Common Landmark Discovery in Urban Scenes

Chokushi Yuuto
University of Fukui

Tanaka Kanji
University of Fukui

Ando Masatoshi
University of Fukui

3-9-1, Bunkyo, Fukui, 910-8507, Japan
e-mail tnkknj@u-fukui.ac.jp

Abstract

In this paper, we introduce a method for unsupervised discovery of landmark objects in urban scenes. Unlike existing methods that deal with pre-defined and typically popular landmarks whose models or training data are available on the internet (e.g. Nortre Dame de Paris, Leaning tower of Pisa, etc.), we develop a new unified framework, common landmark discovery (CLD), which discovers even a priori undefined and general landmarks, which should be defined as commonly recognizable objects, from a given random collection of images acquired in urban scenes. The key idea is to discover landmark objects in a bottom-up manner, as opposed to the conventional top-down supervised approaches. More formally, our approach begins by sampling subimages (or visual phrases) from images, mines common phrases from image pairs, matches regions between phrase pairs, clusters similar region segments, and as a final result, automatically defines the largest cluster as a landmark object. There are four key properties about the proposed techniques: 1) The CLD process is unsupervised, without requiring a priori knowledge on pre-defined landmarks; 2) The CLD method is robust, without depending on good image segmentation, and is able to handle scale variations of the object; 3) An image is semantically characterized as object patterns discovered in it; 4) An image is compactly and discriminatively described in a form of bag-of-bounding-boxes (BoBB), employing bounding box - based object annotation and knowledge transfer. We validate the presented techniques using challenging real images.

1 Introduction

Although the problem of landmark object recognition has been studied in recent years [1]-[3], a robust solution for recognition of *general* landmark objects does not exist yet due to two major reasons: the lack of training data for landmark learning and poor scalability of learning and recognition algorithms. Most previous efforts have focused on predefined and typically popular landmarks (e.g. Nortre Dame de Paris, Leaning tower of Pisa, etc.), whose models or training data are available on the internet. As a result, almost all existing schemes for landmark recognition cannot be applied to the large body of potential applications of landmark recognition, such as geo-localization and tourist guide.

In this work, we develop a new unified framework, called common landmark discovery (CLD), which discovers even a priori undefined and general landmark objects, which should be defined as commonly *recognizable* objects, from a given random collection of images acquired in urban scenes (Fig.1). The key idea is to discover landmark objects in a bottom-up manner, as opposed to the conventional top-down recognition of pre-defined landmarks, by using techniques derived from visual phrases [4], common pattern discovery [5], co-segmentation [6], and region matching algorithms [7]. More formally, our approach begins by extracting visual phrases from images, mines common phrases from image pairs, matches regions between phrase pairs, clusters similar region segments, and as a final result, automatically defines the largest cluster as a landmark object. We validate the presented techniques using challenging real images.

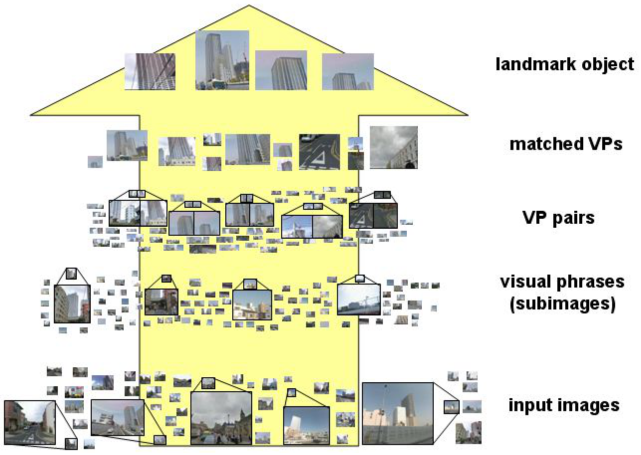

Figure 1. The common landmark discovery (CLD) task. Our approach automatically discovers even a priori undefined and general landmarks in a bottom-up manner, as opposed to the conventional top-down supervised approaches.

There are four key properties about the proposed common landmark discovery (CLD) framework.

1. The CLD process is *unsupervised*, without requiring a priori knowledge on pre-defined landmarks; The estimate can be done with a random collection of images acquired in urban scenes, without having to plan camera viewpoints.
2. The CLD method is *robust*, without depending on good image segmentation, and is able to handle scale variations of the object; It is based on the state-of-the-art technique for common pattern discovery, spatial random partition [4].
3. An image is *semantically* characterized as a small number of object patterns discovered in it; The proposed estimation method beyond conventional common pattern discovery, by efficiently exploring the image space to exploit further information of object-level landmark discovery, via robust region matching framework [7].

