Translation-Invariant Scene Grouping

Pin-Ching Su, Hwann-Tzong Chen
Department of Computer Science
National Tsing Hua University, Taiwan

Koichi Ito, Takafumi Aoki
Graduate School of Information Sciences
Tohoku University, Japan

***Abstract*—We present a new approach to the problem of grouping similar scene images. The proposed method characterizes both the global feature layout and the local oriented edge responses of an image, and provides a translation-invariant similarity measure to compare scene images. Our method is effective in generating initial clustering results for applications that require extensive local-feature matching on unorganized image collections, such as large-scale 3D reconstruction and scene completion. The advantage of our method is that it can estimate image similarity via integrating global and local information. The experimental evaluations on various image datasets show that our method is able to approximate well the similarities derived from local-feature matching with a lower computational cost.**

***Index Terms*—Phase-Only Correlation, Scene Clustering, Image Matching, SIFT Descriptor, Gist Descriptor**

I. Introduction

The state-of-the-art structure-from-motion systems such as [1], [12], [13], [14] can model large-scale 3D structures using unorganized Internet photo collections. One of the key techniques that contribute to the success of those systems is SIFT [8] keypoint matching. The DoG-SIFT keypoint detector is invariant to scaling and rotation. It can be used to extract keypoints stably from various images. Furthermore, SIFT descriptors are distinct for matching, and therefore make the task of finding correspondences among images more robust. Other applications such as image completion from Internet photos [2] might also rely on SIFT matching for finding initial candidates. However, matching SIFT keypoints in a large image dataset is time-consuming: Typically, a keypoint is represented by a 128-dimensional SIFT descriptor, and an image may contain thousands of keypoints. Finding correspondences of keypoints among images would require a huge number of comparisons on all pairs of 128-dimensional SIFT descriptors.

SIFT keypoints and descriptors are local features, and to alleviate the substantial computation of pairwise descriptor similarities, image representations that characterize global information can be used to find the initial clusters. SIFT matching can then be applied to only the images that belong to the same cluster, and thus the computational cost is reduced. For example, the 'gist' descriptor [10] is employed in [7] to group similar views for 3D reconstruction. The scene completion algorithm presented in [4] also uses the gist descriptor to group similar scenes. The gist descriptor aggregates directional filter responses at multiple scales into coarse spatial bins, *e.g.* a 4×4 grid of bins. Although the gist descriptor can model the rough layout of edges and textureness in an image, it is not accurate enough to represent specific image contents. Moreover, if the translation between two images is larger than the width of spatial bin, the sum of squared differences between the gist descriptors of the two images will be very large. Translation has to be specifically handled when two scenes are compared by their gist descriptors, as is done in [11].

We present a new method based on *phase-only correlation* (POC) [5], [6], [9] for comparing and grouping images of similar views. The goal and contribution of our work is to show that the global and local information in a scene can be gracefully integrated by our design of image representation and similarity measure. Our method is able to provide more detailed global layout information of local features, and meanwhile, is translation-invariant owing to the good property of POC functions. Usually our method is about 30 times faster than SIFT matching at computing the similarity matrix for clustering. The experimental evaluations based on nearest-neighbor recall rates also show that our method is more suitable than the gist descriptor for the task of deriving the initial clusters before performing exhaustive SIFT matching.

II. Algorithm

Given a collection of photos of different landmarks, places, and scenes, we would like to divide automatically the photos into subsets of similar views. The photos may vary in lighting conditions and compositions, but we hope to associate each photo with other photos presenting similar views of the same scene. The key issue needed to be addressed would be the design of image representation and similarity measure. To begin with, we crop each photo to get a square image as the input. More specifically, since a photo can be either in 'landscape' or 'portrait' orientation, to obtain a consistent size for subsequent processing, we extract the central square area of which the side is set to 95% of the shorter side of the given photo. We assume that the photos containing a similar view of a scene should all cover a certain portion of the scene at their central areas. This assumption is plausible because, in practice, the locations where the pictures can be taken are not arbitrary but somewhat restricted. Therefore, although the compositions may vary among photos, the photos of a scene often include similar views of some landmarks or main subjects. An example of photo cropping is shown in Fig. 1. The cropped photo is then converted to grayscale and resized to 256×256 pixels, and is taken as the input image for the computation of feature responses.

