View metadata, citation and similar papers at core.ac.uk

brought to you by CORE

provided by Institutional Knowledge at Singapore Management University

Singapore Management University
Institutional Knowledge at Singapore Management University

Research Collection School Of Information Systems | School of Information Systems

7-2016

Learning compact visual representation with canonical views for robust mobile landmark search

Lei ZHU

Jialie SHEN
Singapore Management University, jlshen@smu.edu.sg

Xiaobai LIU

Liang XIE

Liqiang NIE

Follow this and additional works at: https://ink.library.smu.edu.sg/sis_research


Part of the Databases and Information Systems Commons

Citation

ZHU, Lei; SHEN, Jialie; LIU, Xiaobai; XIE, Liang; and NIE, Liqiang. Learning compact visual representation with canonical views for robust mobile landmark search. (2016). *Proceedings of the Twenty-Fifth International Joint Conference on Artificial Intelligence (IJCAI-16): New York, July 9-15, 2016*. 3959-3965. Research Collection School Of Information Systems.
Available at: https://ink.library.smu.edu.sg/sis_research/3544

This Conference Proceeding Article is brought to you for free and open access by the School of Information Systems at Institutional Knowledge at Singapore Management University. It has been accepted for inclusion in Research Collection School Of Information Systems by an authorized administrator of Institutional Knowledge at Singapore Management University. For more information, please email libIR@smu.edu.sg.

Learning Compact Visual Representation with Canonical Views for Robust Mobile Landmark Search

Lei Zhu[†], Jialie Shen[†*], Xiaobai Liu[§], Liang Xie[‡], Liqiang Nie[♮]
[†]School of Information Systems, Singapore Management University, Singapore
[§]Department of Computer Science, San Diego State University, USA
[‡]Department of Mathematics, Wuhan University of Technology, China
[♮]School of Computer Science and Technology, Shandong University, China
{leizhu0608, jialie, xbliu.lhi, whutxl, nieliqiang}@gmail.com

Abstract

Mobile Landmark Search (MLS) recently receives increasing attention. However, it still remains unsolved due to two important issues. One is high bandwidth consumption of query transmission, and the other is the huge visual variations of query images. This paper proposes a Canonical View based Compact Visual Representation (2CVR) to handle these problems via novel three-stage learning. First, a submodular function is designed to measure visual representativeness and redundancy of a view set. With it, canonical views, which capture key visual appearances of landmark with limited redundancy, are efficiently discovered with an iterative mining strategy. Second, multimodal sparse coding is applied to transform multiple visual features into an intermediate representation which can robustly characterize visual contents of varied landmark images with only fixed canonical views. Finally, compact binary codes are learned on intermediate representation within a tailored binary embedding model which preserves visual relations of images measured with canonical views and removes noises. With 2CVR, robust visual query processing, low-cost of query transmission, and fast search process are simultaneously supported. Experiments demonstrate the superior performance of 2CVR over several state-of-the-art methods.

1 Introduction

With the rapid growth of social multimedia and mobile devices, tremendous amount of landmark images have been generated and disseminated in popular social networks. Mobile Landmark Search (MLS) is gaining its importance and increasingly becomes one of the most important techniques to pervasively and intelligently access knowledge about the landmarks of interest [Ji *et al.*, 2012; Chen *et al.*, 2014; Cheng and Shen, 2016].

Mobile devices generally suffer from limited computational power, short battery lifetime, and inefficient or less reliable wireless communication channel. Consequently, a client-server structure is one of the most popular architecture paradigms in existing MLS systems, where query is captured and submitted by mobile devices, computation-intensive search is performed on remote sever with rich computing resources. Since wireless bandwidth is limited, how to generate a compact signature for query to achieve low bit consumption data transmission becomes vital important. On the other hand, visual splendour of a landmark can be photographed by multiple tourists under various circumstances (e.g. a wide sampling of positions, viewpoints, focal lengths, various weather conditions or illuminations.). Besides, landmarks could be comprised of a wide range of regions. In this case, the images taken for different sub-spots in these landmarks may appear with more visual diversity [Zhu *et al.*, 2015a]. All the characteristics of landmark inevitably make the visual appearances of query images very diverse, thus posing great challenges on MLS search system.

Hashing [Wang *et al.*, 2016; Zhu *et al.*, 2015c; Xie *et al.*, 2016b] aims at learning compact binary codes with Hamming distance computation. Thus, it can significantly reduce transmission cost and speedup the search process. Hence, hashing is a promising scheme to support large scale landmark image indexing and retrieval. However, the most existing hashing strategies developed for MLS are based on unimodal visual-words based features and generally suffer from 1) limited discriminative capability and 2) poor robustness against visual variations [Chen *et al.*, 2014; Zhou *et al.*, 2014]. Although general multimodal hashing techniques [Kim and Choi, 2013; Song *et al.*, 2013; Liu *et al.*, 2014; Shen *et al.*, 2015] improve discriminative capability with multimodal fusion [Zhu *et al.*, 2015b; Xie *et al.*, 2016a], they are based on simple low-level feature fusion and enjoy less robustness against visual variations of the query landmark images captured by mobile devices.

This paper proposes a Canonical View based Compact Visual Representation (2CVR) to support efficient and robust MLS. We also develop canonical views as the views which capture key visual appearances of landmark with limited redundancy. Based on them, an arbitrary image can be robustly represented using a specific canonical view or the cross-scenery of multiple particular canonical views. Ac-

*Jialie Shen is corresponding author.

cordingly, varied visual contents of landmark can be effectively characterized using their visual correlations to only fixed canonical views. Through encoding these relations into the binary codes, visual variations of queries can be robustly modeled. Furthermore, the low-cost query transmission and fast search can be well supported.

The contributions of this paper are summarized as follows:

1. A submodular function is designed to measure visual representativeness and redundancy of a view set. With it, an iterative mining strategy is proposed to efficiently identify canonical views of landmarks using multiple modalities. Theoretical analysis demonstrates that it can achieve near-optimal solutions.
2. A novel intermediate representation generated by multi-modal sparse coding is proposed to robustly characterize the visual contents of varied landmark images. It provides a natural and effective connection between the canonical views and binary embedding model.
3. A binary embedding model tailored for canonical views is proposed to preserve visual relations of images into binary codes and thus support efficient MLS with great robustness.