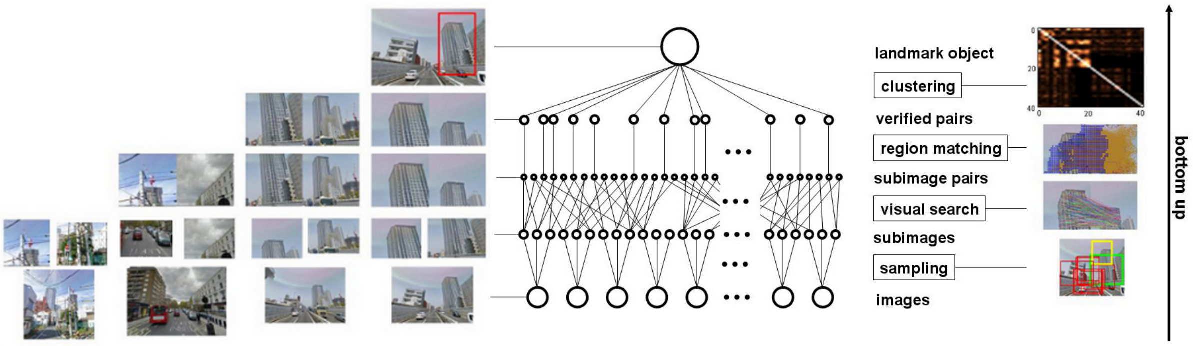

Figure 2. Algorithm pipeline. Our approach begins by extracting visual phrases from images, mines common phrases from image pairs, matches regions between phrase pairs, clusters similar region segments, and as a final result, automatically defines the largest cluster as a landmark object.

4. An image is compactly and discriminatively described in a form of bag-of-bounding-boxes (BoBB), employing traditional bounding box - based object annotation and knowledge transfer [8]; Thus, it is scalable for larger collection of images.

A constraint from our design consideration is the use of bounding boxes (BBs) as a compact and discriminative method for image description. With large-scale applications in mind, we compactly describe a landmark object in an input image by a pair of BBs, which indicate where the landmark object is located w.r.t. an input image and w.r.t. a reference image, respectively. This allows us to describe the input image discriminatively by the pose and the shape (i.e. width, height) of each bounding box that crops a landmark object. The bag-of-bounding-boxes (BoBB) image representation can be potentially used for indexing and retrieving images in large scale applications. We therefore develop a framework which consists of techniques that use the BBs (or subimages) as a primary representation for object, e.g. spatial random partition.

2 Approach

Our bottom-up approach uses different image -based criteria and matching algorithms for common landmark discovery (CLD). Fig.2 describes the overall scheme. The CLD task consists of the following steps. (1) Sampling. For each of the input images, we first obtain a random collection of subimages (or visual phrases) by sampling bounding boxes from the image region. (2) Visual search. We find a set of relevant phrase pairs by efficiently matching local feature descriptors (e.g. SIFT) between each subimage pair. (3) Region matching. Given a collection of relevant phrase pairs, we verify each phrase pair by detailed region matching via a robust correspondence growing algorithm. As a result, we obtain a confusion matrix that carries information about the similarity between each phrase pair. (4) Similarity graph. We obtain a similarity graph whose node represents a phrase in the collection by connecting every phrase pair with high similarity. (5) Clustering. We cluster the nodes in the similarity graph, and find a largest subset (i.e. the target landmark) of the nodes in which each node is connected to every other node in the subset. The following subsections provide detailed explanation of above subtasks.

2.1 Sampling

The sampling subtask aims to sample a pool of subimages from a given input image. The pool of subimages will be used as the image's representation for common pattern discovery (CPD), where each pool of subimages is matched against one another based on the fact that a common pattern is likely to be co-exist in a good number of subimages across images. Our approach follows the recent studies on spatial context [9, 10, 11], where the goal is to enhance the discriminative power of individual local features by bundling co-located visual features together into a visual phrase. In previous studies, there were mainly two ways to select the spatial context to compose the visual phrase, those which use image segmentation as a cue [9] and those which is based on fixed scale visual phrase (e.g. k spatial nearest neighbors) [10, 11]. In contrast, our subimage-based approach is particularly based on the recently developed technique, spatial random partition (SRP) [4], which overcomes the limitations of the previous techniques: It does not depend on good image segmentation; It is able to handle scale variations of the object.