978-1-4577-0121-4/11/$26.00 ©2011 IEEE

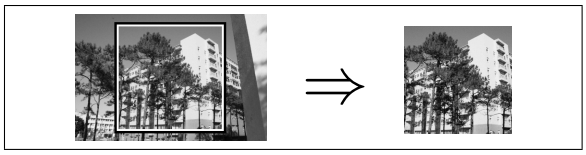

Fig. 1. An input image may be in 'landscape' or 'portrait' orientation, and so we crop it and keep only the central square area. The cropped image is then resized to produce a 256×256-pixel grayscale image for feature extraction.

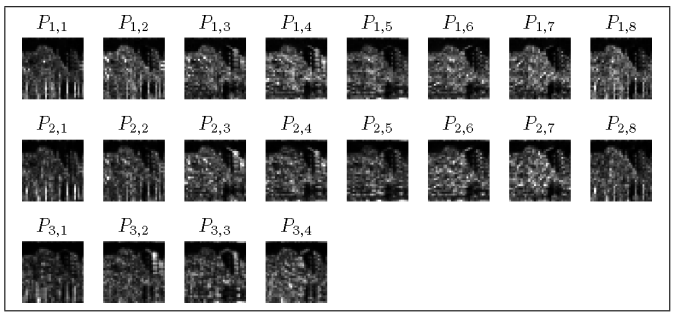

Fig. 2. These 20 feature patches $\{P_{s,t}\}$ are derived from the Gabor-feature pyramid of the input image shown in Fig. 1. The first subscript denotes the level of pyramid, and the second subscript is the index of band (orientation) at each level. Each feature patch contains 32×32 average responses derived from 32×32 spatial bins of each pyramid band.

A. Gabor Feature Pyramid

We use Gabor filters to compute low-level features in images. From an input image of 256×256 pixels, we build a feature pyramid of three levels (scales), and each of the three levels comprises $8, 8, 4$ bands of different directional-filter responses, respectively. Let $\{F_{s,t}\}$ denote the pyramid, where the level is indexed by $s = 1, \cdots, 3$ and the band is indexed by $t = 1, \cdots, 8$ (or $t = 1, \cdots, 4$ for the last level). Hence totally we get 20 filtered outputs. Each filtered output $F_{s,t}$ is then divided into 32×32 spatial bins, and within each spatial bin we compute the average response. The 32×32 spatial bins can thus be viewed as a (32×32)-pixel patch consisting of locally representative features. Consequently, from the three-level pyramid we build 20 feature patches, denoted by $\{P_{s,t}\}$, where $s = 1, \cdots, 3$, and $t = 1, \cdots, 8$ or $t = 1, \cdots, 4$, depending on the level. Such a representation can handle local variations in image. Fig. 2 illustrates the feature patches $\{P_{s,t}\}$ of the input image in Fig. 1.

B. Phase-Only Correlation

The aforementioned representation is able to characterize local features at different scales. Changes in lighting condition or local variations caused by slight rotation, scaling, and translation can be well handled owing to the combination of Gabor feature pyramid and spatial binning. We will further use a robust similarity measure to deal with significant translation due to different photo composition.

We employ the technique of *phase-only correlation* (POC) [5], [6], [9] to compare two images by their feature patches. In particular, we use the *band-limited* POC function [5] for our task. The advantage of POC functions is that they are shift-invariant and insensitive to variations in brightness and contrast. Furthermore, by using the band-limited POC function, we can eliminate meaningless high-frequency components and thus make the similarity measure more reliable. Also note that the maximum value of the correlation peak of the band-limited POC function is normalized to 1, which is convenient for deriving a similarity score.

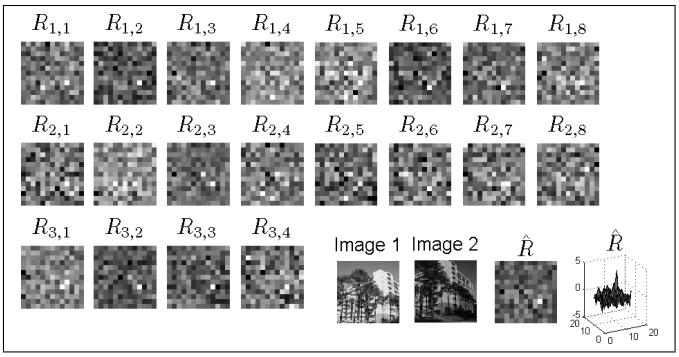

Fig. 3. The outputs $\{R_{s,t}\}$ are produced by the band-limited POC function on the two sets of feature patches $\{P^{(1)}_{s,t}\}$ and $\{P^{(2)}_{s,t}\}$ of the two input images shown at the bottom row. We aggregate $\{R_{s,t}\}$ by taking location-wise average over s, t and obtain $\hat{R}$. It can be seen that $\hat{R}$ contains a significant peak, and the location of the peak reflects the most significant translation between the two input images.

