Binary Multi-view Image Re-ranking

Zhijian Wu, Jun Li(✉), and Jianhua Xu

School of Computer and Electronic Information, Nanjing Normal University, Nanjing 210023, Jiangsu, China
{192235022,lijuncst,xujianhua}@njnu.edu.cn

Abstract. Conventional subspace-based multi-view re-ranking methods essentially handle the Euclidean feature space transformation and tend to be inefficient when dealing with large-scale data, since the cost of computing the similarity between the query item and the database item is prohibitively high. Inspired by Hashing technique, in this paper, we propose an efficient binary multi-view image re-ranking strategy in which the original multi-view features are projected onto a compact Hamming subspace. With the intrinsic structure of the original multi-view Euclidean feature space maintained, the resulting binary codes are consistent with the original multi-view features in similarity measure. Furthermore, coupled with the discriminative learning mechanism, our method leads to compact binary codes with sufficient discriminating power for accurate image re-ranking. Experiments on public benchmarks reveal that our method achieves competitive retrieval performance comparable to the state-of-the-art and enjoys excellent scalability in large-scale scenario.

Keywords: Multi-view image re-ranking · Hamming subspace · Discriminative learning · Binary codes

1 Introduction

In visual search task, image re-ranking aims to update the query model and improve the initial retrieval accuracy by polishing the ranking list in the first place [4,15]. The core of these algorithms is to develop an accurate re-ranking model for re-evaluating the correlation between the target images and the query. Recent research suggests that subspace-based multi-view re-ranking serves as an important line of research in image re-ranking, since the traditional hand-crafted shallow features are supplementary to the deep features. Thus, the complementarity among these two heterogeneous features can be embedded within a unified multi-view learning framework for recovering a latent subspace [12,21,22].

The most representative multi-view learning framework is Canonical Correlation Analysis (CCA) [5], which is aiming at exploring the associations between two

This work was supported by the Natural Science Foundation of China (NSFC) under Grants 61703096, 61773117, 61273246 and the Natural Science Foundation of Jiangsu Province under Grant BK20170691.

Y. Peng et al. (Eds.): ICIG 2021, LNCS 12888, pp. 578–590, 2021.
https://doi.org/10.1007/978-3-030-87355-4_48

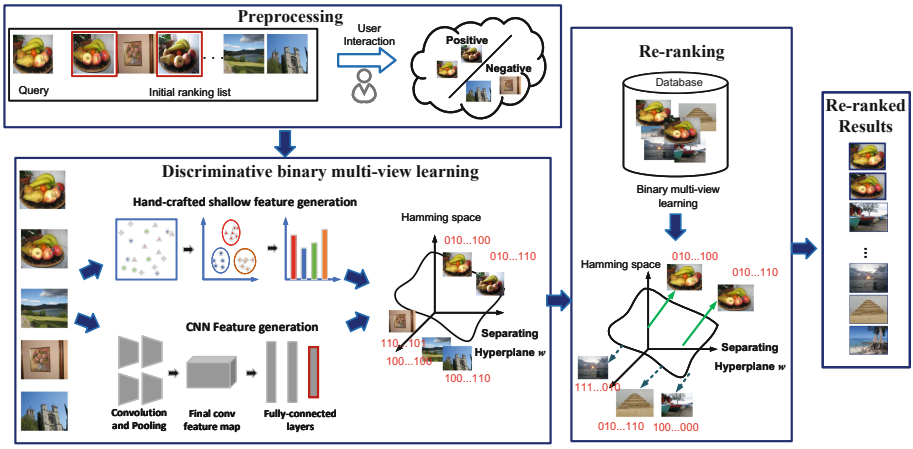


Fig. 1. The system flowchart of our BMVIR approach.

sets of variables. Besides, a wide variety of CCA variants [1,14] are developed and widely used in multi-view learning. Subsequently, multi-view re-ranking methods combining multi-view learning and image re-ranking strategy have been proposed and demonstrated their overwhelming superiority in various visual tasks [19]. In addition to CCA framework, Li et al. [10,11] develop a series of unified discriminative multi-view learning frameworks in which multi-view subspace embedding is combined with discriminative learning for accurate image re-ranking. Despite effective, these methods suffer from expensive computational cost especially in large-scale scenarios, which adversely affects real-world applications.