2.2 Visual Search

The visual search subtask aims at efficiently mining candidates of tentative correspondence between subimage pairs across images. SIFT features are extracted from each subimage and a subimage pair is viewed as a tentative correspondence when there exists at least one SIFT feature pair whose distance of normalized SIFT descriptors is smaller than 0.4 [7].

2.3 Region Matching

The region matching subtask aims to verify each pair of subimages of the tentative correspondence. Our approach follows the recent studies on region matching techniques (e.g. common pattern discovery [5], co-segmentation [6], and correspondence growing [12]). By simultaneously looking at both images, it attempts to find larger object-level correspondence based on the fact that the true correspondences are supported by larger object region than false ones. In particular, our

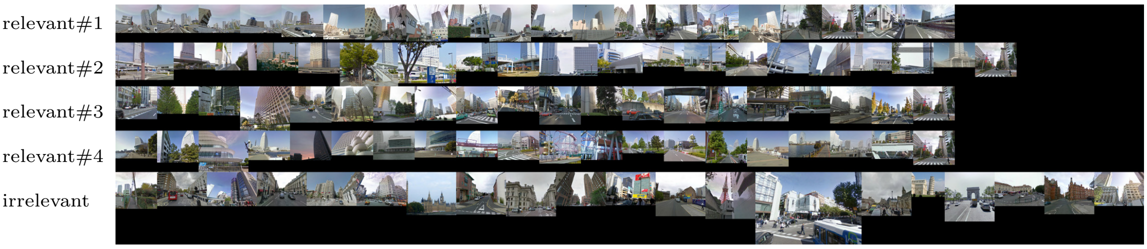

Figure 3. Dataset. Four different image collections are used as datasets. Each image collection consists of 20 relevant images and 20 irrelevant images. Shown in 1st-4th rows in the figure are relevant images, while shown in the 5th row are irrelevant images.

method is based on the match growing algorithm proposed in [7]. It models an image by employing a regular lattice to discretize a query image into grid nodes. It is an iterative algorithm, and designed to maximize a pre-defined objective likelihood function within a probabilistic MCMC framework. For a detailed explanation of the framework, please refer to Appendix A.

2.4 Similarity Graph

Given similarity between each phrase pair, we construct a graph, called similarity graph, whose node represents each phrase and which connects those relevant phrase pairs whose similarity score exceeds a prelearned threshold value.

2.5 Clustering

Given a similarity graph, the clustering subtask aims to find a set of similar groups. It constructs a directed graph whereby an edge connects each node and its most similar neighbor, and then find the largest subset of nodes in which each node is connected to every other node in the subset. We define the largest cluster of nodes or phrases as the common landmark object and output it as the final result.

3 Experiments

The common landmark discovery system has been implemented in C++ with openCV, and has been successfully tested on various scenes. In particular, we did experiments to evaluate our algorithm with four collections of random images acquired in urban scenes in uncontrolled situations and at random viewpoints, as shown in Fig.3. While the proposed framework deals with a priori undefined landmarks in an unsupervised setting, for the sake of evaluation, each image collection is designed so that there is at least one recognizable object from majority of the images. Currently, whether an image is recognizable or not is manually decided (see detail for B). As a primary performance index, we use precision, recall and the f-measure. In this case, precision is the ratio of retrieved (i.e. output) images that are relevant to all retrieved images, while recall is the ratio of the retrieved images that are relevant to the total number of relevant images in the collection. Fig.4 shows an instance of confusion matrix obtained by the conventional and the proposed methods. The comparison of the two confusion matrices shows that the similarity values obtained from the proposed method are higher than the conventional method at the landmark object images. Fig.5 shows an instance result of the common landmark discovery with the proposed method. It can be seen that the image collection consists of confusion images that look similar and have the similar local statistics, where the conventional local feature matching techniques fail. In contrast, the proposed method that mines candidates of relevant visual phrases then verify and cluster them into a landmark object in a bottom up manner successfully discovered the landmark object with high precision in an unsupervised manner. Tab.1 compares results from the three different methods, the standard local feature matching method (LF), the visual phrase method (VP), and the proposed method that combines the advantages of the local feature matching and region matching. One can see that the proposed method clearly outperforms the other two methods in all the cases.

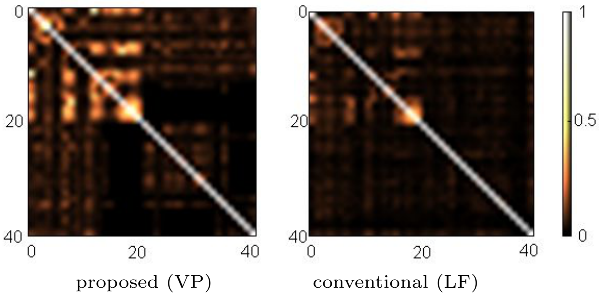

Figure 4. Confusion matrix carrying information of similarity between every pair of input images.