Given the two sets of feature patches $\{P^{(1)}_{s,t}\}$ and $\{P^{(2)}_{s,t}\}$ of two input images, we use the band-limited POC function to compute the correlation between each pair of $P^{(1)}_{s,t}$ and $P^{(2)}_{s,t}$. Let $R_{s,t}$ denote the output of band-limited POC on $P^{(1)}_{s,t}$ and $P^{(2)}_{s,t}$.

The band-limited POC output $R_{s,t}$ is defined by

$$R_{s,t}(m,n) = \frac{1}{(2K+1)^2} \sum_{k=-K}^{K} \sum_{l=-K}^{K} e^{j\Delta\theta(k,l)}\, e^{j\frac{2\pi km}{2K+1}} e^{j\frac{2\pi ln}{2K+1}}, \tag{1}$$

where K is the effective range of frequency spectrum, and $\Delta\theta(k,l)$ denotes the phase difference between the 2D discrete Fourier transforms of $P^{(1)}_{s,t}$ and $P^{(2)}_{s,t}$.

As a result, we can obtain a set $\{R_{s,t}\}$ with $s = 1, \cdots, 3$ and $t = 1, \cdots, 8$ (or $t = 1, \cdots, 4$ for the last level). An example of $\{R_{s,t}\}$ is illustrated in Fig. 3, where we compare the two images of the same building shown at the bottom row. If the two images to be compared are indeed highly correlated, then there should be a consensus among the outputs $\{R_{s,t}\}$ of band-limited POC. We assume that most of the 20 outputs should have common characteristics, and hence we aggregate the outputs $\{R_{s,t}\}$ by taking location-wise average over s, t to obtain $\hat{R} = \frac{1}{20}\sum_{s,t} R_{s,t}$.

An example of $\hat{R}$ is shown in Fig. 3. It can be observed that $\hat{R}$ contains a significant peak, and the location of the peak corresponds to the most significant translation between the two images. Fig. 4 illustrates more results of comparing images by the feature patches using the band-limited POC function. In our task of scene-based image clustering, we use the peak value of $\hat{R}$ as the similarity score for comparing any two input images.

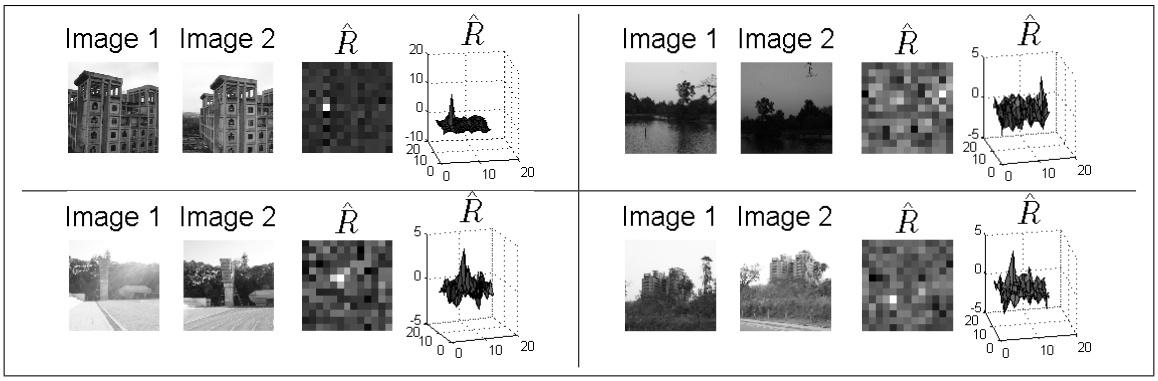

Fig. 4. More examples of comparing images by their feature patches using the band-limited POC function.