To address this issue, in this paper, we propose a novel binary multi-view image re-ranking method (BMVIR) for efficient and accurate image re-ranking. Inspired by image Hashing [8,20], our method attempts to recover a latent Hamming subspace from the original multi-view feature spaces, and the resulting binary codes are used for subsequent image re-ranking. More importantly, coupled with the discriminative learning mechanism, our method is capable of embedding the correlation information of the pairwise images into the generated binary codes, and thus maximally preserving the discriminating power of the compact binary codes. Benefiting from the reliable discriminant information, our method avoids the lossy coding caused by the over-dependence on the data structure for the traditional Hashing methods, and generates the similarity-preserving binary codes with sufficient discriminative power. With the help of the efficient binary codes, fast image search can be carried out via Hamming distance evaluation, which dramatically reduces the computational cost with desirable efficiency. The processing pipeline of our approach is illustrated in Fig. 1.

The rest of the paper is organized as follows. We elaborate our method in details in Sect. 2. Next, we present the experimental evaluations in Sect. 3. Finally, our work is concluded and summarized in Sect. 4.

2 Our Efficient Multi-view Re-ranking Method

2.1 Formulation

Without loss of generality, image re-ranking is defined as the problem of refining the initial retrieval results by updating the query model for improving the retrieval accuracy. Mathematically, given the initial retrieval results R obtained by query model Q, we aim to polish Q with a re-ranking model $\tilde{Q}$, and re-evaluate the query relevance of the target images using $\tilde{Q}$, leading to the re-ranked results $\tilde{R}$. More specifically, the re-ranking model $\tilde{Q}$ is built on the partially labeled samples S collected from the initial image ranks R.

When training $\tilde{Q}$, we make use of multi-view heterogeneous features of S for subspace learning. Given the original multi-view data $Z_v \in \mathbb{R}^{D_v \times n}$ $(v = 1, ..., m)$, the subspace-based multi-view embedding framework seeks to uncover the underlying subspace $X \in \mathbb{R}^{d \times n}$ such that the original multi-view features can be recovered from this subspace via view-specific generation matrix $P_v \in \mathbb{R}^{D_v \times d}$ $(v = 1, ..., m)$. m is the number of data views while n is the number of images. Besides, D_v indicates the view-specific feature dimensionality while d is the dimension of the latent subspace. Mathematically, the subspace recovery process is formulated as:

$$Z_v = P_v X + E_v \tag{1}$$

where $E_v \in \mathbb{R}^{D_v \times n}$ is denoted as the view-dependent reconstruction error. In general, the shared subspace is reconstructed by solving the following formulation:

$$\min_{P_v, X} \sum_{v=1}^{m} \|Z_v - P_v X\|_F^2 + c_1 \sum_{v=1}^{m} \|P_v\|_F^2 + c_2 \|X\|_F^2 \tag{2}$$

where c_1 and c_2 are the parameters controlling the tradeoff among the corresponding regularization terms which are used to prevent overfitting. Although Eq. (2) provides a general paradigm for recovering a potential subspace from multi-view features, the resulting Euclidean subspace is still prone to high similarity calculation costs. Inspired by learning to Hash, we attempt to recover a shared Hamming space directly from the original multi-view feature, yielding compact binary codes. Mathematically, encoding the Hamming subspace with binary discretization constraint is expressed as:

$$\begin{aligned} \min_{P_v, X, B, R} & \sum_{v=1}^{m} \|Z_v - P_v X\|_F^2 + c_1 \sum_{v=1}^{m} \|P_v\|_F^2 + c_2 \|X\|_F^2 + c_3 \left\|B - X^{\mathrm{T}} R\right\|_F^2 \\ s.t. \quad & B \in \{-1, 1\}^{n \times d}, \quad R^T R = I \end{aligned} \tag{3}$$

