when we combine the adaptive local features, the GPS information and the social relationship into a unified framework.

Fig. 9 shows the performance comparison on the 638k perspective frontal images. We can have similar conclusions as on the perspective cental images. Combining GPS information as well as social relationship help to boost the recognition performances. By Figs. 8 and 9, we can see that the performance on PFIs is better than on PCIs. There are mainly two reasons: first, the number of PFIs is about half of PCIs. Generally, the performance decreases with the increase of the database size. Second, the PFIs are generated by shooting a ray through each PCI's central projection with the angle between the viewing direction and the normal direction at the intersection point is less than 45°. The generated PFIs are with reduced geometric distortions between query images and database images, hence are much easier to recognize than PCIs.

To show the roles of the proposed adaptive local features (ALF), the use of GPS information as well as the social relationships (SR) on the perspective frontal images of the San Francisco landmark dataset, we also give the performances of using adaptive local features, GPS and social relationships in Table 3 as well as the performances of combining the three roles with the top N (1, 4, 10, 25 and 40) retrieved images. The performances of using DoG+SIFT and DoG+SIFT+GPS [48] as well as [4] are also provided. We can have similar conclusions as from Table 2 that the use of adaptive local features, GPS information and social relationship can help to improve the performances. We are able to achieve the best performance by combining the adaptive local features, the GPS information and the social relationship together. Note that as the perspective frontal images of the San Francisco landmark dataset are generated from perspective central images and more cleaner and relatively easier, the performances in Table 3 are relatively better than the corresponding methods in Table 2.

To get the recognition results on the whole San Francisco landmark dataset, we linearly merge the matching results on the PFIs and PCIs. We adopt the same performance measurement method as on the PFIs and PCIs. Fig. 10 shows the performance comparison on the whole San Francisco landmark dataset. This hybrid strategy is able to outperform the performances either on the PFIs or PCIs alone. We can see from Fig. 10 that the combination of location information is particular effective for landmark recognition. Besides, the use of adaptive local feature as well as social relationship consistently improves the recognition performance. This again demonstrates the proposed method's effectiveness.

To give an intuitive illustration of the roles of the adaptive local features, the GPS information as well as the usage of social relationship, we give some examples in Fig. 11. The adaptive local feature can help to cope with view point changes to some extent, hence is able to distinguish the two images in Fig. 11(a). However, the visual appearances of the two images of (b) are very similar,