2 Related Work

2.1 Mobile Landmark Search

Ji *et al*. [2012] present a Location Discriminative Vocabulary Coding (LDVC) to compress Bag-of-Visual-Words (BoVW) with location awareness. Duan *et al*. [2013] explore multiple information sources to extract compact landmark image descriptor. Chen *et al*. [2014] develop a soft Bag-of-Visual Phrase (BoVP) to learn category-dependent visual phrases, by capturing co-occurrence features of neighbouring visual-words. Zhou *et al*. [2014] propose Scalable Cascaded Hashing (SCH) to achieve codebook-free large-scale MLS.

All the techniques mentioned above learn compact binary codes from only visual-words based features, without considering complement information from other visual modalities. This limitation makes the generated codes less discriminative.

2.2 Multimodal Hashing

Multimodal hashing has been emerging as a promising technique to generate compact binary code based on multiple features. The earliest study on this topic is Composite Hashing with Multiple Information Sources (CHMIS) [Zhang *et al.*, 2011]. It just post-integrates linear output of features and fails to fully exploit their correlations. Kim *et al*. [2013] present Multi-View Anchor Graph Hashing (MVAGH) by extending Anchor Graph Hashing (AGH) [Liu *et al.*, 2011] to gain robust representation cross multiple images. Song *et al*. [2013] develop Multiple Feature Hashing (MFH). By using the learned hashing hyper-plane, MFH concatenates all the features into a single vector and then maps it into binary codes. Liu *et al*. [2014] propose Compact Kernel Hashing (CKH) by formulating the similarity preserving problem with optimal linearly-combined multiple kernels corresponding to different features. More recently, Multi-View Latent Hashing (MVLH) [Shen *et al.*, 2015] is proposed to incorporate multimodal features in binary representation learning by discovering the latent factors shared by multiple views. Further, distinguished from the existing methods, 2CVR aims to learn informative canonical views by capturing key characteristics of landmarks. Consequently, the generated binary codes can enjoy desirable robustness.

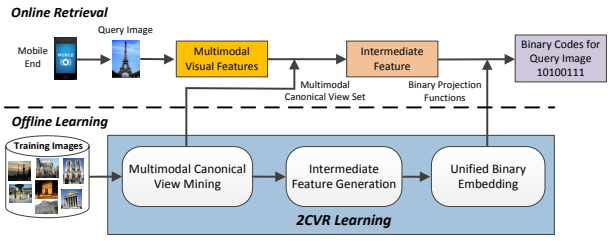


Figure 1: Framework overview of 2CVR learning.

3 The Proposed Method

3.1 Overview

Figure 1 illustrates the basic framework of 2CVR learning. Given a set of landmark images, the computation of compact representation 2CVR, consists of three major steps: Firstly, multimodal canonical view mining is proposed in order to efficiently discover a compact but informative canonical view set from noisy landmark image collections to capture key visual appearances. Then, in order to robustly model diverse visual contents, an intermediate representation is generated by computing multimodal sparse reconstruction coefficients between image and canonical view. Finally, compact binary codes are learned by preserving the discovered visual relations measured on canonical views.

3.2 Multimodal Canonical View Mining

It is common that, in practice, several sceneries of landmarks are frequently photographed and disseminated by different tourists. These views of landmarks are considered as canonical views in this work and applied to cope with visual variations of query images captured by mobile devices. In this subsection, we first formally define mathematical properties of canonical view set, and then give an efficient submodular function based mining algorithm for discovery.

Definition 1. *Let $\mathcal{I}$ denote image space, $\mathcal{I} = \{\mathcal{I}_n\}_{n=1}^{N}$, N is the number of database images. Let $\mathcal{L}$ denote landmark space, $\mathcal{L} = \{\mathcal{L}_m\}_{m=1}^{M}$, M is the number of landmarks in database. $\mathcal{L}_m$ is defined as a set of images which are recorded at the nearby position of the m_{th} landmark. Let $\mathcal{V}$ denote a view set of $\mathcal{L}$. It is defined as a set of images $\{v_i\}_{i=1}^{|\mathcal{V}|}$ belonging to $\mathcal{I}$, $\mathcal{V} \subseteq \mathcal{I}$, $|\mathcal{V}| \ll |\mathcal{I}|$.*

Definition 2. *Let $\mathtt{Rep}(\mathcal{V})$ denote the visual representativeness of view set $\mathcal{V}$ over $\mathcal{L}$. It is defined as $\mathtt{Rep}(\mathcal{V}) = \sum_{v_i \in \mathcal{V}} \mathtt{Rep}(v_i) = \sum_{v_i \in \mathcal{V}} \sum_{v_j \in \mathcal{I}, i \neq j} g_{ij}$, g is the function which measures the feature similarity of two image views, g_{ij} is short for $g(v_i, v_j)$. Let $\mathtt{Red}(\mathcal{V})$ denote the visual redundancy of view set $\mathcal{V}$. It is defined as $\mathtt{Red}(\mathcal{V}) = \sum_{v_i \in V} \mathtt{Red}(v_i) = \sum_{v_i, v_j \in \mathcal{V}, i \neq j} g_{ij}$.*

Definition 3. *Let $\mathcal{C}$ denote the canonical view set of $\mathcal{L}$. The views involved in $\mathcal{C}$ can comprehensively represent diverse visual contents of landmark, and meanwhile, have less visual redundancy. In this paper, it is defined as $\mathcal{C} = \arg\max_{\mathcal{V}\subseteq\mathcal{I},|\mathcal{V}|=T} h(\mathcal{V}), h(\mathcal{V}) = \texttt{Rep}(\mathcal{V}) - \texttt{Red}(\mathcal{V})$, T is cardinality of canonical view set.*

