Semi-Supervised Learning for Cross-Device Visual Location Recognition

Pengcheng Liu*†, Peipei Yang†, Kaiqi Huang†, Tieniu Tan†, Hongwei Hao*
*Interactive Digital Media Technology Research Center
†Center for Research on Intelligent Perception and Computing, National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences
Beijing, China
Email:{pengcheng.liu, hongwei.hao}@ia.ac.cn, {ppyang, kqhuang, tnt}@nlpr.ia.ac.cn

***Abstract*—The aim of this work is to localize a query mobile photograph by utilizing surveillance images, which naturally provide location information. We cast this cross-device visual localization problem as a classification task. By exploiting the surveillance network to collect reference images, the data acquisition process is significantly facilitated. However, the discrepancy between mobile images and surveillance images makes the training samples difficult to be used directly, and the scarcity of training samples caused by the immobility of surveillance cameras further degrades the performance. In contrast to most traditional domain adaptation problems and semi-supervised problems, the scarce labeled data and plentiful unlabeled data exist in different domains. Our location recognition method first exploits the unsupervised subspace alignment to weaken the discrepancy between the two domains, and then adopts the semi-supervised Laplacian SVM to reinforce the discriminant information utilizing the unlabeled mobile images. Experimental results show that our location recognition method significantly outperforms other related methods.**

I. Introduction

With the prevalence of smart mobile phones, the mobile visual location recognition [1]–[6] has attracted more and more attention. They are particularly useful for indoor localization or the cases where GPS signal is unavailable.

The basic idea of mobile visual location recognition methods is to take a photo of a location and query the database of reference images. The result is given by the location corresponding to the best matched image. Therefore, the performance of these methods critically depends on the coverage of candidate locations, and a considerable number of labeled images are necessary for high accuracy. Since both taking photos and assigning labels (locations) for the photos are expensive and time-consuming, the methods are difficult to be applied in practice. Besides, to follow the variation of environment, the database has to be updated once in a while manually, which is also impractical.

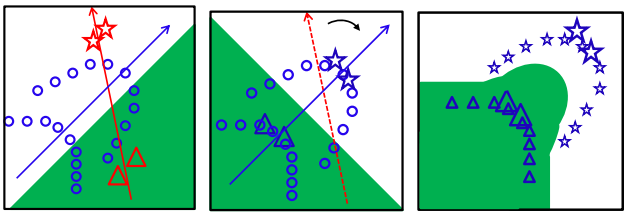
To solve the problems mentioned above, we consider exploiting the surveillance cameras to collect the reference images, which benefits from the following advantages. First, the surveillance cameras spread almost everywhere in the city and can naturally provide reference images with adequate coverage of the city without the time-consuming data acquisition process. Second, since the location of each surveillance camera is fixed and known, the surveillance network indeed constructs a complete geographic coordinate system. The location information can be obtained directly from the coordinate and we do not need to spend time assigning the labels. At last, since a surveillance camera works during the whole day, it can capture any variation of the environment and the database can always keep up to date.

Fig. 1. Illustration of our strategy for mobile visual location recognition. Digital numbers on the sketch are locations of surveillance cameras, the rose color covered area is the visible location of a surveillance camera. Some example images from two locations are shown under the sketch, which are captured by surveillance cameras (first row) and mobile phone cameras (second row) individually. Images inside one rounded rectangle are captured at the same location. Taking a query photo at one location by the mobile phone, if a classifier trained with labeled surveillance images can find out the right surveillance camera that covers this location, then based on the location of the surveillance camera, *we can know where we are*.

Accordingly, we propose a new mobile visual location recognition method utilizing the surveillance network in indoor scenes, as shown in Fig.1. After taking a query photo of the location with a mobile phone, the location is identified by finding out which surveillance camera covers this location. This location recognition process can be solved as a classification problem where the classifier is trained using the surveillance images labeled with locations. Then the location of the query mobile image is given by the classifier taking that image as a test sample. To the best of our knowledge, this is the first attempt to implement a location recognition by jointly utilizing these two mediums.