where B is the resulting d-dimensional Hamming subspace, and $R \in \mathbb{R}^{d \times d}$ is an orthogonal matrix, and c_1, c_2, c_3 are the tuning parameters controlling the tradeoff among respective regularization terms. Basically, since the binary matrix discretization constraint poses a great challenge to the optimization of the resulting Hamming space, it is not easy to generate an accurate mapping from Euclidean space to Hamming space while maximally maintaining the intrinsic correlation among multi-view data without significant loss. In our scheme, we impose the

orthogonal transformation on the resulting Euclidean subspace to minimize its loss with binary code, and thus avoid the problem of direct optimization of binary matrix. The principle behind it can be explained by the invariance property of the orthogonal transformation of Euclidean space. Thus, the optimal multi-view Euclidean subspace structure is retained intact in Hamming space.

In Hashing methods, it is crucial for maintaining the discrimination capability of binary codes, whereas the correlation information contained in the inherent data structure is limited,and thus somewhat affects the performance of Hashing learning. In order to further improve the discriminating power of the binary codes, we introduce discriminative learning to minimize the similarity preserving empirical loss:

$$\min_{w} \frac{1}{2} \|w\|^2 + c_4 \left\|Y - w^{\mathrm{T}} X\right\|^2 \tag{4}$$

where $w \in \mathbb{R}^{d\times 1}$ indicates the separating hyperplane in the latent subspace, whilst $Y \in \{1, -1\}^N$ is the label vector of the training samples. Accordingly, we combine Eq. (4) with Eq. (3), yielding the following binary multi-view feature learning framework:

$$\begin{aligned} \min_{P_v, X, B, R, w} & \sum_{v=1}^{m} \|Z_v - P_v X\|_F^2 + c_1 \sum_{v=1}^{m} \|P_v\|_F^2 + c_2 \|X\|_F^2 + c_3 \left\|B - X^{\mathrm{T}} R\right\|_F^2 \\ & + \frac{1}{2} \|w\|^2 + c_4 \left\|Y - w^{\mathrm{T}} X\right\|^2 \\ & s.t. \quad B \in \{-1, 1\}^{n\times d}, \quad R^T R = I \end{aligned} \tag{5}$$

As shown in Eq. (5), we incorporate Hashing learning and discriminative learning into a unified multi-view embedding framework, solving for the view-dependent generation matrices P_v, the latent representation X, the binary codes B, the decision boundary w and the orthogonal matrix R simultaneously. Thus, the re-ranking model $\tilde{Q}$ is derived. For the on-the-fly re-ranking, we first compute the binary codes $\tilde{B}$ of all the target images by optimizing the following formulation:

$$\min_{\tilde{X}, \tilde{B}} \sum_{v=1}^{m} \|\tilde{Z}_v - P_v \tilde{X}\|_F^2 + c_3 \|\tilde{B} - \tilde{X}^T R\|_F^2 \tag{6}$$

where $\tilde{Z}_v$ denotes the original multi-view features of the target images. Thus, the query relevance can be re-evaluated and re-ordered by calculating $\tilde{B} \cdot w$, leading to the re-ranked results $\tilde{R}$.

2.2 Optimization

The formulation in Eq. (5) requires simultaneous optimization of the five parameters: the view-dependent generation matrix P_v, the low-dimensional subspace embedding X, the orthogonal rotation matrix R, the binary codes B and the decision boundary w. To solve this problem, we design an iterative algorithm to alternate the optimization of the five variables for minimizing the empirical loss as in Eq. (5):

Update P_v by Fixing Others. After removing the irrelevant terms, the formulation in Eq. (5) is reduced to:

$$\min_{P_v} \|Z_v - P_v X\|_F^2 + c_1 \|P_v\|_F^2 \tag{7}$$

Let:

$$\mathcal{L} = \min_{P_v} \|Z_v - P_v X\|_F^2 + c_1 \|P_v\|_F^2 \tag{8}$$

Thus, we take the derivative of L w.r.t. P_v and set the derivative to 0, leading to the close-form solution as follows:

$$P_v = Z_v X^{\mathrm{T}} (X X^{\mathrm{T}} + c_1 I)^{-1} \tag{9}$$

Update X by Fixing Others. With the irrelevant terms discarded, the formulation in Eq. (5) is simplified as:

$$\min_{X} \|Z_v - P_v X\|_F^2 + c_2 \|X\|_F^2 + c_3 \left\|B - X^{\mathrm{T}} R\right\|_F^2 + c_4 \left\|Y - w^{\mathrm{T}} X\right\| \tag{10}$$