Lemma 1. *$h(\mathcal{V})$ is submodular function. That is, $\forall\mathcal{V}_1 \subseteq \mathcal{V}_2 \subseteq \mathcal{V}, \forall v_j \notin \mathcal{V}$, $h(\mathcal{V}_1 \cup v_j) - h(\mathcal{V}_1) \geq h(\mathcal{V}_2 \cup v_j) - h(\mathcal{V}_2)$.*

Proof $\texttt{Rep}(\mathcal{V}_1 \cup v_j) - \texttt{Rep}(\mathcal{V}_1) = \texttt{Rep}(\mathcal{V}_2 \cup v_j) - \texttt{Rep}(\mathcal{V}_2) - (\texttt{Red}(\mathcal{V}_1 \cup v_j) - \texttt{Red}(\mathcal{V}_1)) = -2 \sum_{v_i \in \mathcal{V}_1 \setminus v_j} g_{ij}$

$\geq -2 \sum_{v_i \in \mathcal{V}_2 \setminus v_j} g_{ij} = -(\texttt{Red}(\mathcal{V}_2 \cup v_j) - \texttt{Red}(\mathcal{V}_2))$

$\Rightarrow\ h(\mathcal{V}_1 \cup v_j) - h(\mathcal{V}_1) \geq h(\mathcal{V}_2 \cup v_j) - h(\mathcal{V}_2)$

Lemma 2. *$h(\mathcal{V})$ is monotonically nondecreasing function. That is, $\forall\mathcal{V}_1 \subseteq \mathcal{V}_2 \subseteq \mathcal{V}$, $h(\mathcal{V}_1) \leq h(\mathcal{V}_2)$.*

Proof $h(\mathcal{V}) = \sum_{v_i \in \mathcal{V}} \sum_{v_j \in \mathcal{I}\setminus\mathcal{V}, i\neq j} g_{ij}$

$\Rightarrow\ h(\mathcal{V}_1) = \sum_{v_i \in \mathcal{V}_1} \sum_{v_j \in \mathcal{V}_2\setminus\mathcal{V}_1, i\neq j} g_{ij} + \sum_{v_i \in \mathcal{V}_1} \sum_{v_j \in \mathcal{I}\setminus\mathcal{V}_2, i\neq j} g_{ij}$

$\Rightarrow\ h(\mathcal{V}_2) = \sum_{v_i \in \mathcal{V}_2\setminus\mathcal{V}_1} \sum_{v_j \in \mathcal{I}\setminus V_2, i\neq j} g_{ij} + \sum_{v_i \in \mathcal{V}_1} \sum_{v_j \in \mathcal{I}\setminus\mathcal{V}_2, i\neq j} g_{ij}$

Since in our case, $|\mathcal{V}_1|, |\mathcal{V}_2| \ll \mathcal{I}$

$\Rightarrow \sum_{v_i \in \mathcal{V}_1} \sum_{v_j \in \mathcal{V}_2\setminus\mathcal{V}_1, i\neq j} g_{ij} \leq \sum_{v_i \in \mathcal{V}_2\setminus\mathcal{V}_1} \sum_{v_j \in \mathcal{I}\setminus\mathcal{V}_2, i\neq j} g_{ij}$

$\Rightarrow\ h(\mathcal{V}_1) \leq h(\mathcal{V}_2)$

As indicated in **Definition 3**, to discover optimal canonical view set, the function $h(\mathcal{V})$ should be maximized. However, since $h(\mathcal{V})$ is submodular function, the maximization of it is a NP-complete optimization problem. Fortunately, $h(\mathcal{V})$ is monotonically nondecreasing with a cardinality constraint. Canonical views can be discovered near optimally by greedy strategy as following steps

Step 1*: Extract visual feature for all landmark images.*
Step 2*: Set canonical view set as empty, $\mathcal{C} = \emptyset$.*
Step 3*: Iterate the following two steps for T times.*
3.1*: Compute $\texttt{diff}(\mathcal{I}_n) = h(\mathcal{C} \cup \mathcal{I}_n) - h(\mathcal{I}_n)$ for each landmark image $\mathcal{I}_n \in \mathcal{I}$.*
3.2*: Select the image with the maximum $\texttt{diff}$ into $\mathcal{C}$, $\mathcal{I}^* = \arg\max_{I_n \in \mathcal{I}} \texttt{diff}(I_n)$, and simultaneously remove it from $\mathcal{I}$, $\mathcal{C} \leftarrow \mathcal{C} \cup \mathcal{I}^*$, $\mathcal{I} \leftarrow \mathcal{I} \setminus \mathcal{I}^*$.*

Theorem 1. *[Nemhauser* et al.*, 1978] Let $\mathcal{S}^*$ denote the global optimal solution that solves the combinatorial optimization problem $\arg\max_{\mathcal{S}\subseteq\mathcal{L}, |\mathcal{S}|=T} h(\mathcal{S})$, $\mathcal{S}$ denote approximate solution found by the greedy algorithm. If $h(\mathcal{S})$ is nondecreasing submodular function with $h(\emptyset) = 0$, we can have*

$$h(\mathcal{S}) \geq h(\mathcal{S}^*)\frac{\zeta - 1}{\zeta}$$

where ζ refers to the natural exponential.

As validated by **Theorem 1**, this greedy algorithm can achieve a result that is no worse than a constant fraction $\frac{\zeta-1}{\zeta}$ away from the optimal value. The time complexity of canonical view mining is reduced to $O(NT)$. The canonical view discovery process can be completed efficiently. To comprehensively cover visual appearances of landmarks, canonical view mining is performed in multiple modalities, obtaining canonical view set $\{\mathcal{C}^p\}_{p=1}^P$, P is number of modalities. We concatenate features of canonical views and construct a matrix $E^p = [e_1^p, ..., e_T^p] \in \mathbb{R}^{d_p \times T}$ in modality p, d_p is the corresponding feature dimension.