To realize the cross-device visual location recognition, there are two problems to solve since the performance is usually poor by directly training the classifier using surveillance images. On one hand, there exists obvious discrepancy between

1051-4651/14 $31.00 © 2014 IEEE
DOI 10.1109/ICPR.2014.495



(a) Original data (b) Domain adaptation (c) Laplacian SVM

Fig. 2. Illustration of our model to incorporate the domain adaptation and semi-supervised learning for cross-domain mobile visual location recognition. (a) The two class data captured at two locations by different device. The red data points are labeled surveillance images, the blue data points are unlabeled mobile images. (b) A better separation hyperplane is learned after the surveillance data has been aligned to the mobile data by domain adaption. (c) Our final location recognition result with a Laplacian SVM classifier utilizing the intrinsic structure of aligned unlabeled mobile data.

the data acquired by the two devices as shown in Fig.1, which will affect the classification accuracy. On the other hand, due to its immobility, few differences exist between the photos taken by one surveillance camera except a bit of illumination variation as shown in Fig.1, and thus the number of valid labeled training samples for each class is indeed very small. As a result, the classifier learned from only surveillance images is incapable to represent the variety of the mobile images, which seriously deteriorates the performance of location recognition as illustrated on Fig. 2a.

In the perspective of machine learning, the first problem caused by the differences between two devices is a domain adaptation problem where the surveillance images and mobile images constitute a *source domain* and *target domain* respectively, and the classifier learned in the source domain should be adapted to the target domain. The second problem caused by the scarcity of labeled samples is usually solved by the semi-supervised learning since there are a lot of unlabeled data available from user's queries. However, our problem obviously differs from the traditional domain adaption or semi-supervised learning problems in that the scarce labeled data and the plentiful unlabeled data are in different domains. In other words, the domain adaptation and semi-supervised learning are twisted in our problem, which makes the problem complex.

In this paper, instead of directly learning a classifier in the source domain, we first apply an unsupervised domain adaptation based on subspace alignment [7] to the data so that the labeled samples in the source domain are adapted to the target domain. By this process, the adapted labeled data and unlabeled data are aligned in the target domain as shown in Fig. 2b and the classification accuracy is improved. Then we adopt the semi-supervised Laplacian SVM [8] which naturally utilizes the unlabeled samples to learn a more accurate classifier by discovering the intrinsic structure of data, as shown in Fig. 2c. In this way, the variety of unlabeled samples in target domain is well incorporated into the classifier learned from the labeled samples in source domain.

The rest of the paper is organized as follows. Section II is devoted to the presentation of our location recognition method. Section III provides experimental details and shows the benefits of our method, and the paper is concluded in Section IV.

II. The Semi-supervised Cross-device Location Recognition

In this section, we present our cross-device location recognition method in detail. After the problem definition with some necessary notations, we elaborate the subspace alignment based domain adaptation approach for weakening the data discrepancy between two kinds of devices in II-B, and the semi-supervised Laplacian SVM classifier to overcome the scarcity of labeled training samples in II-C respectively. At last, the whole algorithm flow is presented in II-D.

A. Problem Definition and Notations

As explained above, since the surveillance images and mobile images constitute the *source domain* and *target domain* respectively, we use $\mathcal{X}^s = \{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ and $\mathcal{X}^t = \{\mathbf{x}_i^t\}_{i=1}^{n_t}$ to denote the sample sets of the source domain and target domain respectively, where y_i^s is the label of $\mathbf{x}_i^s$ indicating its location. Then the objective of location recognition is to learn a classifier using the surveillance images $\mathcal{X}^s$ and predict the label (location) y_i^t for each mobile image $\mathbf{x}_i^t$.

Using the notations above, the first problem of our location recognition is that the classifier learned using $\mathcal{X}^s$ cannot be directly applied to $\mathcal{X}^t$ due to the discrepancy between the distributions of samples. The second problem is that the training samples corresponding to the same label are all similar to each other, i.e., $\mathbf{x}_i^s \approx \mathbf{x}_j^s, \forall y_i^s = y_j^s$, while the test samples corresponding to the same label may be very different. This makes the training samples incapable to capture the variety of test samples. In the following parts, we exploit the domain adaptation method and semi-supervised method to solve the problems.