Analogously, we take the derivative of L w.r.t. X and set the derivative to 0 for obtaining the following close-form solution:

$$X = (\sum_{v=1}^{m} P_v^{\mathrm{T}} P_v + c_2 I + c_2 R R^{\mathrm{T}} + c_4 w w^{\mathrm{T}})^{-1} (\sum_{v=1}^{m} P_v^{\mathrm{T}} Z_v + c_3 R B^{\mathrm{T}} + c_4 w Y) \tag{11}$$

Update B by Fixing Others. While updating B, we only maintain the terms regarding B, and thus the problem is reduced to:

$$\min_{B} \left\|B - X^{\mathrm{T}} R\right\|_F^2 \quad s.t. \quad B \in \{-1, 1\}^{n \times d} \tag{12}$$

Expanding Eq. (12), we have:

$$\begin{aligned} &\min_{B} \|B\|_F^2 + \left\|X^{\mathrm{T}}\right\|_F^2 - 2tr(B R^{\mathrm{T}} X) \\ =&\min_{B} nd + \left\|X^{\mathrm{T}}\right\|_F^2 - 2tr(B R^{\mathrm{T}} X) \end{aligned} \tag{13}$$

Since $n \cdot d$ is constant while the projected data matrix X is fixed, minimizing Eq. (13) is equivalent to maximizing

$$tr(B R^{\mathrm{T}} X) = \sum_{i=1}^{n} \sum_{j=1}^{d} B_{ij} \tilde{X}_{ij} \tag{14}$$

where $\tilde{X}_{ij}$ denotes the elements of $\tilde{X} = X^{\mathrm{T}} R$. To maximize this formulation with respect to B, we have $B_{ij} = 1$ whenever $\tilde{X}_{ij} \geq 0$ and -1 otherwise.

Update R by Fixing Others. After removing the irrelevant variables, Eq. (5) is reduced to:

$$\min_{R} \left\| B - X^{\mathrm{T}} R \right\|_F^2 \quad s.t. \quad R^T R = I \tag{15}$$

The objective function shown in Eq. (15) is in spirit the classic orthogonal Procrustes problem [6]. In our method, Eq. (15) is minimized by computing the SVD of the $n \times n$ matrix $B^{\mathrm{T}} X^{\mathrm{T}}$ as $S\Omega\hat{S}^{\mathrm{T}}$ and letting $R = \hat{S}S^{\mathrm{T}}$.

Update w by Fixing Others. Similar to the above-mentioned steps, we remove the irrelevant terms and rewrite Eq. (5) as follows:

$$\min_{w} \frac{1}{2} \|w\|^2 + c_4 \left\| Y - w^{\mathrm{T}} X \right\|^2 \tag{16}$$

we also take the derivative of L w.r.t. w and set the derivative to 0 for obtaining the following close-form solution:

$$w = (I + 2c_4 X X^{\mathrm{T}})^{-1} 2c_4 X Y^{\mathrm{T}} \tag{17}$$

We iteratively alternate the above steps until the convergence of the algorithm is reached.

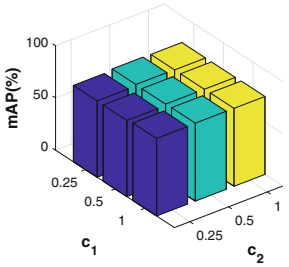
Time Complexity. In our method, computing P_v requires $O(D_v \cdot n \cdot d) + O(D_v \cdot d^2) + O(d^3) + O(d^2 \cdot n)$ time cost, which can be further approximated by $O(D_v \cdot n \cdot d) + O(D_v \cdot d^2)$ because $D_v >> d$ in our case. n denotes the number of database images. In term of the time complexity of updating X, it amounts to $m \cdot (O(D_v \cdot n \cdot d) + O(D_v \cdot d^2)) + 2O(d^3) + 2O(d^2 \cdot n)$ which can be also approximated by $m \cdot O(D_v \cdot n \cdot d) + m \cdot O(D_v \cdot d^2))$. Meanwhile, optimizing w requires approximately $O(d^2 \cdot n + d^3)$ time cost. In addition, the time complexity of computing B is $O(d \cdot n)$, while optimizing R requires $O(n^3)$ time cost. Besides, the feature matching based on Hamming distance has a time complexity of $O(d \cdot n)$.