3.3 Intermediate Representation Generation

With the discovered canonical views, an arbitrary recorded landmark image either describes a specific canonical view or the cross-scenery among several particular canonical views. In both cases, visual contents of the image can be principally represented with its relations to several particular canonical views. Motivated by the observations, we calculate multimodal sparse reconstruction coefficients between image and canonical views. And the auto-generated response coefficients construct the dimensions of intermediate representation. The concrete mathematical form is

$$\begin{aligned} \min_{\{Y^p\}_{p=1}^P} & \sum_{p=1}^{P} ||X^p - E^p Y^p||_F^2 + \alpha \sum_{p=1}^{P}\sum_{n=1}^{N} ||d_n^p \otimes y_n^p||_F^2 \\ s.t.\ & 1_T^\top y_n^p = 1, d_n^p = \exp(\frac{\texttt{dist}(x_n^p, E^p)}{\rho}), \forall p, n \end{aligned} \tag{1}$$

where $\alpha > 0$ is a constant factor that adjusts the balance between terms, ρ is set to be the mean of pairwise distances, $\otimes$ denotes the element-wise product, $1_T \in \mathbb{R}^\top$ denotes a column vector with all ones. $X^p = [x_1^p, ..., x_N^p] \in \mathbb{R}^{d_p \times N}$ denotes features of database images in modality p. $Y^p = [y_1^p, ..., y_N^p] \in \mathbb{R}^{T \times N}$ denotes modality-specific canonical view based intermediate representation. Each column has r non-zeros coding coefficients. $\texttt{dist}(x_n^p, E^p) = [\texttt{dist}(x_n^p, e_1^p), ..., \texttt{dist}(x_n^p, e_T^p)]$. The problem in (1) can be efficiently solved by using the Alternating Direction Method of Multipliers (ADMM) [Elhamifar and Vidal, 2013]. After solving, we concatenate the calculated Y^p and construct dimensions of intermediate representation

$$Y = [Y^1; ...; Y^P] \in \mathbb{R}^{TP \times N} \tag{2}$$

Remark. The intermediate representation can effectively characterize diverse visual contents by adaptively adjusting the response coefficients on canonical views. It lays the solid foundation for subsequent binary embedding.

3.4 Binary Embedding Model

Based on the intermediate representation, we design a binary embedding model to learn final binary representation.

Due to approximate canonical view mining and multimodal information integration, the intermediate representation generation inevitably brings about noises and redundancies. It is very important to remove them during binary embedding to avoid disturbance. To achieve this goal, matrix

factorization is applied in this paper to decompose intermediate features into the latent binary bits and guarantee them to be orthogonal to each other. Besides, a graph regularizer is constructed to preserve visual relationships among images. That is, if two landmark images have similar visual distributions on canonical views, they are constrained to be projected to close points in hamming space. Moreover, since queries are out of database and continuously flowing into database when time passes by, we learn linear projection for out-of-sample extension. Therefore, by integrating the aforementioned considerations, the overall binary embedding is formulated as

$$\begin{aligned}\min_{W,V^*} \; & Tr(V^*L(V^*)^{\mathrm{T}}) + \lambda||Y - UV^*||_F^2 + \beta(||V^* - \\ & W^{\mathrm{T}}Y||_F^2 + \gamma||W||_F^2) \\ s.t.\; & V^*(V^*)^{\mathrm{T}} = I_c, V^* \in \{-1,1\}^{c\times N}\end{aligned} \tag{3}$$

where $\lambda, \beta, \gamma > 0$ adjust the balance of terms. $||Y - UV^*||_F^2$ is matrix factorization term, $V^*(V^*)^{\mathrm{T}} = I_c$ is bit orthogonality constraint. $Tr(V^*L(V^*)^{\mathrm{T}})$ is graph regularizer which preserves visual relations of landmark images on canonical views. $Tr(\cdot)$ is trace operation, L is graph Laplacian matrix, which is constructed based on intermediate representation. It measures visual relations of images on canonical views. $||V^* - W^{\mathrm{T}}Y||_F^2 + \gamma||W||_F^2$ is binary projection learning term, $W \in \mathbb{R}^{TP\times c}$ is linear projection matrix to be learned, c is binary code length, $||V^* - W^{\mathrm{T}}Y||_F^2$ is to reduce the loss between binary codes and the projected values. It is worth noting that as linear projection is leveraged, the online mobile landmark search process can be efficient.

When directly imposing $V^* \in \{-1,1\}^{c\times N}$, problem (3) will become NP-hard. Thus we relax this discrete constraint to $V \in \mathbb{R}^{c\times N}$ and derive the following relaxed form

$$\begin{aligned}\min_{W,VV^{\mathrm{T}}=I_c} \; & Tr(VLV^{\mathrm{T}}) + \lambda||Y - UV||_F^2 + \beta(||V - \\ & W^{\mathrm{T}}Y||_F^2 + \gamma||W||_F^2)\end{aligned} \tag{4}$$

The optimal W, U that solves Eq.(4) can be expressed in terms of Y and V. We can derive the following theorem.

Theorem 2. *Let W, U, V and Y be defined as before. Then the optimal W that solves learning problem in (4) is given by $W = (YY^T + \gamma I)^{-1}YV^T, U = YV^T$. The relaxed minimum problem in (4) is equivalent to the following simple problem*

$$\min_{VV^{\mathrm{T}}=I_c} Tr(VAV^T) \tag{5}$$

where $A = L - \lambda Y^TY + \beta(I - Y^TQY), Q = (YY^T + \gamma I)^{-1}$.