B. Subspace Alignment Based Domain Adaptation

For the same location, images collected from a surveillance camera are different from those taken by a mobile phone camera due to the differences of devices perspectives, illuminations, etc. This results in the discrepancy between source domain (surveillance camera) and target domain (mobile phone), and the classifier trained on $\mathcal{X}^s$ will likely fail to classify $\mathcal{X}^t$ correctly. In order to build a robust classifier, it is necessary to take into account the shift of distributions between the two domains, which is referred as domain adaptation (DA).

There already exist many studies on this topic in the field of computer vision, and subspace based DA has demonstrated a good performance in recent years. For example, Wang [9] aligns the two domains from two different manifolds so that they can be projected to a common subspace. In [10], data of the source domain are transformed into an intermediate representation by low-rank reconstruction technology, and then each transformed source sample can be linearly reconstructed by the target samples. In order to model the distribution shift process, both Gopalan [11] and Gong [12] explore the idea of using geodesic flows to derive intermediate subspaces that interpolate between the source and target domains.

Since there is no label information available in the target domain, we select an unsupervised subspace alignment based domain adaptation (SADA) method [7] in our method. Since $\mathcal{X}^s$ and $\mathcal{X}^t$ are captured at the same locations by different

devices, they should follow the related marginal distributions. In order to learn the transformation of distributions between these two domains rather than working on the original data directly, the SADA suggests a more robust representation of $\mathcal{X}^s$ and $\mathcal{X}^t$ in their principal subspaces defined by the projection matrices $S_s, S_t \in \mathbb{R}^{D\times d}$, whose columns are composed of the d principal components of $\mathcal{X}^s$ and $\mathcal{X}^t$ respectively.

The SADA assumes that the subspaces of the source domain and the target domain are related by a linear transformation. Because the strategy of other DA methods mentioned above is projecting the data to a common subspace, this may lead to information loss in both source and target domains [7]. The SADA method projects $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ to their respective subspaces by the operations $S_s^\top \mathbf{x}_i^s$ and $S_t^\top \mathbf{x}_i^t$. Then, a linear transformation matrix is learned to align the bases of the two subspaces. Denoting M as the transformation matrix from S_s to S_t, the subspace alignment is equivalent to solving the minimization problem

$$M^* = \arg\min_M \|S_s M - S_t\|_{\mathrm{F}}^2,$$

where $\|.\|_{\mathrm{F}}$ is the Frobenius norm.

By simple matrix algebra calculations, the optimal solution of M has a closed form $M^* = S_s^\top S_t$, which implies that the space of the new target-aligned source domain is spanned by the columns of $S_a = S_s M^* = S_s S_s^\top S_t$. Projecting the source data $\mathcal{X}^s$ using S_a into the target-aligned source subspace and the target data $\mathcal{X}^t$ into the target subspace using S_t, we can obtain the new representations $\tilde{\mathcal{X}}^s = \{(\tilde{\mathbf{x}}_i^s, y_i^s)\}_{i=1}^{n_s}$ where $\tilde{\mathbf{x}}_i^s = S_a^\top \mathbf{x}_i^s$ and $\tilde{\mathcal{X}}^t = \{\tilde{\mathbf{x}}_i^t\}_{j=1}^{n_t}$ where $\tilde{\mathbf{x}}_i^t = S_t^\top \mathbf{x}_i^t$ of the original data. Since $\tilde{\mathcal{X}}^s$ and $\tilde{\mathcal{X}}^t$ have been aligned in the common subspace, a classifier trained on the former is expected to have a better performance for the latter.

C. Semi-supervised Laplacian SVM for Classification

When the data of the two domains have been aligned, there still exists a problem that only a few valid training samples in $\tilde{\mathcal{X}}^s$ are available for each class (location). Thus even if $\tilde{\mathcal{X}}^s$ and $\tilde{\mathcal{X}}^t$ are aligned, the classifier learned using $\tilde{\mathcal{X}}^s$ still gives unsatisfactory recognition accuracy on $\tilde{\mathcal{X}}^t$ because the scarce training samples are incapable to reflect the true distribution. However, with the increasing number of visual location recognition users, there will be more and more unlabeled samples from mobile phone cameras. Although there is no label information, they can help to discover the true marginal distribution of $\tilde{\mathcal{X}}^t$. Thus, we consider to exploit these unlabeled images to enhance the accuracy of the classifier.