3 Experiments

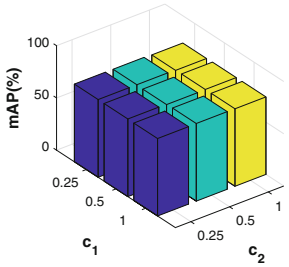
In this section, we evaluate our BMVIR framework for interactive image re-ranking. We first introduce the public benchmark datasets along with the experimental setting and the evaluation protocols. Next, comprehensive quantitative analyses are carried out to illustrate the performance of our algorithm. In addition, comparative study also demonstrates the significant advantages of our algorithm over the existing image re-ranking methods.

3.1 Datasets

We evaluate our approach on two public benchmark datasets for landmark retrieval, namely Oxford5K [16] and Paris6K [17]. Both the two datasets include



$c_3 = 0.0001, c_4 = 2$



$c_3 = 0.001, c_4 = 2$

Fig. 2. The performance of our algorithm with different parameter values $\{c_1, c_2, c_3, c_4\}$ on query "triomphe".

11 famous landmarks, each of which has five query instances, yielding a total of 55 query groups. In terms of the performance measure, we compute average precision (AP) for each query group and obtain the mean average precision (mAP) by averaging all AP scores for the overall evaluation. In addition, to assess the scalability of our algorithm in large-scale scenarios, we respectively merge Oxford5K and Paris6K with the Flickr100K [16] for large-scale evaluation. Flickr100K has a total of 100,071 images crawled from Flickr's 145 most popular tags, and is typically merged with other datasets for large-scale retrieval task.

3.2 Multi-view Features

Two heterogeneous image signatures are involved in our evaluation, namely CNN and TEDA [7]. Besides, we also make use of the VLAD+ [3] feature for the subsequent evaluation in large-scale retrieval scenario. As for the CNN feature, we directly adopt the deep model which is specifically fine-tuned for landmark retrieval and recognition task [2] and generate a 4,096-dimensional vector for feature representation. Besides, we follow [7] to compute TEDA signature which is represented by a 8,064-dimensional vector. In terms of VLAD+, we reproduce the method in [3] and also use a vocabulary of 256 visual words for producing 16,384-dimensional image feature.

3.3 Experimental Setting

As shown in Eq. (5), there are four regularization parameters c_1, c_2, c_3 and c_4 involved in the model selection. Figure 2 shows the impact of different parameters on the performance of BMVIR for query "triomphe". In implementation, we empirically set c_1 and c_2 as 0.25 and 0.5 respectively, while parameter c_3 is set to 0.001 to compromise between the multi-view reconstruction error and the binary quantization loss. In addition, c_4 is set as 2 empirically to put more weights on the similarity-preserving term for improving the discriminating power of our BMVIR model. Besides, we set the subspace dimension and the binary code length as 256 for compact representation and efficient retrieval.

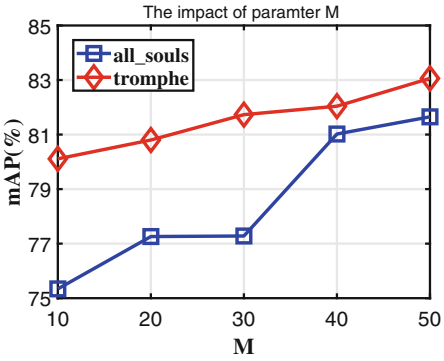


Fig. 3. The impact of the parameter M on the proposed BMVIR algorithm.