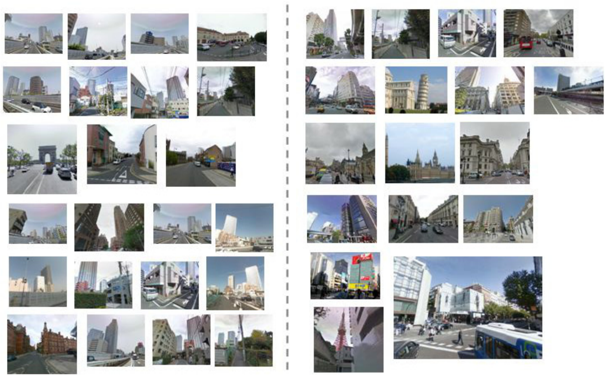

Figure 5. Results of common landmark discovery from an experiment where the image collection is split into those which belong to the landmark object (left panel) and the rest (right panel).

Table 1. Comparison of different methods for four different datasets. LF: local features (conventional). VP: visual phrase. VP+RM: visual phrase with region matching (proposed method).

dataID	method	precision	recall	f-measure
1	LF	1.000	0.400	0.571
	VP	0.667	0.700	0.683
	VP+RM	1.000	0.600	0.750
2	LF	1.000	0.300	0.462
	VP	0.857	0.900	0.878
	VP+RM	0.857	0.900	0.878
3	LF	0.556	0.250	0.345
	VP	0.643	0.900	0.750
	VP+RM	0.692	0.900	0.783
4	LF	1.000	0.250	0.400
	VP	0.722	0.650	0.684
	VP+RM	1.000	0.550	0.710

4 Conclusions

Based on the current techniques in the field of computer vision, we have proposed a novel contribution in the landmark recognition. Unlike existing methods that deal with pre-defined and typically popular landmarks whose models or training data are available on the internet (e.g. Nortre Dame de Paris, Leaning tower of Pisa, etc.), our approach can automatically discover even a priori undefined and general landmarks, which should be defined as commonly recognizable objects, from a given random collection of images acquired in urban scenes. The key idea is to discover landmark objects in a bottom-up manner, as opposed to the conventional top-down supervised approaches, by using techniques derived from common pattern discovery, co-segmentation, visual phrases, and correspondence-growing algorithms. We have further evaluated the presented techniques using challenging real images and found that the proposed method outperforms the previous techniques.

Acknowledgement

This work was partially supported by MECSST Grant (23700229, 30325899), by KURATA grants and by TATEISI Science And Technology Foundation.

References

[1] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Modeling the world from internet photo collections. *IJCV*, 80(2):189–210, 2008.

[2] Noah Snavely, Steven M. Seitz, and Richard Szeliski. Photo tourism: Exploring photo collections in 3d. In *SIGGRAPH Conference Proceedings*, pages 835–846, New York, NY, USA, 2006. ACM Press.

[3] Qiang Hao, Rui Cai, Zhiwei Li, Lei Zhang, Yanwei Pang, and Feng Wu. 3d visual phrases for landmark recognition. In *CVPR*, pages 3594–3601, 2012.

[4] Yuning Jiang, Jingjing Meng, and Junsong Yuan. Randomized visual phrases for object search. In *CVPR*, pages 3100–3107, 2012.

[5] Hung-Khoon Tan and Chong-Wah Ngo. Common pattern discovery using earth mover's distance and local flow maximization. *Computer Vision, IEEE International Conference on*, 2:1222–1229, 2005.

[6] Gunhee Kim, Eric P. Xing, Fei-Fei Li, and Takeo Kanade. Distributed cosegmentation via submodular optimization on anisotropic diffusion. In *ICCV*, pages 169–176, 2011.

[7] Minsu Cho, Young Min Shin, and Kyoung Mu Lee. Unsupervised detection and segmentation of identical objects. In *CVPR*, pages 1617–1624, 2010.

[8] Matthieu Guillaumin and Vittorio Ferrari. Large-scale knowledge transfer for object localization in imagenet. In *CVPR*, pages 3202–3209, 2012.

[9] Zhong Wu, Qifa Ke, Michael Isard, and Jian Sun. Bundling features for large scale partial-duplicate web image search. In *CVPR*, pages 25–32, 2009.

[10] Josef Sivic and Andrew Zisserman. Efficient visual search of videos cast as text retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 31(4):591–606, 2009.