Proof. We calculate the derivation of the objective function in Eq.(4) *w.r.t* U and set it to be zero. Thus, we have

$$-2YV^{\mathrm{T}} + 2U = 0 \Rightarrow U = YV^{\mathrm{T}}$$

By replacing U into Eq.(4), we can derive that

$$\begin{aligned}&||Y - UV||_F^2 = Tr((Y - UV)(Y - UV)^{\mathrm{T}}) \\ &= Tr(YY^{\mathrm{T}}) - 2Tr(YV^{\mathrm{T}}U^{\mathrm{T}}) + Tr(UVV^{\mathrm{T}}U^{\mathrm{T}}) \\ &= Tr(YY^{\mathrm{T}}) - 2Tr(YV^{\mathrm{T}}VY^{\mathrm{T}}) + Tr(YV^{\mathrm{T}}VY^{\mathrm{T}}) \\ &= Tr(YY^{\mathrm{T}}) - Tr(YV^{\mathrm{T}}VY^{\mathrm{T}}) \\ &\Rightarrow \min_{VV^{\mathrm{T}}=I_c} \lambda||Y - UV||_F^2 = \min_{VV^{\mathrm{T}}=I_c} -\lambda Tr(YV^{\mathrm{T}}VY^{\mathrm{T}}) \\ &= \min_{VV^{\mathrm{T}}=I_c} -\lambda Tr(VY^{\mathrm{T}}YV^{\mathrm{T}})\end{aligned}$$

Similarly, by calculating the derivation of the objective function in Eq.(4) *w.r.t* W and setting it to be zero, we can have

$$YY^{\mathrm{T}}W - YV^{\mathrm{T}} + \gamma W = 0 \Rightarrow W = (YY^{\mathrm{T}} + \gamma I)^{-1}YV^{\mathrm{T}} \tag{6}$$

Let $Q = (YY^{\mathrm{T}} + \gamma I)^{-1}$, then $W = QYV^{\mathrm{T}}$. By replacing W into Eq.(4), we can derive that

$$||V - W^{\mathrm{T}}Y||_F^2 + \gamma||W||_F^2 = Tr(V(I - Y^{\mathrm{T}}QY)V^{\mathrm{T}})$$

By summing three terms together, we find that the relaxed minimum problem in (4) is equivalent to

$$\begin{aligned}&\min_{VV^{\mathrm{T}}=I_c} Tr(VLV^{\mathrm{T}}) - \lambda Tr(VY^{\mathrm{T}}YV^{\mathrm{T}}) + \beta Tr(V(I - Y^{\mathrm{T}} \\ &(YY^{\mathrm{T}} + \gamma I)^{-1}Y)V^{\mathrm{T}}) = \min_{VV^{\mathrm{T}}=I_c} Tr(VAV^{\mathrm{T}})\end{aligned}$$

where $A = L - \lambda Y^{\mathrm{T}}Y + \beta(I - Y^{\mathrm{T}}QY), Q = (YY^{\mathrm{T}} + \gamma I)^{-1}$. This completes the proof of the theorem. □

It turns out that the problem in Eq.(4) has a close-form solution. The rows of optimal solution (We denote it as $\tilde{V}$) are given as the c eigenvectors with minimal eigenvalues of the matrix A. After that, the optimal binary projection W can be calculated as Eq.(6).

However, solving continuous $\tilde{V}$ will generate binary quantization error. Inspired by [Gong *et al.*, 2013], we could apply orthogonal transformation to rotate $\tilde{V}$ to align with hypercube $\{\pm 1\}^{c\times N}$ as close as possible, and thus reduce the quantization error. But, it would be feasible only if the rotation does not change the minimum of Eq.(5). Fortunately, it can be easily validated that, for an arbitrary orthogonal rotation matrix $\tilde{R}$, we can have $\min_{VV^{\mathrm{T}}=I_c} Tr(VAV^{\mathrm{T}}) = \min_{RVV^{\mathrm{T}}R^{\mathrm{T}}=I_c} Tr(RVAV^{\mathrm{T}}R^{\mathrm{T}})$. Hence, it is possible to learn an orthogonal rotation matrix R which guarantees that $R\tilde{V}$ can simultaneously achieve the minimum objective function value in Eq.(5) and binary embedding errors. Formally, R is learned by

$$\min_{V^*,R} ||V^* - R\tilde{V}||_F^2 \;\; s.t.\; V^* = \{-1,1\}^{c\times N}, RR^{\mathrm{T}} = I_c$$

This reduces to the Orthogonal Procrustes Problem (OPP) [Yu and Shi, 2003]. A local optimal solution can be obtained by alternating minimization between V^* and R. Afterwards, the final binary projection matrix is adjusted as $W = WR$. Given a landmark image q, we first extract intermediate representation Y_q as Eq.(2). Its binary codes are calculated as $V_q^* = \mathtt{sgn}(W^{\mathrm{T}}Y_q)$, where $\mathtt{sgn}(x)$ denotes the sign function. It returns 1 if $x > 0$ and -1 otherwise.

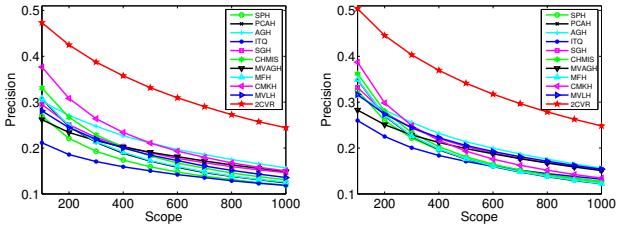
4 Experiments

Experimental Datasets and Setting. Three real landmark datasets are applied for empirical study: *Oxford5K* [Philbin *et al.*, 2007], *Paris6K* [Philbin *et al.*, 2008], and *Paris500K* [Weyand and Leibe, 2013][1]. *Oxford5K* is comprised of 5,062 *Oxford* landmark images in 17 categories. *Paris6K* consists of 6,412 *Paris* landmark images in 12 categories. *Paris500K* contains 41,673 images with clustering ground truth which

[1]In this paper, the maximum number of images in each category is limited to 2000 to avoid bias.

Table 1: mAP of all approaches on three datasets. The best performance in each column is marked with bold.