3.4 Training Samples Collection

In our method, since it is essential to indicate the query-relevance of the initial shortlisted images from scratch, we collect the query-relevant and query-irrelevant instances from the original image ranks to obtain the partial supervised information for the subspace learning. It is well known that high-ranked images are likely to be query-relevant, whereas the low-scored images are negative examples for a specific query instance in the initial retrieval results. In implementation, we use user interaction to indicate the relevance of top returned candidates in the shortlist of size M and consider the relevant images as positive training examples, whilst label L images at the bottom of the rank list as negative samples. Since our method mainly depends on the size of the shortlist M in the initial ranking list shown to the user, we discuss the influence of M in Fig. 3. It is observed that increasing M allows more positive examples involved in the training process, and thus leads to significant performance promotion. For the sake of the balance between efficiency and performance, we set M and L as 50 and 1000 respectively in practice.

3.5 Results

Baselines. We produce two separate baseline retrieval results in the first place by combining single image representation with efficient cosine similarities, which are denoted as CNN_COS, and TE_COS for short respectively. As shown in Table 1, CNN_COS consistently reports higher retrieval accuracy than TE_COS across all the performance measures, suggesting the promise of the TEDA signature and the significant heterogeneity between these two features.

Re-ranked Results by Using BMVIR. Given the initial retrieval results, we impose the proposed BMVIR on the original ranking list produced from TE_COS for accurate re-ranking. Table 2 reports the dramatic performance gains provided

Table 1. Comparison of different baselines on Oxford5K and Paris6K.

Datasets	CNN_COS	TE_COS
Oxford5K	0.6809	0.6204
Paris6K	0.7596	0.6176

Table 2. Comparison of TE_COS baseline method and TE_BMVIR re-ranking approach on some representative query groups of Paris6K.

Query	TE_COS (Baseline)	TE_BMVIR (Re-ranking)
	AP	AP
defense	0.3656	0.7233
eiffel	0.4438	0.6605
moulinrouge	0.3317	0.6532
triomphe	0.5265	0.8307
mean	0.6176	0.7731

by TE_BMVIR on the two benchmarks. It can be observed that our method significantly improves the baseline. For instance, on Paris6k, TE_BMVIR outperforms TE_COS by 35.77% on query "defense". The overall mAP score also increases from 61.76% to 77.31% accordingly. Similar advantage of TE_BMVIR is also shown on Oxford5K, indicating the mAP score is reported at 78.34% which considerably outperforms the corresponding baseline score at 62.04%. This implies that our BMVIR strategy fully explores the complementarity of heterogeneous features and combines discriminant learning to produce discriminative binary codes, and significantly improves the performance and efficiency of baselines.

Comparative Studies. To further demonstrate the advantages of our approaches, we have compared our method with other related works.

- **BMSL**: We use the recent binary multi-view fusion method BMSL [20] to produce compact binary codes from the original multi-view features. It is worth noting that this is an unsupervised multi-view learning method without discriminative learning strategy involved.
- **DMINTIR-ITQ**: In this method, we utilize the classical multi-view reranking method DMINTIR [11] to generate discriminating subspace. Next, we encode the resulting subspace representation into the binary codes by using the ITQ algorithm [6].

For the sake of consistency, we adopt the same setup of multi-view features as our method. As show in Fig. 4, our approach consistently achieves superior performance on both datasets. In particular, our method reports 78.34% mAP score on Oxford5K, and significantly exceeds the competing method BMSL reporting

Table 3. Comparison of TE_COS baseline method and TE_BMVIR re-ranking approach on some representative query groups of Oxford5K.