This idea is known as semi-supervised learning, which has attracted considerable attention in recent years. Some cluster assumption based methods mainly focus on looking for an optimal separation boundary that lies in the low density region of the data space, such as the TSVM [13] and S^{3}VM [14]. Some manifold assumption based methods mainly consider the marginal distribution of data lying on a low-dimensional manifold embedded in a high-dimensional space. Some transductive semi-supervised learning algorithms [15]–[17] have been proposed based on the manifold assumption, but the model has to be retrained for each new test sample. In the mobile location recognition problem, retraining the classifier for each query costs too much time, and thus we focus on the Laplacian SVM (LapSVM) algorithm [8], [18], which supports a natural out-of-sample extension to novel examples and has proved to perform well in many semi-supervised classification problems.

The Laplacian SVM is based on the manifold assumption [8]: if two points $\mathbf{x}_1, \mathbf{x}_2 \in \mathbb{X}$ are close in the intrinsic geometry of $\mathcal{P}_{\mathbb{X}}$ (i.e., with respect to geodesic distance on manifold), the conditional distributions $\mathcal{P}(y|\mathbf{x}_1)$ and $\mathcal{P}(y|\mathbf{x}_2)$ should be similar (i.e., should have the same label). In other words, the $\mathcal{P}(y|\mathbf{x})$ should vary smoothly along the geodesics in the intrinsic geometry of $\mathcal{P}_{\mathbb{X}}$. As a result, the LapSVM uses these geometric intuitions to extend the classical SVM, and the classifier of our location recognition problem is learned from

$$f^* = \arg\min_{f\in\mathcal{H}_K} \frac{1}{l}\sum_{i=1}^{l} \max(1-y_i^s f(\tilde{\mathbf{x}}_i^s), 0) + \gamma_A \|f\|_K^2 + \gamma_I \|f\|_I^2. \tag{1}$$

In the above objective function, the first part is the hinge loss encouraging a large margin from samples to the separating plane, the regularization $\gamma_A\|f\|_{\mathrm{K}}^2$ imposes smoothness conditions on possible solutions, and $\gamma_I\|f\|_I^2$ reflects the intrinsic structure of $\mathcal{P}_X$ and penalizes f along the manifold that the probability distribution is supported on.

An appropriate choice for $\|f\|_I^2$ is $\int_{\mathbf{x}\in\mathcal{M}} \|\nabla_{\mathcal{M}} f\|^2 \mathrm{d}\mathcal{P}_{\mathbb{X}}(x)$, which is empirically estimated by the graph Laplacian [19] associated with labeled and unlabeled examples. Since the labeled surveillance samples and unlabeled mobile samples are united to construct $\|f\|_I^2$, we join them into one training set with totally $n_s + n_t$ samples defined by

$$\begin{aligned} \tilde{\mathbf{x}}_i &= \tilde{\mathbf{x}}_i^s, \qquad \forall i = 1, \ldots, n_s; \\ \tilde{\mathbf{x}}_i &= \tilde{\mathbf{x}}_{i-n_s}^t, \quad \forall i = n_s+1, \ldots, n_s+n_t. \end{aligned} \tag{2}$$

Then the manifold based regularization can be formulated as

$$\begin{aligned} \|f\|_I^2 &= \frac{1}{(n_s+n_t)^2} \sum_{i,j=1}^{n_s+n_t} (f(\tilde{\mathbf{x}}_i) - f(\tilde{\mathbf{x}}_j))^2 W_{ij} \\ &= \frac{1}{(n_s+n_t)^2} \mathbf{f}^\top L \mathbf{f}, \end{aligned}$$

where W_{ij} are edge weights in the data adjacency graph, $\mathbf{f} = [f(\tilde{\mathbf{x}}_1), ..., f(\tilde{\mathbf{x}}_{n_s+n_t})]^\top$, $L = D - W$ is the graph Laplacian matrix, and the diagonal matrix D is given by $D_{ii} = \sum_{j=1}^{n_s+n_t} W_{ij}$.

According to Theorem 2 of [8], the optimal solution to problem (1) in $\mathcal{H}_K$ is

$$f^*(\mathbf{x}) = \sum_{i=1}^{n_s+n_t} \alpha_i K(\tilde{\mathbf{x}}_i, \mathbf{x}),$$

where coefficients α_is are parameters of the LapSVM classifier and $K(\cdot,\cdot)$ is the kernel function. We choose the RBF kernel in our experiments.