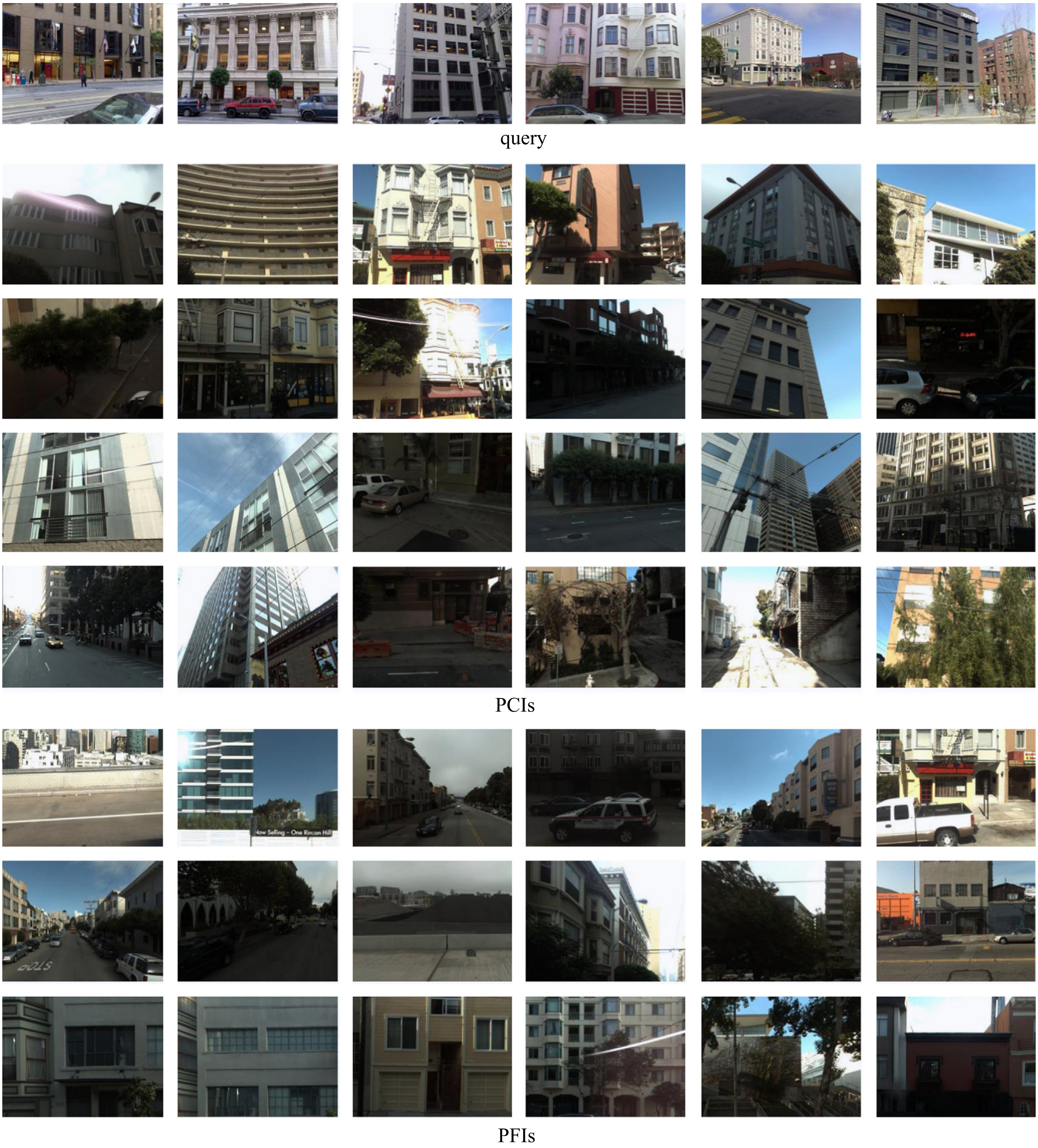


Fig. 7. Example images of the San Francisco landmark dataset.

only using the adaptive local feature is not enough. Fortunately, the GPS positions of the two images of (b) are dissimilar which helps to separate them apart. The GPS information is also able to cope with visual dissimilarities, such as (c) and (d). The two images are taken with different viewpoints by different people, it is very hard to recognize them well with visual features alone while the location information can help to achieve this goal. Although the adaptive local features and GPS information are very effective, they cannot solve the problems as in (e) and (f). Both (e) and (f) are images at the same location but with large viewpoint change, only using the visual feature and GPS information is not enough to recognize the corresponding landmarks well. However, the two images are taken by friends at the same location. By incorporating this social relationship into consideration, we can improve the recognition results.

4.3. Computational complexity analysis

Since we use the server to compute visual features, the main computation cost is the calculation and encoding of adaptive local features. As to the dataset images to be retrieved, their features

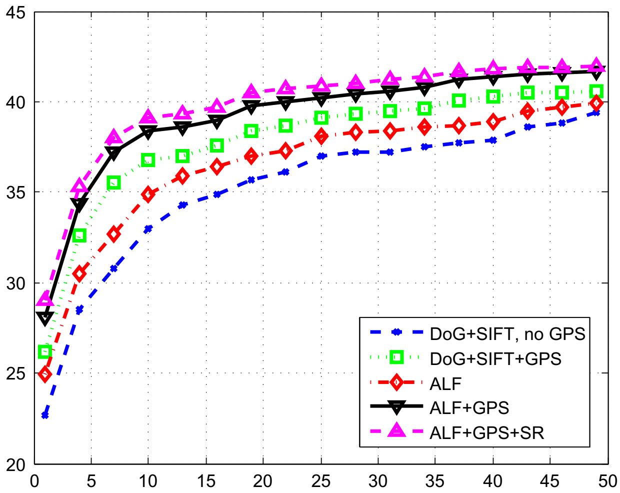


Fig. 8. Performance comparison on the 1.06M perspective central images. The horizontal axis represents the number of top *N* candidate images. The vertical axis represents the average classification rates (percent) of the query images. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 2
The performances (percent) of using adaptive local features, GPS information and social relationships individually on the perspective central images of the San Francisco landmark dataset with the top *N* (1, 4, 10, 25 and 40) retrieved images. The performance of combining adaptive local features, GPS information and social relationships is also provided along with the baseline methods of DoG+SIFT and DoG+SIFT+GPS [48] as well as [4].