Query	TE_COS (Baseline)	TE_BMVIR (Re-ranking)
	AP	AP
all_souls	0.6105	0.7682
ashmolean	0.4988	0.6641
bodleian	0.3728	0.8430
hertford	0.6705	0.8755
radcliffe_camera	0.9020	0.9525
mean	0.6204	0.7834

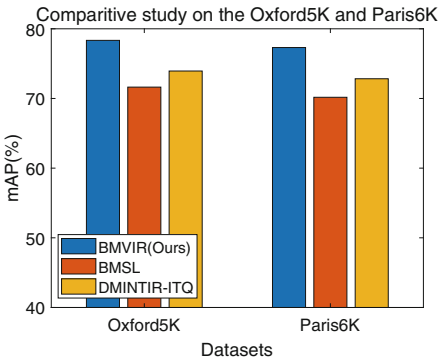


Fig. 4. Comparison of different approaches on Oxford5K and Paris6K benchmarks

score at 71.63%. This implies that the beneficial effect of the discriminative learning in our method can significantly improve the discriminant power of the generated visual representation. Meanwhile, the proposed BMVIR outperforms the DMINTIR-ITQ by 4.4% and 4.48% on respective dataset, which suggests that the unified subspace learning framework allows discriminative and compact binary representations.

In addition, we also compare our BMVIR algorithm with the state-of-the-art re-ranking methods on the two benchmarks. As presented in Table 4, the proposed approach bests all competing methods on Oxford5K including recent deep models. Surprisingly, our algorithm lags behind the state-of-the-arts on Paris6K. We argue that the disadvantage is due to the inferior performance of the original multi-view features and the insufficient complementarity between the two heterogeneous features. Since the performance of our method is achieved with very compact binary codes, this reveals the promise and the competitive performance of BMVIR along with the advantage in retrieval efficiency.

Table 4. Comparison of our method with the state-of-the-arts.

Method	Oxford5K	Paris6K
R-MAC+R+QE [18]	77.3	**86.5**
LME [13]	67.5	-
CroW+QE [9]	74.9	83.31
Ours	**78.3**	77.3

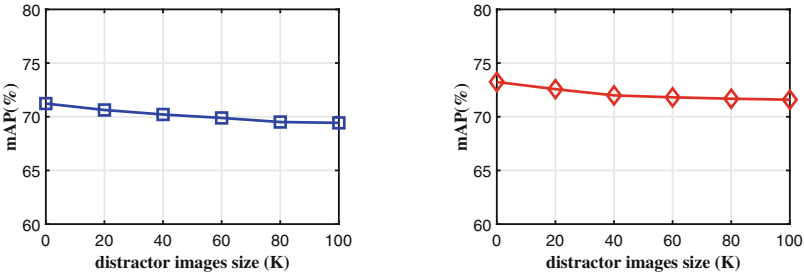


Fig. 5. Large-scale evaluation of BMVIR on Oxford5K (left) and Paris6K (right).

Scalability to Large Database. In order to evaluate the scalability of our approach, we respectively merge the Oxford5K and Paris6K with Flickr100K for large-scale evaluations. Different from the aforementioned setup of multi-view features, we use the CNN and VLAD+ as the multi-view data to explore the effect of different combinations of multiple heterogeneous features. Figure 5 illustrates the performance of our approach when the distractor images are incrementally added to Oxford5K and Paris6K. It is shown that our approach is hardly affected with the distractor images added incrementally, which substantially demonstrates the promising scalability of our approach.

Computational Cost. In implementation, training our model costs roughly 20 s, while the similarity matching takes approximately 3 ms. In contrast, the traditional retrieval scheme with cosine similarity takes about 10 ms in feature matching. This sufficiently demonstrates the real-time performance of our method using compact binary codes for instance retrieval. All the experiments are conducted under the Matlab environment using a laptop with CPU Intel Core i5-5200U 2.2 GHz and 4 GB memory.

4 Conclusions

In this paper, we have proposed a unified binary multi-view learning framework for accurate image re-ranking. In particular, we take advantage of multi-view learning paradigm to integrate the binary encoding and discriminative learning

into a unified framework, resulting in compact binary codes with sufficient discriminating power for efficient and effective image re-ranking. The evaluations on the public benchmarks and large-scale scenarios reveal that our approach achieves the promising performance with desirable scalability.