Methods	*Oxford5K*				*Paris6K*				*Paris500K*			
	32	48	64	128	32	48	64	128	32	48	64	128
SPH	0.2930	0.2989	0.2997	0.3279	0.2868	0.3111	0.3252	0.3449	0.3080	0.3656	0.3948	0.4756
PCAH	0.2837	0.2959	0.2979	0.3151	0.2964	0.3204	0.3288	0.3393	0.3512	0.4168	0.4447	0.5186
AGH	0.3007	0.3162	0.3098	0.3070	0.3276	0.3383	0.3520	0.3276	0.3530	0.3721	0.3872	0.3944
ITQ	0.2878	0.2672	0.2870	0.3105	0.2761	0.2988	0.2960	0.3339	0.2219	0.2820	0.3049	0.3565
SGH	0.3029	0.3160	0.3148	0.3283	0.3220	0.3458	0.3579	0.3725	0.3309	0.3714	0.4071	0.4628
CHMIS	0.2977	0.3092	0.2955	0.3121	0.3247	0.3278	0.3410	0.3697	0.3937	0.4431	0.4684	0.5317
MVAGH	0.3039	0.2997	0.3104	0.3058	0.2521	0.2733	0.2864	0.3197	0.2916	0.3200	0.3384	0.3631
MFH	0.2728	0.2882	0.3021	0.3203	0.2909	0.3111	0.3191	0.3597	0.3539	0.4200	0.4461	0.5203
CMKH	0.2947	0.3152	0.2983	0.3041	0.3337	0.3449	0.3408	0.3426	0.4114	0.4547	0.5095	0.5497
MVLH	0.2895	0.3286	0.3008	0.3232	0.3320	0.3355	0.3366	0.3977	0.3168	0.3385	0.3827	0.4163
2CVR	**0.3176**	**0.3371**	**0.3458**	**0.3586**	**0.3644**	**0.3856**	**0.4022**	**0.4173**	**0.4480**	**0.5191**	**0.5645**	**0.6139**



(a) 64 bits (b) 128 bits

Figure 2: *Precision-Scope* curves on *Paris500K*.

describe 79 landmarks. For *Oxford5K* and *Paris6K*, 10%, 20%, and 70% images are used as query images, training images, and database images, respectively. For *Paris500K*, the corresponding ratios are 10%, 10%, and 80%. Both query and database images appear with great visual diversity in three datasets. Each image is represented by features in 5 visual modalities: 81-D Color Moments (CM) [Yu *et al.*, 2002], 58-D Local Binary Pattern (LBP) [Wang *et al.*, 2009], 80-D Edge Direction Histogram (EDH) [Park *et al.*, 2000], and 1,000-D BoVW[2] [Sivic and Zisserman, 2003], 512-D GIST [Oliva and Torralba, 2001].

Evaluation Metrics. In experimental study, mean Average Precision (mAP) [Liu *et al.*, 2014; Shen *et al.*, 2015] is adopted as main evaluation metric. The number of returned images is set as 100 to collect experimental results. Furthermore, *Precision-Scope* curve is also reported to demonstrate the retrieval performance variations with the number of retrieved images. Binary code length c on all datasets is varied in the range of $\{32, 48, 64, 128\}$, and the search scope is ranged from 100 $\sim$1000 with step size 100.

Competitors. Since 2CVR is learned by unsupervised learning on multiple visual modalities, we compare it with state-of-the-art unsupervised multimodal binary representation generation approaches. They include[3]: CHMIS [Zhang *et al.*, 2011], MVAGH [Kim and Choi, 2013], MFH [Song *et al.*, 2013], CMKH [Liu *et al.*, 2014], and MVLH [Shen *et al.*, 2015]. Besides, we also compare 2CVR with several state-of-the-art unimodal approaches: SPH [Weiss *et al.*, 2008], PCAH [Wang *et al.*, 2010], AGH [Liu *et al.*, 2011], ITQ [Gong *et al.*, 2013], and SGH [Jiang and Li, 2015]. For them, multimodal features are concatenated into a unified vector for subsequent learning. The involved parameters of the compared approaches are strictly adjusted to report the maximum performance according to the relevant literature. For 2CVR, the best performance is achieved when $\lambda = 1, \beta = 10^4, \gamma = 10^4$. The best canonical view size T is set to 100 on *Oxford5K* and *Paris6K*, and 300 on *Paris500K*. The best number of nearest canonical views r in Eq.(1) is set to 70 on *Oxford5K* and *Paris6K*, and 200 on *Paris500K*. α in Eq.(1) is set to 10^{-4} to maximize the performance.

Performance Comparison. We report mAP results and *Precision-Scope* curves on *Paris500K* of all approaches in Table 1 and Figure 2, respectively. From the presented results, we can easily find that 2CVR consistently outperforms the competitors on all code lengthes and datasets. It is interesting to find that, even with less binary bits, 2CVR can achieve higher mAP than many competitors with longer binary codes. Further, Figure 2 shows that, on *Paris500K* and 128 bits, the precision gain of 2CVR over the second best approach is more than 10%, and it becomes larger when more images are returned. Moreover, we observe that performance improvement on *Paris500K* is more than that obtained on *Oxford5K* and *Paris6K*. Since images in *Paris500K* have more diverse visual appearances, this experimental phenomenon validates the desirable property of 2CVR on accommodating the visual variations. Finally, we observe that the retrieval performance of 2CVR is steadily improved when binary code length increases. However, we don't gain similar observations for many approaches studied in this experimental study. This is because 2CVR ensures bit orthogonality constraint in binary code learning. The design guarantees the learned binary bits to achieve less information redundancy. More binary bits will enable 2CVR to gain higher discriminative capability.

[2] 128-D SIFT [Lowe, 2004] is employed as local descriptor.

[3] For CHMIS, MFH, CMKH, SPH, PCAH, AGH, ITQ, and SGH, implementation codes of them are provided by the authors. For MVAGH and MVLH, we implement them according to the relevant literature.

Table 2: Canonical views improve the robustness of 2CVR. 2CVR-II denotes binary code learning without canonical views.

Methods	*Oxford5K*				*Paris6K*				*Paris500K*			
	32	48	64	128	32	48	64	128	32	48	64	128
2CVR-II	0.3080	0.3208	0.3230	0.3188	0.3357	0.3525	0.3722	0.3815	0.1949	0.2527	0.3037	0.4573
2CVR	**0.3176**	**0.3371**	**0.3458**	**0.3586**	**0.3644**	**0.3856**	**0.4022**	**0.4173**	**0.4480**	**0.5191**	**0.5645**	**0.6139**

Table 3: Effects of canonical view mining in multiple modalities.