Methods	Top 1	Top 4	Top 10	Top 25	Top 40
DoG+SIFT [48]	22.7	28.5	33.0	37.0	37.9
DoG+SIFT+GPS [48]	26.2	32.6	36.8	39.1	40.3
Xiao [4]	18.4	23.8	26.2	31.9	32.2
ALF	24.9	30.5	34.9	38.1	38.9
GPS	27.9	34.2	38.3	40.2	41.1
SR	23.8	29.4	33.7	37.5	38.6
ALF+GPS+SR	**29.0**	**35.3**	**39.1**	**40.9**	**41.8**

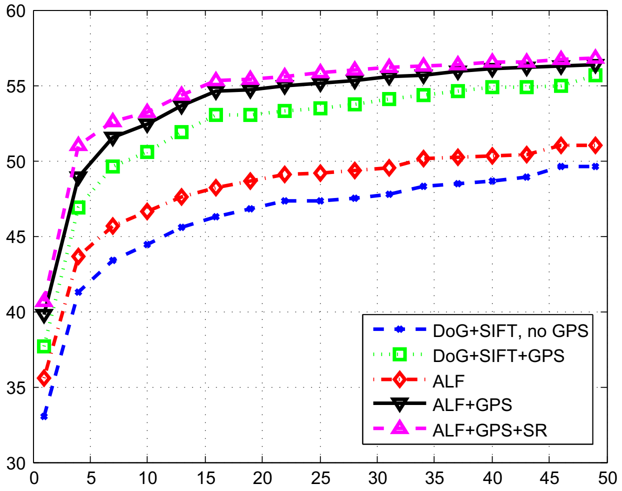


Fig. 9. Performance comparison on the 638k perspective frontal images. The horizontal axis represents the number of top *N* candidate images. The vertical axis represents the average classification rates (percent) of the query images. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

Table 3
The performances (percent) of using adaptive local features, GPS information and social relationships individually on the perspective frontal images of the San Francisco landmark dataset with the top *N* (1, 4, 10, 25 and 40) retrieved images. The performance of combining adaptive local features, GPS information and social relationships is also provided along with the baseline method of DoG+SIFT and DoG+SIFT+GPS [48] as well as the method of [4].

Methods	Top 1	Top 4	Top 10	Top 25	Top 40
DoG+SIFT [48]	33.0	41.3	44.4	47.3	48.6
DoG+SIFT+GPS [48]	37.7	46.9	50.6	53.5	54.9
Xiao [4]	28.4	36.5	38.9	41.2	42.4
ALF	35.6	43.6	46.6	49.1	50.3
GPS	39.3	48.5	51.7	54.5	55.6
SR	35.7	43.4	46.2	48.8	50.2
ALF+GPS+SR	**40.7**	**51.0**	**53.2**	**55.8**	**56.5**

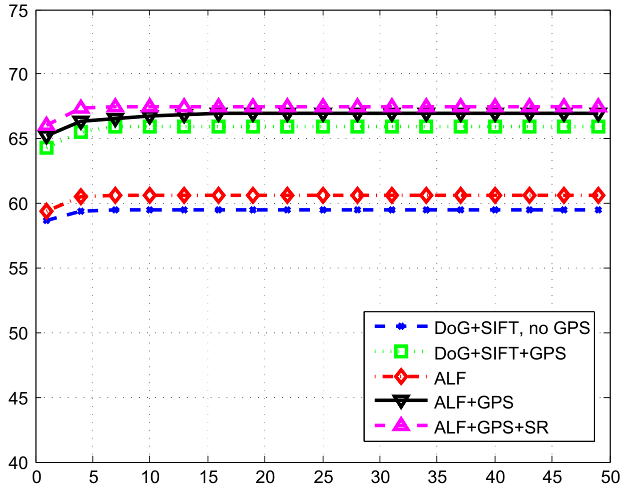


Fig. 10. Performance comparison on the whole San Francisco landmark dataset. The horizontal axis represents the number of top *N* candidate images. The vertical axis represents the average classification rates (percent) of the query images. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