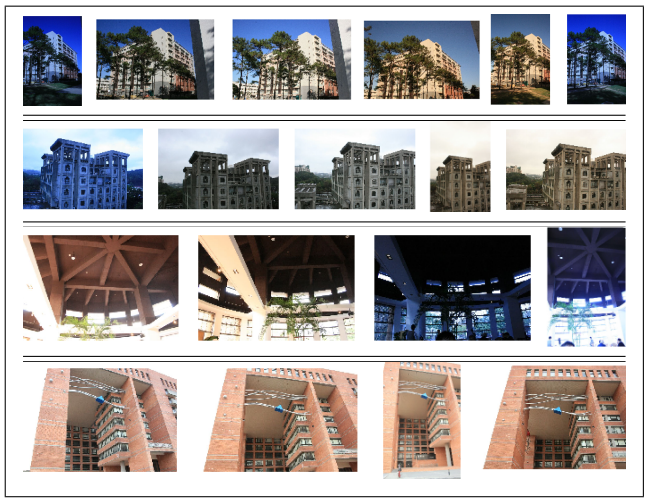

Fig. 5. Four of the clusters obtained by our method on a dataset containing 163 images of 25 different scenes (the NTHU dataset).

C. Clustering

Given an image collection, we use the band-limited POC function to compute a similarity matrix on all pairs of images. Each element of the similarity matrix consists of the similarity score $\hat{R}$ between the corresponding pair of images. We may then apply some standard clustering algorithm such as k-means or spectral clustering to the similarity matrix, and obtain the clusters of similar scenes. In this work, we choose to use *affinity propagation* [3] for clustering. Affinity propagation is very efficient, and is also convenient to use in that the number of clusters can be automatically determined rather than being specified beforehand like k-means. Some example results of grouping 163 images of 25 different scenes are shown in Fig. 5.

III. Evaluations

Four datasets are used to evaluate the performance of our method: NTHU, Golden Temple, Colosseum Rome, Trevi Fountain. The NTHU dataset is created by ourselves and the other three are downloaded from Flickr using their names as the search keywords. The size of each dataset is listed in Table I. Since the aim of our method is to provide initial clustering for narrowing down the search range of SIFT matching, we use the results of exhaustive SIFT matching on the whole dataset as the 'ground truth'. For SIFT, images are resized to 769×512 pixels. We use Lowe's C implementation of SIFT [8] to do keypoint extraction and matching. The number of keypoints extracted in each image is around 700 to 1200. We define the SIFT-based similarity score between two images as the number of matched keypoints found in the image pair. More matched keypoints being found implies that the two images are more similar. The timing results of computing the similarity matrices using SIFT matching and our band-limited POC matching are summarized in Table I (in boldface). It can be seen that extracting SIFT keypoints is fast but matching SIFT keypoints is very time-consuming. Our method is about 30 times faster than SIFT matching at computing the similarity matrices. The experiments are done on a 2.8GHz PC. Note that our method currently is implemented in Matlab and the computation time may be further improved by C implementation. Figs. 6-8 show some more results of grouping the Golden Temple, Colosseum Rome, and Trevi Fountain datasets.

Based on the similarity scores derived from exhaustive SIFT matching over the whole dataset (the 'ground truth'), we may analyze the neighborhood of each image, and see if the neighborhood defined by our method is similar to the neighborhood defined by SIFT matching. More specifically, we evaluate the performance of our method by measuring the probability of observing *any of the k'-nearest neighbors suggested by SIFT* within *the k-nearest neighbors suggested by our method*. We call this probability the nearest-neighbor recall rate. The evaluations of our method on the four datasets are illustrated in Fig. 9 (red solid lines). For comparison, we also show in Fig. 9 (blue dashed lines) the recall rates yielded by computing the similarities using the gist descriptor of 4×4 spatial bins with $8, 8, 4$ orientations at three scales. Although the gist descriptors are easy to compute and very fast to match (*e.g.*, overall 150 sec for Trevi Fountain), its nearest-neighbor recall rates are not good. The results imply that the image neighborhoods obtained by our method are more consistent with SIFT. Our method is more suitable for initial clustering of SIFT matching as far as the recall rate is concerned.

Fig. 6. Four of the clusters obtained by our method on the Golden Temple dataset.