Once the parameters are solved, the label of any unlabeled query sample can be obtained by $y(\mathbf{x}_i^t) = \mathrm{sign}(f^*(\tilde{\mathbf{x}}_i^t))$. Additionally, since there are multiple classes in our problem, we extend LapSVM to the one-vs-one multi-class classifier. Furthermore, the newly enquired samples always enrich the

unlabeled dataset, which further improves the performance of location recognition. Therefore, our location recognition method can update the unlabeled data once a period to balance the computing load and recognition accuracy.

D. Algorithm Flow

We have introduced our cross-device visual location recognition utilizing the SADA and LapSVM to solve the two main problems in the recognition process. The algorithm flow of our method is presented in Algorithm 1.

Algorithm 1 Our location recognition

Input: n_s labeled images $\{(\mathbf{x}_i^s, y_i^s)\}_{i=1}^{n_s}$ captured by surveillance cameras, n_t unlabeled images $\{\mathbf{x}_i^t\}_{i=1}^{n_t}$ captured by mobile phone cameras.

Output: Predict the locations $\{y_i^t\}_{i=1}^{n_t}$.

1: Subspace generation: $S_s \leftarrow \mathrm{PCA}(\{\mathbf{x}_i^s\})$, $S_t \leftarrow \mathrm{PCA}(\{\mathbf{x}_i^t\})$.
2: Subspace alignment: $S_a = S_s S_s^\top S_t$; $\tilde{\mathbf{x}}_i^s = S_a^\top \mathbf{x}_i^s$, $\tilde{\mathbf{x}}_i^t = S_t^\top \mathbf{x}_i^t$.
3: Construct data adjacency graph with $\mathbf{x}_i^s$ and $\mathbf{x}_i^t$ by KNN, and choose heat kernel for edge weights W_{ij}.
4: Compute graph Laplacian matrix $L = D - W$.
5: Choose a kernel function $K(\mathbf{x}, \mathbf{y})$. Compute the Gram matrix $K_{ij} = K(\tilde{\mathbf{x}}_i, \tilde{\mathbf{x}}_j)$ where $\tilde{\mathbf{x}}_i$ are obtained by combining $\tilde{\mathcal{X}}^s$ and $\tilde{\mathcal{X}}^t$ as (2).
6: Choose γ_A and γ_I.
7: Solve $\{\alpha^*\}$ using the method proposed by [18].
8: Output $y_i^t \leftarrow \mathrm{sign}(f^*(\mathbf{x})) = \mathrm{sign}(\sum_{j=1}^{n_s+n_t} \alpha_j^* K(\tilde{\mathbf{x}}_i, \mathbf{x}_j))$

III. Experiments

In this section, our proposed method is evaluated on real datasets. After the introduction of the datasets and data preparation, we experimentally validate the existence of the two problems to apply a mobile visual location recognition using surveillance images. Then, we evaluate the influences on the performance of applying the SADA and LapSVM respectively to prove their effectiveness. Finally, the location recognition accuracy of our method is presented.

A. Datasets and Data Preparation

Because there is no public dataset for cross-device visual location recognition, we construct a dataset by collecting images on a floor of our building. We select 6 distinguishable locations among the covered areas of 10 surveillance cameras shown in Fig.1, including elevator room (under camera1), working area1 (under camera4), meeting room (under camera5), corridor (under camera6), working area2 (under camera8) and coffee room (under camera7). Images captured by surveillance cameras and mobile phones are denoted by S and M respectively. Limited by the space of an indoor scene, we cannot take too many representative photos at one location. Thus we randomly select 60 frames in a two hours long video of every surveillance camera, and the size of S is 360. A total of 80 images are captured by the camera of a mobile phone at one location, and the size of M is 480.

To prepare the experimental data, we resize all images to 320×240 and extract two types of features. The first is the bag-of-visual-words (BoW, [20]) feature based on SIFT descriptors. To extract such a feature, the SIFT features are extracted from the images, and a codebook of size 300 is generated by k-means clustering on S. Then the images from S and M are represented by a 300 bin histogram corresponding to the codebook. For the second type of feature, each image is directly represented by a 512-dimensional GIST feature [21], which models a holistic representation of the location.