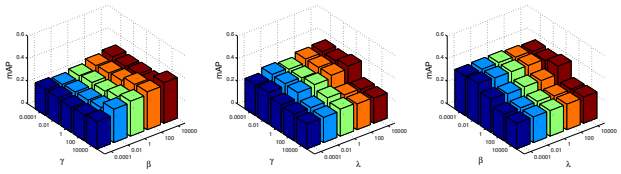
Methods	*Oxford5K*				*Paris6K*				*Paris500K*			
	32	48	64	128	32	48	64	128	32	48	64	128
CM	0.2159	0.2130	0.2266	0.2224	0.2018	0.2179	0.2167	0.2262	0.1665	0.1932	0.2107	0.2440
LBP	0.2749	0.2703	0.3873	0.3008	0.2359	0.2524	0.2503	0.2638	0.2682	0.3074	0.3314	0.3761
EDH	0.2537	0.2681	0.2798	0.2696	0.2361	0.2471	0.2578	0.2735	0.2879	0.3281	0.3588	0.4099
BOVW	0.3059	0.3176	0.3220	0.3408	0.3480	0.3548	0.3630	0.3820	0.4217	0.4618	0.4953	0.5445
GIST	0.2688	0.2686	0.2689	0.2869	0.2813	0.2834	0.2862	0.3114	0.3376	0.3956	0.4408	0.4981
2CVR	**0.3176**	**0.3371**	**0.3458**	**0.3586**	**0.3644**	**0.3856**	**0.4022**	**0.4173**	**0.4480**	**0.5191**	**0.5645**	**0.6139**

Canonical View or Not? To see how the canonical view mining can benefit compact representation learning, we first compare the performance of 2CVR with the one which performs binary embedding Eq.(3) directly on raw concatenated multiple low-level features. Table 2 presents the detailed comparative results. From it, we easily find that 2CVR can consistently yield better performance. On three datasets, the maximum search precision improvements reach about 3%, 4%, and 26%, respectively. The performance improvement is attributed to the fact that, canonical views capture key visual contents of landmarks, diverse visual contents can be robustly accommodated, and thus binary codes have better robustness.

Then, we investigate the effects of canonical view mining in multiple modalities. We compare the performance of 2CVR with the approaches that perform binary embedding Eq.(3) on only unimodal canonical view set. We denote them directly with the corresponding modality names: CM, LBP, EDH, BOVW, and GIST respectively. Table 3 presents the main results. It demonstrates that 2CVR can achieve the best performance. The reason is that, with multimodal learning, canonical view set can cover more visual variations and thus 2CVR can enjoy better robustness. All the above results clearly demonstrate that 2CVR adopts a reasonable strategy by employing canonical views for MLS.

Finally, we validate the effects of the proposed submodular function based canonical view selection approach. We compare the performance of 2CVR with two variants of method which discover canonical views by random selection and k-means, respectively. The detailed comparison results are presented in Table 4. It can be easily observed that 2CVR can consistently achieve better performance. These results demonstrate the effectiveness of submodular function on discovering canonical views of landmarks.

Parameter Study. We investigate the performance variations of 2CVR with parameters. Due to the space limit, we observe the performance variations with λ, β, and γ. They are used in Eq.(3) to play trade-off between terms. We observe the performance variations with respect to two parameters while fixing the remaining one parameter. We report results on *Oxford5K* when binary code length is 128. Similar results can be found on other code lengths and datasets. Figure 3 presents results. From it, we can find the performance of 2CVR is stable on a range of parameters.



(a) λ is fixed to 1 (b) β is fixed to 10^4 (c) γ is fixed to 10^4

Figure 3: mAP variations with parameters on *Oxford* when binary code length is 128. This figure is best viewed with pdf magnification.

5 Conclusions

This paper proposes 2CVR to support efficient and robust mobile landmark search. The design of 2CVR has basis in reality that only canonical views are frequently photographed and disseminated by different tourists, and they naturally provide effective visual representation basis of landmarks. Experimental results on three real landmark datasets demonstrate that our proposed approach can achieve superior performance compared with several state-of-the-art approaches.

Acknowledgments

Lei Zhu and Jialie Shen are supported by the Singapore Ministry of Education Academic Research Fund Tier 2 (MOE Ref: MOE2013-T2-2-156)

References

[Chen *et al.*, 2014] Tao Chen, Kim-Hui Yap, and Dajiang Zhang. Discriminative soft bag-of-visual phrase for mobile landmark recognition. *IEEE Trans. Multimedia*, 16(3):612–622, 2014.

[Cheng and Shen, 2016] Zhiyong Cheng and Jialie Shen. On very large scale test collection for landmark image search benchmarking. *Signal Process.*, 124:13–26, 2016.

[Duan *et al.*, 2013] Ling-Yu Duan, Jie Chen, Rongrong Ji, Tiejun Huang, and Wen Gao. Learning compact visual descriptors for

Table 4: Effects of submodular function based canonical view discovery.

Methods	*Oxford5K*				*Paris6K*				*Paris500K*			
	32	48	64	128	32	48	64	128	32	48	64	128
Random	0.3109	0.3057	0.3177	0.3206	0.3476	0.3600	0.3709	0.3925	0.4293	0.4836	0.5176	0.5697
K-means	0.3170	0.3108	0.3251	0.3357	0.3592	0.3612	0.3716	0.4010	0.4329	0.5005	0.5358	0.5860
2CVR	**0.3176**	**0.3371**	**0.3458**	**0.3586**	**0.3644**	**0.3856**	**0.4022**	**0.4173**	**0.4480**	**0.5191**	**0.5645**	**0.6139**

low bit rate mobile landmark search. *AI Magazine*, 34(2):67–85, 2013.

[Elhamifar and Vidal, 2013] Ehsan Elhamifar and René Vidal. Sparse subspace clustering: Algorithm, theory, and applications. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(11):2765–2781, 2013.

[Gong *et al.*, 2013] Yunchao Gong, S. Lazebnik, A. Gordo, and F. Perronnin. Iterative quantization: A procrustean approach to learning binary codes for large-scale image retrieval. *IEEE Trans. Pattern Anal. Mach. Intell.*, 35(12):2916–2929, 2013.