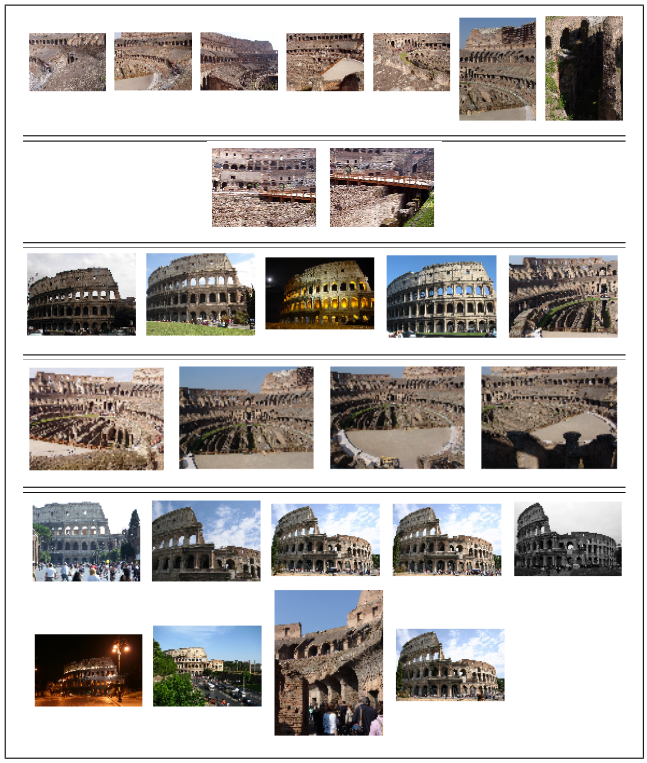

Fig. 7. Five of the clusters obtained by our method on the Colosseum Rome dataset.

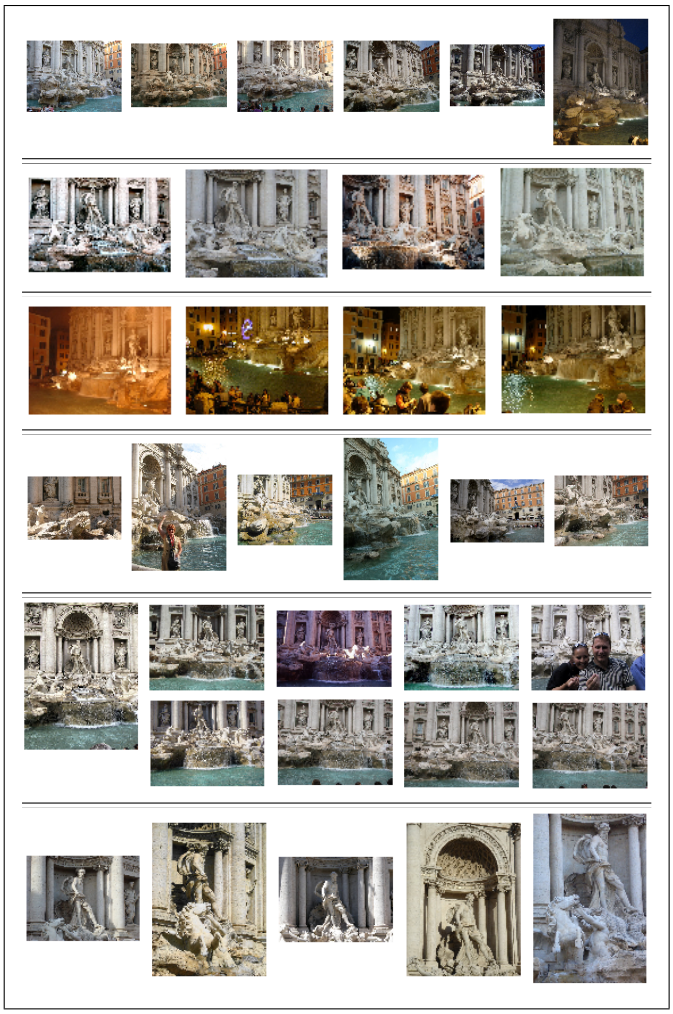

Fig. 8. Six of the clusters obtained by our method on the Trevi Fountain dataset.