TABLE I. Location recognition accuracy using an SVM classifier.

Method	BoW	GIST
$M \rightarrow M$	98.33	95.00
$S \rightarrow M$	53.33	43.33

B. Experimental Validation of the Two Problems

We denote the strategy of our visual location recognition by $S \rightarrow M$, where the training images and test images are from S and M respectively. Similarly, we denote the general visual location recognition by $M \rightarrow M$, where both training and test images are from M. There are 60 training samples and 20 test samples for each of the 6 locations. Preliminarily, we compare the performance of our strategy and the general visual location recognition method by measuring the classification precision in one randomized trial, where an RBF kernel based SVM classifier is used and the results are shown in TABLE I. The performance of mobile location recognition is significantly degraded by using surveillance images instead of mobile images as training samples. Thus, there may exist some problems if a classifier trained on S is directly used on M.

In order to further explore the problems existing in our location recognition method, the number of valid labeled training samples is first evaluated in Fig. 3. Along with the increasing of training samples, the number of query images that are located accurately starts to level off when the number of training samples for each location is greater than 2 (resp. 11) using the GIST (resp. BoW) feature. It indicates that there are only a few valid labeled samples in S, which are incapable to represent the variety of the mobile images and seriously deteriorate the performance of location recognition.

Second, we randomly select one data sample from S and 60 data samples from M for each location. By embedding the GIST data in a low-dimensional manifold with the Laplacian Eigenmaps [19], we present a visual example of data distribution between S and M in Fig.4a. Ideally, if a surveillance image $s \in S$ and a mobile image $m \in M$ are captured at one location and generated from the same marginal distribution, they are more liable to lie on the same manifold. However, as shown in Fig.4a, the surveillance images in S (represented by pentagrams) are far away from their corresponding manifold generated by the mobile images in M at the same location (points with the same color as the pentagram), especially for camera 4 and camera 8. This phenomenon proves the discrepancy between the distributions of S and M, which makes a classifier trained on S perform poorly on M.

C. Evaluation of SADA and LapSVM Individually

We now empirically study the influence on performance by introducing SADA or LapSVM individually. First, the

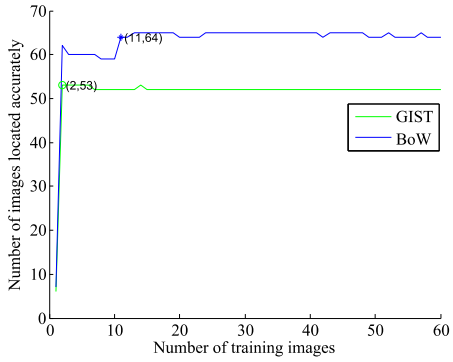


Fig. 3. Location recognition results on different number of training images using the SVM classifier.

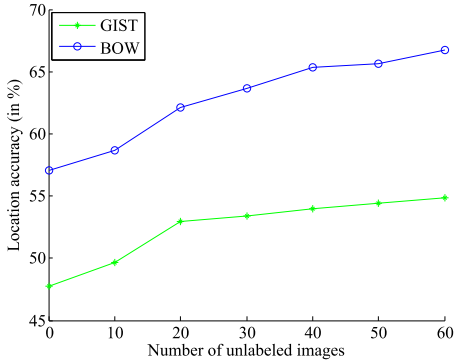


Fig. 5. Location recognition accuracy on different number of unlabeled images using a LapSVM classifier.

effect of SADA on eliminating the data differences between S and M is evaluated. We also use the experimental setup outlined in the previous section and present the results in Fig. 4. While most surveillance images are far away from the manifold corresponding to the same location in Fig.4a, both the surveillance images and mobile images of the same location lie on the one manifold in Fig.4b. In addition, most of the manifolds are smoother and more compact after SADA, which is also helpful for the following LapSVM learning. More numerical experimental results will be shown in Section III-D.

We then experiment with LapSVM when adding more and more unlabeled images from M to training set S. This process indeed simulates the situation that there are more and more mobile users uploading query images to the server for location recognition. For each location, we randomly select one labeled image from S and different numbers (from 0 to 60) of unlabeled images from M, and then apply the LapSVM classification. The average location recognition accuracy of LapSVM over 50 random trials is shown in Fig.5. It is notable that when the number of unlabeled images from M is 0, the LapSVM degrades to the baseline SVM that training sets are only consisted of images from S. The recognition accuracy of LapSVM is improved gradually while there are more and more unlabeled mobile images in training set. Thus, if there are enough unlabeled samples in M, the LapSVM is an effective way to solve the problem that there are only a few valid labeled training samples in S.