can be pre-computed and stored on the servers. Hence, the computational concern mainly lies in the fast adaptive local feature extraction and encoding of user queries. By using the sub-region dividing technique proposed in Section 3.1, we can reduce the histogram of gradients (HoG) calculation to 3 × 3=9 times of standard HoG. Besides, since this sub-region calculation is independent of each other, they can be run in parallel using nine times memory than standard HoG extraction with better performance. The encoding of these extracted adaptive local features can also be implemented in a similar parallel way. This can be alleviated with the rapid development of fast storage devices (e.g. solid state disk, SSD). Moreover, the memory usage of this adaptive local feature extraction can be further reduced by first storing the gradients of the 4 × 4 pixels sub-region. The HoGs of the 3 × 3 pixels and 2 × 2 pixels within this sub-region can then be computed by choosing the gradients at the corresponding pixels.

Besides the computational efficiency of the proposed method, since the query images are uploaded to the server, this can also enable the processing of query images afterwards which can be used to improve the performance of the system with the increase of queries. Moreover, the computational power of the server is still relatively faster than mobile devices whose computational power also increases rapidly in recently. The use of the server enables us to take advantages of more computational power to achieve better performance. Besides, if we combine the indexing and approximate nearest neighbor search techniques [42,43], we can extend the approach to the large scale case. We will try to combine these techniques in our future work.