IV. Conclusion

We have presented a new method of computing image similarities for scene grouping. The proposed method incorporates an effective image representation that can model global and local features in an image, and the representation enables the use of phase-only correlation (POC) for measuring the similarity between two images. Owing to the shift-invariant property of POC functions, our method can well handle image translation. It is clear that there is a trade-off between using local-feature matching for higher accuracy and using global representation for greater efficiency. Through the experimental evaluations, we have shown that our method provides a useful perspective on how to effectively integrate local and global information for matching and grouping scenes.

Acknowledgment. P.-C. Su and H.-T. Chen were supported in part by grants NTHU 100F2242EA and NSC 98-2221-E-007-083-MY3.

TABLE I

THE TIMING RESULTS OF COMPUTING THE SIMILARITY MATRICES USING SIFT MATCHING (LOWE'S C IMPLEMENTATION) AND OUR BAND-LIMITED POC MATCHING (OUR MATLAB IMPLEMENTATION). EXTRACTING SIFT KEYPOINTS IS FAST BUT MATCHING SIFT KEYPOINTS IS VERY TIME-CONSUMING. OUR METHOD IS ABOUT 30 TIMES FASTER THAN SIFT MATCHING AT COMPUTING THE SIMILARITY MATRICES.

Dataset	# of images	SIFT extraction	SIFT similarity	Gabor Filters	POC similarity
NTHU	163	61 sec	**0.5 hr**	223 sec	**93 sec**
Golden Temple	226	113 sec	**1.3 hr**	313 sec	**180 sec**
Colosseum Rome	267	130 sec	**2.9 hr**	372 sec	**251 sec**
Trevi Fountain	623	318 sec	**18 hr**	868 sec	**1368 sec**