TABLE II. LOCATION RECOGNITION PERFORMANCE COMPARISON.

Method	BoW	GIST
SVM	56.87 ± 0.7	47.97 ± 0.5
SADA + SVM	62.01 ± 0.8	52.14 ± 0.7
LapSVM	67.21 ± 0.8	55.44 ± 1.8
SADA + LapSVM	**75.52 ± 0.9**	**65.92 ± 1.2**

D. Location Recognition Result

To evaluate the performance of our location recognition method, we compare against a number of baselines described below.

- **SVM** : An RBF kernel based SVM classifier trained using only the original samples in S.
- **SADA + SVM** : An RBF kernel based SVM classifier trained using the aligned samples in S by SADA.
- **LapSVM** : A Laplacian SVM classifier trained using both the original labeled samples in S and original unlabel samples in M.

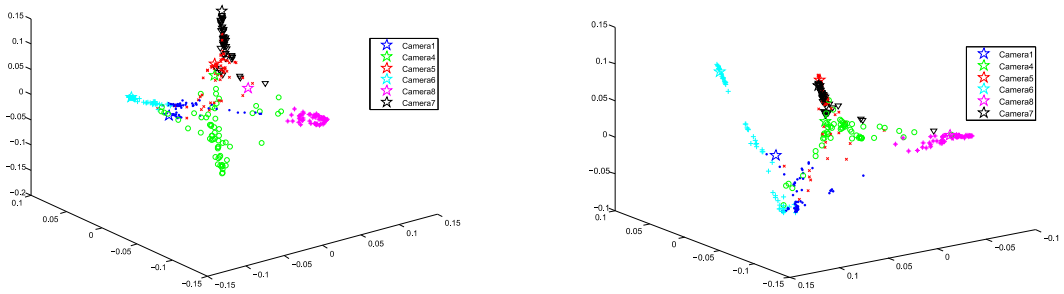
Additionally, our method trains a LapSVM classifier using both the labeled samples in S and unlabeled samples in M after SADA, denoted by **SADA + LapSVM**. The algorithm hyperparameters for all methods include dimension d for SADA, number of nearest neighbors k in graph construction, $\gamma_A = 10^{-5}$ and $\gamma_I = 1.0$ for LapSVM. The parameters for GIST feature are $d = 20$ and $k = 2$, while $d = 17$ and $k = 5$ for BoW feature. Because there are very few valid training data in S, we randomly select one labeled example for each location in our experiments. This means that there are a total of only 6 labeled examples available in the training set.

We randomly select 60 unlabeled examples from M for each location as historical queries existing in the recognition system, and the rest are considered as new queries. For each method, the model is learned using surveillance images as labeled data and the historical mobile queries as unlabeled data. Then the locations of the new queries are recognized by each method, and the average accuracy over 50 random trials is shown in TABLE II.

In TABLE II we observed that both **SADA + SVM** and **LapSVM** give a better performance than the simple **SVM** over different representation of images. Meanwhile, it is interesting to note that the performance improvement of **SADA + SVM** is not as obvious as **LapSVM**. This is to be expected since there are only a few valid training samples to learn a classifier. However, as mentioned above, the SADA is helpful for semi-supervised LapSVM learning. As a result, our location recognition method has significant performance increase over other methods. This again validates that the two problems do exist in our location search method.

IV. CONCLUSION

In this paper, we presented a new visual location recognition method which jointly uses the surveillance cameras and mobile phone cameras. By unifying a subspace alignment based domain adaption and the Laplacian SVM, we effectively solved the twisted domain adaptation and semi-supervised learning problem where the scarce labeled data and plentiful



(a) Before SADA (b) After SADA

Fig. 4. Comparison of data distributions before and after SADA (best viewed in color). Each color corresponds to one location, and pentagrams represent surveillance images while others represent the mobile images.

unlabeled data exist in different domains. Experimental results showed the superiority of our method over other related methods. For the future work, we will consider the situation that there are foreground moving targets in the scene, and the more difficult outdoor location recognition.