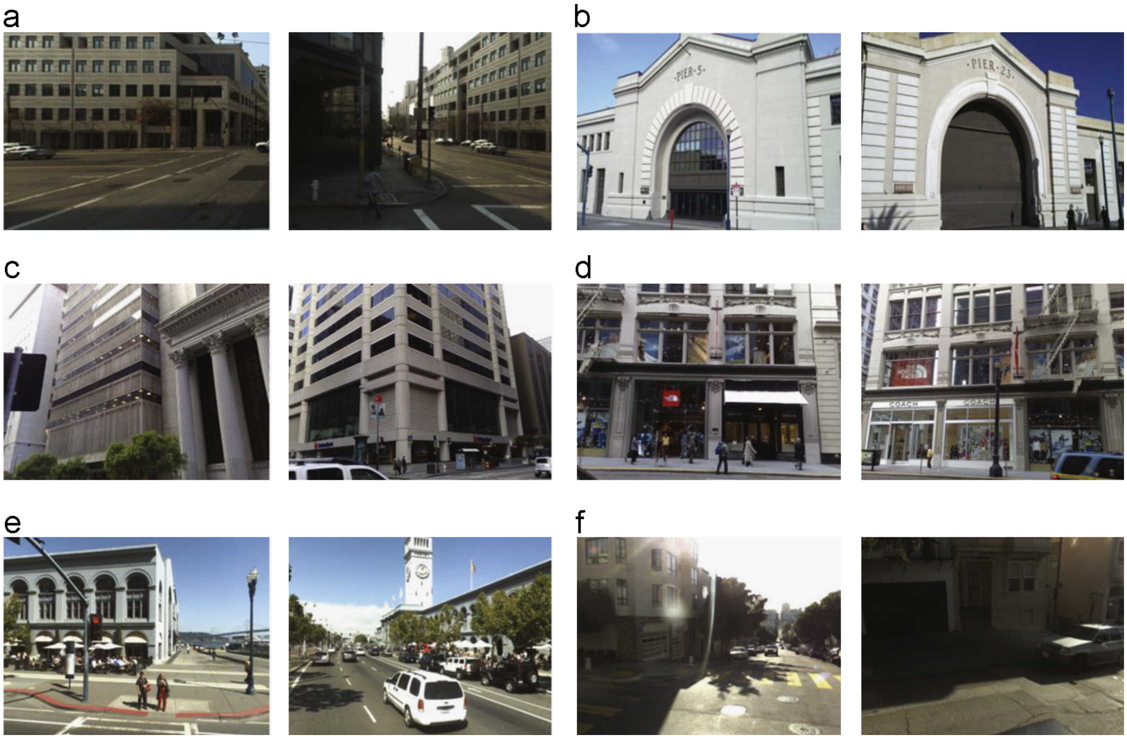


Fig. 11. Illustrative examples of the usage of adaptive local features, the GPS information and the social relationship. Adaptive visual feature can help recognize (a) while the GPS information plays an important role for (b), (c) and (d). The social relationship can help to improve the performance by incorporating the information of images with large variation at the same location. It is best viewed in color. (For interpretation of the references to color in this figure caption, the reader is referred to the web version of this paper.)

5. Conclusion

This paper proposes a novel socio-mobile landmark recognition method using local features with adaptive region selection. We adopt the popular client/server architecture and do the local feature extraction and landmark recognition on the server side. In order to get reliable recognition performance, we adopt the dense local feature extraction strategy and use the pixel gradients to represent each local region. In order to cope with varied poses and camera orientations of mobile images, we propose to adaptively choose the proper size of each subregion. This is achieved by minimizing the encoding loss of the extracted local features. Besides, a geo-discriminative codebook is generated to take advantages of images' location information. Moreover, we jointly consider the visual distances as well as user friends' matching results to further boost the final performance. Experimental results on two public datasets prove the effectiveness and efficiency of the proposed socio-mobile visual recognition method using adaptive region selection based local features.

Our future work consists of three aspects. First, how to speed up the computation of the adaptive local features will be studied. Second, we will explore new strategies [61,62] which can make use of the social relationship more effectively. Third, the design of more discriminative and representative features will also be considered.

Acknowledgment

This work is supported in part by National Basic Research Program of China (973 Program): 2012CB316400. This work is supported by National Natural Science Foundation of China: 61303154, 61202325, 61402023, 61379100, 61133003 and 61332016.

References

[1] V. Chandrasekhar, D. Chen, S. Tsai, N. Cheung, H. Chen, G. Takacs, Y. Reznik, R. Vedantham, R. Grzeszczuk, J. Bach, B. Girod, The Stanford mobile visual search data set, in: ACM Multimedia Systems Conference, 2011.
[2] J. He, J. Feng, X. Liu, T. Cheng, T. Lin, H. Chung, S.F. Chang, Mobile product search with bas of hash bits and boundary reranking, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2012.
[3] 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 (3) (2012) 290–314.
[4] X. Xiao, C. Xu, J. Wang, M. Xu, Enhanced 3-D modeling for landmark image classification, IEEE Trans. Multimedia 14 (4) (2012) 1246–1258.
[5] A. Bimbo, F. Dini, G. Lisanti, F. Pernici, Exploiting distinctive visual landmark maps in pan-tilt-zoom camera networks, Comput. Vis. Image Underst. 114 (6) (2010) 611–623.
[6] Google, Google Goggles ⟨http://www.google.com/mobile/goggles/⟩, 2009.
[7] Nokia, Nokia Point and Find ⟨http://www.pointandfind.nokia.com⟩, 2006.
[8] Amazon, SnapTell ⟨http://www.smaptell.com⟩, 2007.
[9] Kooaba, Kooaba ⟨http://www.kooaba.com⟩, 2007.
[10] D. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.
[11] K. Mikolajczyk, C. Schmid, Performance evaluation of local descriptors, IEEE Trans. Pattern Anal. Mach. Intell. 27 (10) (2005) 1615–1630.
[12] H. Bay, T. Tuytelaars, L. Gool, SURF: speeded up robust features, in: European Conference on Computer Vision, 2006, pp. 404–417.
[13] V. Chandrasekhar, G. Takacs, D. Chen, S. Tsai, R. Grzeszczuk, B. Girod, CHoG: compressed histogram of gradients, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2009, pp. 2504–2511.
[14] Y. Jiang, J. Wang, S.F. Chang, Lost in binarization: query-adaptive ranking for similar image search with compact codes, in: ACM International Conference on Multimedia Retrieval, 2011.
[15] H. Liu, T. Mei, J. Luo, H. Li, S. Li, Finding perfect rendezvous on the go: accurate mobile visual localization and its applications to routing, in: ACM Multimedia Conference, 2012, pp. 9–18.
[16] ⟨http://www.flickr.com⟩.
[17] ⟨http://www.youtube.com⟩.
[18] ⟨https://twitter.com⟩.
[19] Y. Lin, K. Candan, H. Sundaram, L. Xie, SCENT: scalable compressed monitoring of evolving multirelational social networks, ACM Trans. Multimedia Comput. Commun. Appl. 7S (1) (2011).