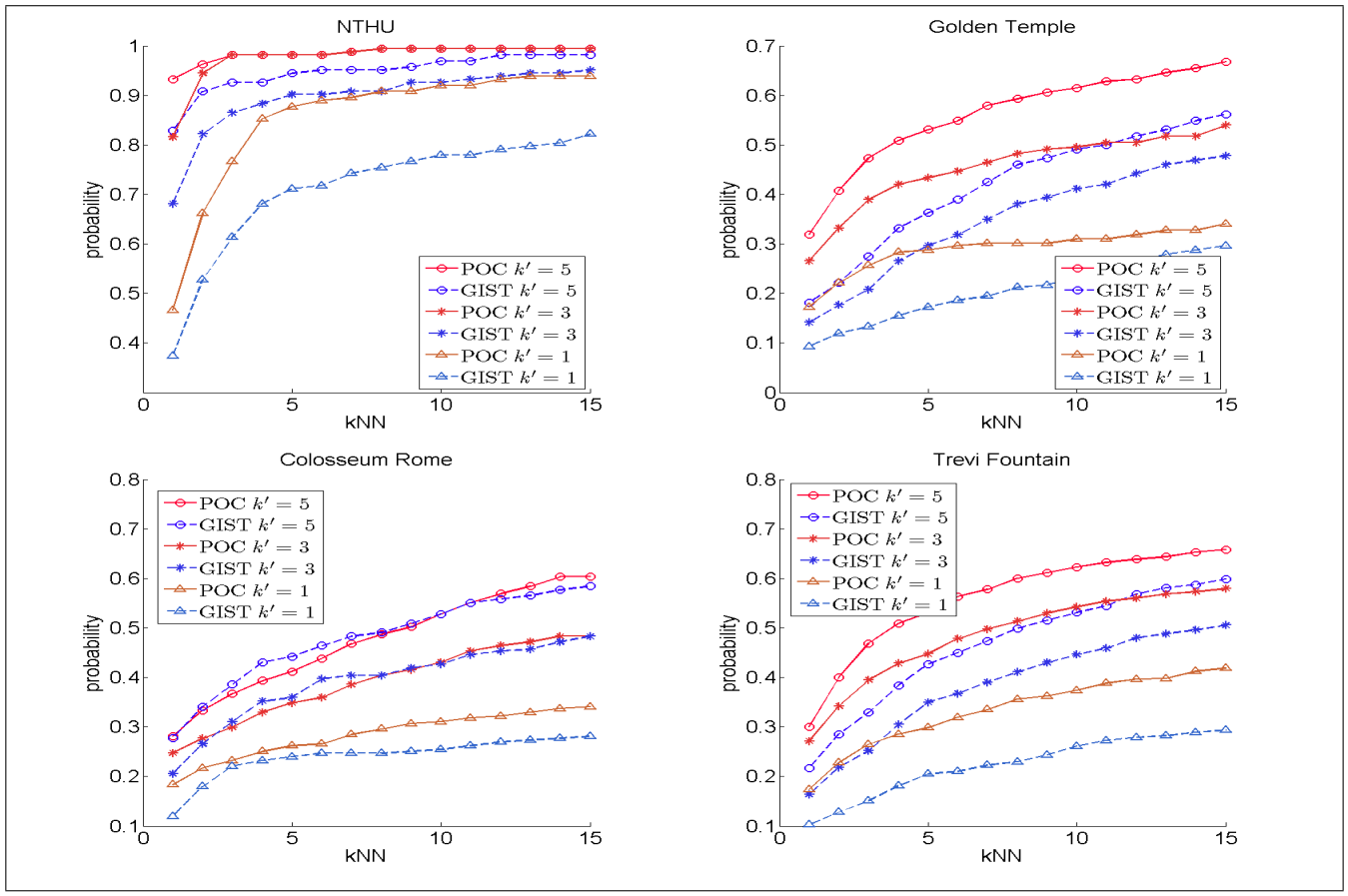

Fig. 9. Nearest-neighbor recall rates of the four datasets. The results of our method are plotted as red solid lines, and the results of gist are plotted as blue dashed lines. Please see the description in Section III for the meanings of nearest-neighbor recall rates, k', and kNN. Note that the Colosseum Rome dataset is hard since it contains many ambiguous images of similar brick walls.

REFERENCES

[1] S. Agarwal, Y. Furukawa, N. Snavely, B. Curless, S. M. Seitz, and R. Szeliski. Reconstructing rome. *IEEE Computer*, 43(6):40–47, 2010.

[2] H. Amirshahi, S. Kondo, K. Ito, and T. Aoki. An image completion algorithm using occlusion-free images from internet photo sharing sites. *IEICE Transactions*, 91-A(10):2918–2927, 2008.

[3] B. J. Frey and D. Dueck. Mixture modeling by affinity propagation. In *NIPS*, 2005.

[4] J. Hays and A. A. Efros. Scene completion using millions of photographs. *Commun. ACM*, 51(10):87–94, 2008.

[5] K. Ito, H. Nakajima, K. Kobayashi, T. Aoki, and T. Higuchi. A fingerprint matching algorithm using phase-only correlation. *IEICE Transactions*, 87-A(3):682–691, 2004.

[6] K. Ito, A. Nikaido, T. Aoki, E. Kosuge, R. Kawamata, and I. Kashima. A dental radiograph recognition system using phase-only correlation for human identification. *IEICE Transactions*, 91-A(1):298–305, 2008.

[7] X. Li, C. Wu, C. Zach, S. Lazebnik, and J.-M. Frahm. Modeling and recognition of landmark image collections using iconic scene graphs. In *ECCV (1)*, pages 427–440, 2008.

[8] D. G. Lowe. Distinctive image features from scale-invariant keypoints. *International Journal of Computer Vision*, 60(2):91–110, 2004.

[9] K. Miyazawa, K. Ito, T. Aoki, K. Kobayashi, and H. Nakajima. An effective approach for iris recognition using phase-based image matching. *IEEE Trans. Pattern Anal. Mach. Intell.*, 30(10):1741–1756, 2008.

[10] A. Oliva and A. Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *International Journal of Computer Vision*, 42(3):145–175, 2001.

[11] J. Sivic, B. Kaneva, A. Torralba, S. Avidan, and W. T. Freeman. Creating and exploring a large photorealistic virtual space. In *First IEEE Workshop on Internet Vision, associated with CVPR*, 2008.

[12] N. Snavely, R. Garg, S. M. Seitz, and R. Szeliski. Finding paths through the world's photos. *ACM Trans. Graph.*, 27(3), 2008.

[13] N. Snavely, S. M. Seitz, and R. Szeliski. Photo tourism: exploring photo collections in 3d. *ACM Trans. Graph.*, 25(3):835–846, 2006.

[14] N. Snavely, S. M. Seitz, and R. Szeliski. Modeling the world from internet photo collections. *International Journal of Computer Vision*, 80(2):189–210, 2008.