ACKNOWLEDGMENT

We thank Junge Zhang and Yanming Zhang for their insightful suggestions. This work is funded by the National Basic Research Program of China (Grant No. 2012CB316302) and National Natural Science Foundation of China (Grant No. 61322209).

REFERENCES

[1] J. Z. Liang, N. Corso, E. Turner, and A. Zakhor, "Image Based Localization in Indoor Environments," in *International Conference on Indoor Positioning and Indoor Navigation*, 2013.

[2] B. Girod, V. Chandrasekhar, D. Chen, and et al, "Mobile Visual Search," *IEEE Signal Processing Magazine*, vol. 28(4), pp. 61–76, 2011.

[3] J. Z. Liang, N. Corso, E. Turner, and A. Zakhor, "Reduced-Complexity Data Acquisition System for Image Based Localization in Indoor Environments," in *International Conference on Indoor Positioning and Indoor Navigation*, 2013.

[4] D. Chen, G. Baatz, K. Koser, and et al, "City-Scale Landmark Identification on Mobile Devices," in *CVPR*, 2011.

[5] F. X. Yu, R. Ji, and S.-F. Chang, "Active Query Sensing for Mobile Location Search," in *Proceedings of the 19th ACM International Conference on Multimedia*, 2011.

[6] N. Corso and A. Zakhor, "Indoor Localization Algorithms for an Ambulatory Human Operated 3D Mobile Mapping System," *Remote Sensing*, vol. 5(12), pp. 6611–6646, 2013.

[7] B. Fernando, A. Habrard, M. Sebban, and T. Tuytelaars, "Unsupervised Visual Domain Adaptation Using Subspace Alignment," in *ICCV*, 2013.

[8] M. Belkin, P. Niyogi, and V. Sindhwani, "Manifold Regularization: A Geometric Framework for Learning from Labeled and Unlabeled Examples," *The Journal of Machine Learning Research*, pp. 2399–2434, 2006.

[9] C. Wang and S. Mahadevan, "Manifold Alignment without Correspondence," in *IJCAI*, 2009.

[10] I.-H. Jhuo, D. Liu, D. T. Lee, and S.-F. Chang, "Robust Visual Domain Adaptation with Low-rank Reconstruction," in *CVPR*, 2012.

[11] R. Gopalan, R. Li, and R. Chellappa, "Domain Adaptation for Object Recognition: An Unsupervised Approach," in *ICCV*, 2011.

[12] B. Gong, Y. Shi, F. Sha, and K. Grauman, "Geodesic Flow Kernel for Unsupervised Domain Adaptation," in *CVPR*, 2012.

[13] T. Joachims, "Transductive Inference for Text Classification Using Support Vector Machines," in *ICML*, 1999.

[14] O. Chapelle, V. Sindhwani, and S. Keerthi, "Optimization Techniques for Semi-supervised Support Vector Machines," *The Journal of Machine Learning Research*, pp. 203–233, 2008.

[15] T. Joachims, "Transductive Learning via Spectral Graph Partitioning," in *ICML*, 2003.

[16] X. Zhu, Z. Ghahramani, and J. Lafferty, "Semi-supervised Learning Using Gaussian Fields and Harmonic Functions," in *ICML*, 2003.

[17] M. Belkin and P. Niyogi, "Using Manifold Stucture for Partially Labeled Classification," *Advances Neural Information Processing Systems*, pp. 953–960, 2003.

[18] S. Melacci and M. Belkin, "Laplacian Support Vector Machines Trained the Primal," *Journal of Machine Learning Research*, vol. 12, pp. 1149–1184, 2011.

[19] M. Belkin and P. Niyogi, "Laplacian Eigenmaps for Dimensionality Reduction and Data Representation," *Neural Computation*, vol. 15(6), pp. 1373–1396, 2003.

[20] J. Sivic and A. Zisserman, "Video Google: A Text Retrieval Approach to Object Matching in Videos," in *ICCV*, 2003.

[21] A. Oliva and A. Torralba, "Modeling the Shape of the Scene: A Holistic Representation of the Spatial Envelope," *International Journal of Computer Vision*, vol. 42(3), pp. 145–175, 2001.