[20] P. Rota, N. Conci, N. Sebe, Real time detection of social interactions in surveillance video, in: ECCV Workshops (3), 2012, pp. 111–120.

[21] Y. Yang, F. Wu, F. Nie, H. Shen, Y. Zhuang, A. Hauptmann, Web & personal image annotation by mining label correlation with relaxed visual graph embedding, IEEE Trans. Image Process. 21 (3) (2012) 1339–1351.

[22] X. Li, C. Snoek, M. Worring, Learning social tag relevance by neighbor voting, IEEE Trans. Multimedia 11 (7) (2009) 1310–1322.

[23] S. Bao, S. Xu, L. Zhang, R. Yan, D. Han, Z. Su, Y. Yu, Mining social emotions from affective text, IEEE Trans. Knowl. Data Eng. 24 (9) (2012) 1658–1670.

[24] D. Liu, G. Ye, C. Chen, S. Yan, S.-F. Chang, Hybrid social media network, in: ACM Multimedia Conference, 2012.

[25] L. Duan, D. Xu, S.-F. Chang, Exploiting web images for event recognition in consumer videos: a multiple source domain adaptation approach, in: IEEE International Conference on Computer Vision and Pattern Recognition, 2012.

[26] M. Aly, A. Hatch, V. Josifovski, V. Narayanan, Web-scale user modeling for targeting, in: International Conference Companion on World Wide Web, 2012, pp. 3–12.

[27] G. Singh, N. Parikh, N. Sundaresan, Rewriting null e-commerce queries to recommend products, in: International Conference Companion on World Wide Web, 2012, pp. 73–83.

[28] J. Gall, N. Razavi, Luc Van Gool, On-line adaption of class-specific codebooks for instance tracking, in: British Machine Vision Conference, 2010.

[29] C. Eitzinger, M. Gmainer, W. Heidl, E. Lughofer, Increasing classification performance with adaptive features, in: Proceedings of ICVS, 2008, pp. 445–453.

[30] C. Eitzinger, S. Thumfart, Optimizing Feature Calculation in Adaptive Machine Vision Systems, Learning in Non-Stationary Environments: Methods and Applications, Springer, New York (2012) 349–374.

[31] N. Dalal, B. Triggs, Histograms of oriented gradients for human detection, in IEEE International Conference on Computer Vision and Pattern Recognition, 2005, pp. 886–893.

[32] A. Andoni, M. Datar, N. Immorlica, P. Indyk, V. Mirrokni, Locality-sensitive hashing Scheme based on p-Stable Distributions, Nearest Neighbor Methods in Learning and Vision: Theory and Practice, MIT Press, London, UK, 2006.

[33] A. Andoni, P. Indyk, Near-optimal hashing algorithms for near neighbor problem in high dimensions, in: Proceedings of the Symposium on Foundations of Computer Science, 2006.

[34] L. Pauleve, H. Jegou, L. Amsaleg, Locality sensitive hashing: a comparison of hash function types and querying mechanisms, Pattern Recognit. Lett. 31 (11) (2010) 1348–1358.

[35] K. Mikolajczyk, C. Schmid, Scale and affine invariant interest point detectors, Int. J. Comput. Vis. 60 (1) (2004) 63–86.

[36] S. Lazebnik, C. Schmid, J. Ponce, Beyond bags of features: spatial pyramid matching for recognizing natural scene categories, in: Proceedings of Computer Vision and Pattern Recognition, 2006, pp. 2169–2178.

[37] J. Yang, K. Yu, Y. Gong, T. Huang, Linear spatial pyramid matching using sparse coding for image classification, in: Proceedings of Computer Vision and Pattern Recognition, 2009, pp. 1794–1801.