[11] Yimeng Zhang, Zhaoyin Jia, and Tsuhan Chen. Image retrieval with geometry-preserving visual phrases. In *CVPR*, pages 809–816, 2011.

[12] Jan Cech, Jiri Matas, and Michal Perdoch. Efficient sequential correspondence selection by cosegmentation. *IEEE Trans. Pattern Anal. Mach. Intell.*, 32(9):1568–1581, 2010.

[13] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2007.

A Region Matching Algorithm

The match growing is an iterative algorithm. For each step of iteration, either of two moves are possible: (1) correspondence initialization and (2) region expansion. In the former case, a new correspondence of SIFT interest points is established and a correspondence region pair is initialized. In the latter case, a *frontier node*, a node at the boundary inside the current region within the query image is randomly sampled, then one of its (at most 4) neighbor nodes that lie outside the region is considered as a potential member of the region, and as a next candidate of the expansion attempt. A pre-defined parameter $\omega = 0.5$ controls the ratio between the two moves: the former move is chosen with probability ω and the latter is chosen with probability $1-\omega$.

The likelihood function to maximize is defined as follows. Consider that we have an input subimage pair, a query image I^Q and a candidate image I^i in the collection. The image I^Q contains a lattice with a set of N^Q nodes and edges. An object hypothesis h is represented by a pairing (X^h, Y^h), of the corresponding node labels $X^h = \{x_j\}_{j=1,\cdots,N^Q}$ and a set of affine transformations $Y^h = \{y_k\}_{k=1,\cdots,N^Q}$. A node label x_j is a binary indicator which labels the corresponding node as belonging to the common object or not. An affine transformation (i.e. rotation, translation, scaling, etc.) y_k maps the location of the k-th node on the query subimage I^Q to the corresponding location on the candidate subimage I^i. Using the above terminology, the problem of common pattern discovery is formulated as finding an optimal set $Z = \{(X^h, Y^h)\}$ of landmark hypotheses that maximizes an objective function in the form:

$$P(Z|I^Q, I^i) = P(I^Q, I^i|Z)P(Z). \quad (1)$$

We explain each term in (1) in the following expressions (2)-(4).

We consider $P(Z)$ as a prior model for the object, and predict it from the size and the shape of the object:

$$P(Z) = \prod_i \exp\{-R_i/R^O\}\exp\{-S_i/S^O\}. \tag{2}$$

R_i is the number of nodes in the current region. S_i is to evaluate the similarity of the configuration of nodes in the regions between the query and the candidate images. Considering the fact that nodes on the query image lie on a regular lattice by definition, we can evaluate the similarity S_i in terms of *lattice-ness* of nodes in the candidate image, which is defined as:

$$S_i = \sum_{T_i}\sum_{j=1}^{4} \left| p_j - p_{(j\%4+1)} \right|, \tag{3}$$

where % is the mod operation, T_i is a set of nodes in the current region, and that have four neighbors that are denoted as $\{p_j\}_{j=1}^4$ (i.e. not the boundary nodes).

In (1), $P(I^Q, I^i|Z)$ is a likelihood function based on similarity of appearance between the correspondence region of interest on the query image and its corresponding region (i.e. obtained via affine transformations Y^h) on the candidate image. An averaged color distance D_i in the normalized RGB color-space is employed as the similarity measure:

$$P(I^Q, I^i|Z) = \prod_{i=1}^{N^Q} \exp\{-D_i/D^O\}. \tag{4}$$

In the above expressions (2)-(4), symbols R^O, S^O and D^O are used as normalizer constants. In order to sample in the high dimensional distribution represented by (1), some form of MCMC such as Gibbs are required. To deal with this problem, we employ a deterministic sampling strategy, a family of strategies employed in [7] and other recent studies. In particular, we employ a strategy that accepts only those moves that expand or increase the size of correspondence region, and automatically reject those ones that decrease the sizes.

B Definition of Recognizable Objects

In the experiment, the ground truth is given by using the notion "recognizable", which is currently manually defined by the authors. In practice, definition of recognizable objects is very important: For instance, it is not difficult for us to achieve 100% recognition rate, if we can exclude objects as unrecognizable when we have got objects that are not recognized by the system. In the case of Oxford building dataset [13], each image is classified into four different categories: Good - A nice, clear picture of the object/building; OK - More than 25% of the object is clearly visible; Bad - The object is not present; Junk - Less than 25% of the object is visible, or there are very high levels of occlusion or distortion. Intuitively, our definition of "recognizable" is near to the "Good" within the above categorization. However, clear definition of recognizable objects is itself a difficult problem and will have to be addressed in our future study.