References

1. Andrew, G., Arora, R., Bilmes, J., Livescu, K.: Deep canonical correlation analysis. In: ICML, pp. 1247–1255 (2013)
2. Arandjelovic, R., Gronat, P., Torii, A., Pajdla, T., Sivic, J.: NetVLAD: CNN architecture for weakly supervised place recognition. In: CVPR, pp. 5297–5307 (2016)
3. Arandjelovic, R., Zisserman, A.: All about VLAD. In: CVPR (2013)
4. Bai, S., Tang, P., Torr, P.H., Latecki, L.J.: Re-ranking via metric fusion for object retrieval and person re-identification. In: IEEE CVPR, pp. 740–749 (2019)
5. Blaschko, M.B., Lampert, C.H.: Correlational spectral clustering. In: CVPR, pp. 1–8 (2008)
6. Gong, Y., Lazebnik, S., Gordo, A., Perronnin, F.: Iterative quantization: a procrustean approach to learning binary codes for large-scale image retrieval. IEEE TPAMI **35**(12), 2916–2929 (2012)
7. Jégou, H., Zisserman, A.: Triangulation embedding and democratic aggregation for image search. In: CVPR, pp. 3310–3317 (2014)
8. Jiang, Q.Y., Li, W.J.: Scalable graph hashing with feature transformation. IJCAI. **15**, 2248–2254 (2015)
9. Kalantidis, Y., Mellina, C., Osindero, S.: Cross-dimensional weighting for aggregated deep convolutional features. In: Hua, G., Jégou, H. (eds.) ECCV 2016, Part I. LNCS, vol. 9913, pp. 685–701. Springer, Cham (2016). https://doi.org/10.1007/978-3-319-46604-0_48
10. Li, J., Xu, C., Yang, W., Sun, C., Kotagiri, R., Tao, D.: ROMIR: robust multi-view image re-ranking. IEEE TKDE **31**(12), 2393–2406 (2018)
11. Li, J., Xu, C., Yang, W., Sun, C., Tao, D.: Discriminative multi-view interactive image re-ranking. IEEE TIP **26**(7), 3113–3127 (2017)
12. Li, J., Yang, B., Yang, W., Sun, C., Zhang, H.: When deep meets shallow: subspace-based multi-view fusion for instance-level image retrieval. In: ROBIO, pp. 486–492 (2018)
13. Li, Y., Geng, B., Tao, D., Zha, Z.J., Yang, L., Xu, C.: Difficulty guided image retrieval using linear multiple feature embedding. IEEE TOM **14**(6), 1618–1630 (2012)
14. Michaeli, T., Wang, W., Livescu, K.: Nonparametric canonical correlation analysis. In: ICML, pp. 1967–1976 (2016)
15. Ouyang, J., Zhou, W., Wang, M., Tian, Q., Li, H.: Collaborative image relevance learning for visual re-ranking. IEEE TMM 1 (2020)
16. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Object retrieval with large vocabularies and fast spatial matching. In: CVPR, pp. 1–8. IEEE (2007)
17. Philbin, J., Chum, O., Isard, M., Sivic, J., Zisserman, A.: Lost in quantization: improving particular object retrieval in large scale image databases. In: CVPR, pp. 1–8. IEEE (2008)
18. Tolias, G., Sicre, R., Jégou, H.: Particular object retrieval with integral max-pooling of CNN activations. arXiv preprint arXiv:1511.05879 (2015)

19. Wang, L., Qian, X., Zhang, Y., Shen, J., Cao, X.: Enhancing sketch-based image retrieval by CNN semantic re-ranking. IEEE Trans. Cybern. **50**(7), 3330–3342 (2020)
20. Wu, Z., Li, J., Xu, J.: Efficient binary multi-view subspace learning for instance-level image retrieval. In: Yang, H., Pasupa, K., Leung, A.C.-S., Kwok, J.T., Chan, J.H., King, I. (eds.) ICONIP 2020, Part IV. CCIS, vol. 1332, pp. 59–68. Springer, Cham (2020). https://doi.org/10.1007/978-3-030-63820-7_7
21. Xie, Y., et al.: Joint deep multi-view learning for image clustering. IEEE TKDE 1 (2020)
22. Xu, C., Tao, D., Xu, C.: Multi-view intact space learning. IEEE TPAMI **37**(12), 2531–2544 (2015)