[38] K. Sande, T. Gevers, C. Snoek, Evaluating color descriptors for object and scene recognition, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1582–1596.

[39] J. Morel, G. Yu, ASIFT: a new framework for fully affine invariant image comparison, SIAM J. Imaging Sci. 2 (2) (2009) 438–469.

[40] N. Kulkarni, B. Li, Discriminative affine sparse codes for image classification, in: Proceedings of Computer Vision and Pattern Recognition, 2011, pp. 1609–1616.

[41] H. Xu, J. Wang, X. Hua, S. Li, Tag refinement by regularized LDA, in: ACM Multimedia, 2009, pp. 573–576.

[42] J. Wang, J. Wang, X. Hua, S. Li, Scalable similar image search by joint indices, in: ACM Multimedia, 2012, pp. 1325–1326.

[43] J. Wang, S. Li, Query-driven iterated neighborhood graph search for large scale indexing, in: ACM Multimedia, 2012, pp. 179–188.

[44] J. McAuley, J. Leskovec, Image labeling on a network: using social-network metadata for image classification, in: Proceedings of European Conference on Computer Vision, 2012, pp. 828–841.

[45] L. Terveen, W. Hill, Beyond Recommender Systems: Helping People Help Each Other, HCI in the New Millennium, Addison-Wesley, Indianapolis, Indiana, 2001.

[46] Z. Zha, M. Wang, Y. Zheng, Y. Yang, R. Hong, T. Chua, Interactive video indexing with statistical active learning, IEEE Trans. Multimedia 14 (1) (2012) 17–27.

[47] Z. Zha, L. Yang, T. Mei, M. Wang, Z. Wang, Visual query suggestion, in: ACM Multimedia, 2009, pp. 15–24.

[48] D. Chen, G. Baatz, K. Koser, S. Tsai, J. Bach, M. Polleyfeys, B. Girod, R. Grzeszczuk, City-scale landmark identification on mobile devices, in: Proceedings of Computer Vision and Pattern Recognition, 2011.

[49] H. Lee, A. Battle, R. Raina, A. Ng, Efficient sparse coding algorithms, in: Advances in Neural Information Processing Systems, 2007, pp. 801–808.

[50] 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: Proceedings of ACM International Conference on Multimedia Information Retrieval, 2008.

[51] K. Wang, B. Babenko, S. Belongie, End-to-end scene text recognition, in: Proceedings of ICCV, 2011.

[52] Z. Zha, H. Zhang, M. Wang, H. Luan, T. Chua, Detecting group activities with multi-camera context, IEEE Trans. Circuits Syst. Video Tech. 23 (5) (2013) 856–869.

[53] L. Gool, T. Moons, D. Ungureanu, Affine/photometric invariants for planar intensity patterns, in: European Conference on Computer Vision, 1996, pp. 642–651.

[54] G. Stockman, L. Shapiro, Computer Vision, Prentice Hall, Upper Saddle River, NJ, 2001.

[55] G. Schroth, R. Huitl, D. Chen, M. Abu-Alqumsan, A. Al-Nuaimi, E. Steinbach, Mobile visual location recognition, IEEE Signal Process. Mag. 28 (4) (2011) 77–89.

[56] J. Wang, J. Yang, K. Yu, F. Lv, T. Huang, Y. Gong, Locality-constrained linear coding for image classification, in: Proceedings of Computer Vision and Pattern Recognition, 2010.

[57] C. Zhang, J. Liu, Q. Tian, C. Xu, H. Lu, S. Ma, Image classification by non-negative sparse coding, low-rank and sparse decomposition, in: Proceedings of Computer Vision and Pattern Recognition, 2011, pp. 1673–1680.

[58] T. Serre, L. Wolf, T. Poggio, Object recognition with features inspired by visual cortex, in: Proceedings of CVPR, 2005.

[59] K. Mikolajczyk, Software for computing Hessian-affine interest points and SIFT descriptor, 2010.

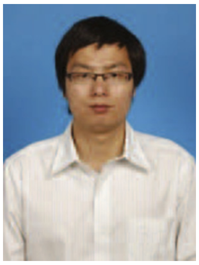
[60] A. Vedaldi, B. Fulkerson, VLFeat: an open and portable library of computer vision algorithms ⟨http://www.vlfeat.org⟩, 2008.