[Ji *et al.*, 2012] Rongrong Ji, Ling-Yu Duan, Jie Chen, Hongxun Yao, Junsong Yuan, Yong Rui, and Wen Gao. Location discriminative vocabulary coding for mobile landmark search. *Int. J. Comput. Vision*, 96(3):290–314, February 2012.

[Jiang and Li, 2015] Qing-Yuan Jiang and Wu-Jun Li. Scalable graph hashing with feature transformation. In *IJCAI*, pages 2248–2254, 2015.

[Kim and Choi, 2013] Saehoon Kim and Seungjin Choi. Multi-view anchor graph hashing. In *ICASSP*, pages 3123–3127, 2013.

[Liu *et al.*, 2011] Wei Liu, Wang Jun, Sanjiv Kumar, and Shih-Fu Chang. Hashing with graphs. In *ICML*, pages 1–8, 2011.

[Liu *et al.*, 2014] Xianglong Liu, Junfeng He, and Bo Lang. Multiple feature kernel hashing for large-scale visual search. *Pattern Recogn.*, 47(2):748–757, 2014.

[Lowe, 2004] David G. Lowe. Distinctive image features from scale-invariant keypoints. *Int. J. Comput. Vision*, 60(2):91–110, 2004.

[Nemhauser *et al.*, 1978] G.L. Nemhauser, L.A. Wolsey, and M.L. Fisher. An analysis of approximations for maximizing submodular set functionsi. *Math. Program.*, 14(1):265–294, 1978.

[Oliva and Torralba, 2001] Aude Oliva and Antonio Torralba. Modeling the shape of the scene: A holistic representation of the spatial envelope. *Int. J. Comput. Vision*, 42(3):145–175, 2001.

[Park *et al.*, 2000] Dong Kwon Park, Yoon Seok Jeon, and Chee Sun Won. Efficient use of local edge histogram descriptor. In *MM*, pages 51–54, 2000.

[Philbin *et al.*, 2007] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, pages 1–8, 2007.

[Philbin *et al.*, 2008] James Philbin, Ondrej Chum, Michael Isard, Josef Sivic, and Andrew Zisserman. Lost in quantization: Improving particular object retrieval in large scale image databases. In *CVPR*, pages 1–8, 2008.

[Shen *et al.*, 2015] Xiaobo Shen, Fumin Shen, Quan-Sen Sun, and Yun-Hao Yuan. Multi-view latent hashing for efficient multimedia search. In *MM*, pages 831–834, 2015.

[Sivic and Zisserman, 2003] Josef Sivic and Andrew Zisserman. Video google: A text retrieval approach to object matching in videos. In *ICCV*, pages 1470–1477, 2003.

[Song *et al.*, 2013] Jingkuan Song, Yi Yang, Zi Huang, Heng Tao Shen, and Jiebo Luo. Effective multiple feature hashing for large-scale near-duplicate video retrieval. *IEEE Trans. Multimedia*, 15(8):1997–2008, 2013.

[Wang *et al.*, 2009] Xiaoyu Wang, Tony X. Han, and Shuicheng Yan. An hog-lbp human detector with partial occlusion handling. In *ICCV*, pages 32–39, 2009.

[Wang *et al.*, 2010] Jun Wang, Ondrej Kumar, and Shih-Fu Chang. Semi-supervised hashing for scalable image retrieval. In *CVPR*, pages 3424–3431, 2010.

[Wang *et al.*, 2016] J. Wang, W. Liu, S. Kumar, and S. Chang. Learning to hash for indexing big data-a survey. *Proceedings of the IEEE*, 104(1):34–57, 2016.

[Weiss *et al.*, 2008] Yair Weiss, Antonio Torralba, and Robert Fergus. Spectral hashing. In *NIPS*, pages 1753–1760, 2008.

[Weyand and Leibe, 2013] Tobias Weyand and Bastian Leibe. Discovering details and scene structure with hierarchical iconoid shift. In *ICCV*, pages 3479–3486, 2013.

[Xie *et al.*, 2016a] Liang Xie, Lei Zhu, and Guoqi Chen. Unsupervised multi-graph cross-modal hashing for large-scale multimedia retrieval. *Multimedia Tools and Applications*, pages 1–20, 2016.

[Xie *et al.*, 2016b] Liang Xie, Lei Zhu, Peng Pan, and Yansheng Lu. Cross-modal self-taught hashing for large-scale image retrieval. *Signal Process.*, 124:81–92, 2016.

[Yu and Shi, 2003] Stella X. Yu and Jianbo Shi. Multiclass spectral clustering. In *ICCV*, pages 313–, 2003.

[Yu *et al.*, 2002] Hui Yu, Mingjing Li, Hong-Jiang Zhang, and Jufu Feng. Color texture moments for content-based image retrieval. In *ICIP*, pages 929–932, 2002.

[Zhang *et al.*, 2011] Dan Zhang, Fei Wang, and Luo Si. Composite hashing with multiple information sources. In *SIGIR*, pages 225–234, 2011.

[Zhou *et al.*, 2014] Wengang Zhou, Ming Yang, Houqiang Li, Xiaoyu Wang, Yuanqing Lin, and Qi Tian. Towards codebook-free: Scalable cascaded hashing for mobile image search. *IEEE Trans. Multimedia*, 16(3):601–611, 2014.

[Zhu *et al.*, 2015a] Lei Zhu, Jialie Shen, Hai Jin, Liang Xie, and Ran Zheng. Landmark classification with hierarchical multimodal exemplar feature. *IEEE Trans. Multimedia*, 17(7):981–993, 2015.

[Zhu *et al.*, 2015b] Lei Zhu, Jialie Shen, Hai Jin, Ran Zheng, and Liang Xie. Content-based visual landmark search via multimodal hypergraph learning. *IEEE Trans. Cybernetics*, 45(12):2756–2769, 2015.

[Zhu *et al.*, 2015c] Lei Zhu, Jialie Shen, and Liang Xie. Topic hypergraph hashing for mobile image retrieval. In *MM*, pages 843–846, 2015.