[61] Z. Zha, J. Yu, J. Tang, M. Wang, T. Chua, Product aspect ranking and its applications, IEEE Trans. Knowl. Data Eng. 26 (5) (2014) 1211–1224.

[62] J. Tang, Z. Zha, D. Tao, T. Chua, Semantic-gap-oriented active learning for multilabel image annotation, IEEE Trans. Image Process. 21 (4) (2012) 2354–2360.

Chunjie Zhang received his Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences, China, in 2011. He received his B.E. degree from Nanjing University of Posts and Telecommunications, Jiangsu, China, in 2006. He worked as an Engineer in the Henan Electric Power Research Institute during 2011–2012. He is currently working as a Postdoc at School of Computer and Control, University of Chinese Academy of Sciences, Beijing, China. Dr. Zhang's current research interests include image processing, machine learning, cross media content analysis, pattern recognition and computer vision.

Yifan Zhang received the B.E. degree in Automation from Southeast University in 2004, and the Ph.D. degree in Pattern Recognition and Intelligent Systems from Institute of Automation, Chinese Academy of Sciences in 2010. Then he joined National Laboratory of Pattern Recognition (NLPR), Institute of Automation, Chinese Academy of Sciences, Beijing, China, where he is currently an Assistant Professor. From 2011 to 2012, he was a Postdoctoral Research Fellow with the Department of Electrical, Computer, and Systems Engineering, Rensselaer Polytechnic Institute (RPI), Troy, New York. His research interests include probabilistic graphical models, activity recognition and video semantic analysis.



Xiaobin Zhu received his Ph.D. degree in Institute of Automation, Chinese Academy of Sciences, China in 2013. He received his M.E. degree in 2006 from Beijing Normal University. He is currently working at Beijing Technology and Business University. Dr. Zhu's current research interests include machine learning, video analysis and object tracking, etc.

Zhe Xue received the B.S. degree in Electronic Engineering from Civil Aviation University of China, Tianjin, China, in 2010. He is currently pursuing the Ph.D. degree in University of Chinese Academy of Sciences, Beijing, China. His research interests include image processing, topic detection and multimedia data mining.

Lei Qin received the B.S. and M.S. degrees in Mathematics from the Dalian University of Technology, Dalian, China, in 1999 and 2002, respectively, and the Ph.D. degree in Computer Science from the Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China, in 2008. He is currently an Associate Professor with the Key Laboratory of Intelligent Information Processing, Institute of Computing Technology, Chinese Academy of Sciences, Beijing, China. His research interests include image/video processing, computer vision, and pattern recognition. He has authored or coauthored over 40 technical papers in the area of computer vision. Dr. Qin is a reviewer for IEEE Transactions on Multimedia, IEEE Transactions on Circuits and Systems for Video Technology, and IEEE Transactions on Cybernetics. He has served as a TPC member for various conferences, including ECCV, ICPR, ICME, PSIVT, ICIMCS, PCM.

Qingming Huang received the Ph.D. degree in Computer Science from the Harbin Institute of Technology, Harbin, China, in 1994. He was a Postdoctoral Fellow with the National University of Singapore from 1995 to 1996 and was with the Institute for Infocomm Research, Singapore, as a Member of Research Staff from 1996 to 2002. He joined the Chinese Academy of Sciences, Beijing, China, under Science100 Talent Plan in 2003, and is currently a Professor with the Graduate University, Chinese Academy of Sciences. His current research areas are image and video analysis, video coding, pattern recognition, and computer vision.

Qi Tian received his Ph.D. degree in Electrical and Computer Engineering from University of Illinois at Urbana-Champaign, Illinois in 2002. He received his M.S. degree from Drexel University, Philadelphia, Pennsylvania, 1996, and B.E. degree from Tsinghua University, China, 1992, respectively. He is currently a Professor in the Department of Computer Science at the University of Texas at San Antonio, and Adjunct Professor in Zhejiang University and Xidian University. Dr. Tian's current research interests include Multimedia Information Retrieval, Computational Systems Biology, Biometrics, and Computer Vision. He is a Senior Member of IEEE, and a Member of ACM.