Cognitive Computation (2021) 13:859–872
https://doi.org/10.1007/s12559-019-09687-3

Visual Cognition–Inspired Multi-View Vehicle Re-Identification via Laplacian-Regularized Correlative Sparse Ranking

Aihua Zheng[1] · Jiacheng Dong[1] · Xianmin Lin[1] · Lidan Liu[1] · Bo Jiang[1,2] · Bin Luo[1]

Received: 18 November 2018 / Accepted: 23 September 2019 / Published online: 27 November 2019

Abstract

Vehicle re-identification has gradually gained attention and widespread applications. However, most of the existing methods learn the discriminative features for identities by single-feature channel only. It is worth noting that visual cognition of the human eyes is a multi-channel system which usually seeks a sparse representation. Therefore, integrating the multi-view information in sparse representation is a natural way to boost computer vision tasks in challenging scenarios. In this paper, we propose to mine multi-view deep features via Laplacian-regularized correlative sparse ranking for vehicle re-identification. Specifically, first, we employ multiple baseline networks to generate features. Then, we explore the feature correlation via enforcing the correlation term into the multi-view Laplacian sparse ranking framework. The original rankings are obtained by the reconstruction coefficients between the probe and gallery. Finally, we utilize a re-ranking technique to further boost performance. Experimental results on public benchmark VeRi-776 and VehicleID datasets demonstrate that our approach outperforms state-of-the-art approaches. The Laplacian-regularized correlative sparse ranking as a general framework can be used in any multi-view feature fusion and will obtain more competitive results.

Keywords Vehicle re-identification · Laplacian-regularized correlative sparse ranking · Multi-view · Deep feature

Introduction

With the great progress of computer vision [1, 2], vehicle re-identification (Re-ID) has recently drawn much more attention due to its potential applications such as intelligent transportation, urban computing and intelligent monitoring. The vehicle Re-ID aims to identify the same vehicle across non-overlapping cameras, where the license plate of the vehicle is scarcely possible to be identified due to motion blur, challenging camera view, etc. In addition to person Re-ID, vehicle Re-ID has particular challenges: different identities, especially from the same manufacturer, may have similar colors and types.

The research on visual cognition of the human eyes shows that the human visual system is a multi-channel system [3] and usually seeks a sparse representation [4]. The neuroscientists at Vanderbilt University have discovered that color cognition is processed with isolation and other comprehensive visual attributes [3]. Inspired by this, many researchers used multi-view theory in computer vision tasks such as 3D shape recognition task [5], face alignment [6], visual recognition [7], pose prediction [8], and cross-view classification [9]. Meanwhile, Ravello et al. [4] have discovered that brain cognition will produce a sparser code while preserving important information, based on which many researchers developed this sparse mechanism theory in image decomposition [10], image compression [11], visual tracking [12, 13], facial expression recognition [14], etc.

The recent dramatic increase in different kinds of deep learning networks extracts discriminative feature representation for Re-ID. However, most of the existing methods focus on single-view representation learning. It is worth mentioning that visual cognition of the human eyes is a multi-channel system [3]. Generally speaking, different views from the same identity usually contain complementary information. Therefore, compared with single-view learning, multi-view learning can exploit more expressive representation. A comprehensive survey of multi-view learning refers to [5].

✉ Bo Jiang
zeyiabc@163.com

[1] Key Laboratory of Intelligent Computing and Signal Processing of Ministry of Education, School of Computer Science and Technology, Anhui University, Hefei, China

[2] Institute of Physical Science and Information Technology, Anhui University, Hefei, China

Some researchers have also discovered that visual cognition of the human eyes seeks a sparse representation for the incoming image [4]. And many algorithms based on sparse ranking in single processing have been proposed for computer vision task in the past years [15–17]. The main idea of sparse ranking is to approximately transform input data to the weighted linear combination of a small number of basis vectors from the dictionary. These basis vectors thus contain high-level patterns in the input data, while the coefficients consist of the sparse representation of the input data. Not only will this method simplify learning tasks and reduce the complexity of learning models, but it can also be used in multi-tasks learning since the multiple views are assumed to have the same sparsity pattern in their sparse representation vectors.

In this paper, we propose to explore the multi-view deep feature correlation based on the sparse ranking framework for vehicle Re-ID. Specifically, we mine the correlation between multiple feature space via correlative sparse ranking by enforcing the correlation constraint into the multi-view sparse ranking framework. Furthermore, we introduce a Laplacian regularization to preserve the local manifold structure. It can be regarded as a general framework for multi-view feature fusion for any existing networks. Furthermore, inspired by the satisfactory performance of the re-ranking techniques in person Re-ID, we further utilize the Expanded Cross Neighborhood (ECN) [18] based on the re-ranking technique to boost the performance of the proposed method.

A preliminary version of this work appeared in [19]. In this work, apart from using correlative sparse representation to fuse multi-view features, we further consider enforcing a graph Laplacian regularization into the multi-view sparse ranking framework to preserve the local manifold structure. In addition, more comprehensive experiments have been implemented, including more baseline networks, more experimental demonstration, and more evaluations on additional larger vehicle re-identification dataset: VehicleID.

Related Work

Inspired by the human visual system, we propose to mine multi-view deep features via Laplacian regularized correlative sparse ranking for vehicle re-identification in this paper. Therefore, we briefly introduce the related works on vehicle Re-ID and multi-view learning in this section.

Vehicle Re-Identification

Along with the blossom of person Re-ID [20–22], vehicle re-identification has drawn much attention recently. Several vehicle re-identification datasets have been collected to boost the research. Liu et al. [23] proposed a big dataset VeRi-776 for vehicle Re-ID, and extracted the Fusion of Attributes and Color features (FACT). Liu et al. [24] proposed a large surveillance-nature dataset (VehicleID) and explored Coupled Clusters Loss to measure the distance of arbitrary two input vehicle images. Yang et al. [25] designed CompCars dataset for vehicle model classification which can also be used for vehicle Re-ID task. Liu et al. [26] learned a structured feature embedding for vehicle Re-ID and captured nearly 1 million vehicle images to build the Vehicle-1M dataset.

Deep learning is an active method on vehicle re-identification; most of the existing methods focus on designing of effective network architectures for vehicle Re-ID. Zapletal et al. [27] learned a linear classifier on color histograms and histograms of oriented gradients by vehicle 3D bounding boxes. Zhang et al. [28] designed a classification-oriented loss and triplet sampling method based on the triplet-wise network. Kanacı et al. [29] proposed to transfer the vehicle model representation for more fine-grained Re-ID tasks via a so-called cross-level vehicle recognition method. Zhu et al. [30] proposed a shortly and densely connected convolutional neural network to combine the advantages of VGGNet and DenseNet to improve Re-ID performance.

Furthermore, some works tried to integrate other auxiliary information such as the spatio-temporal information into the vehicle Re-ID process [31, 32]. Considering that vehicles have specific attributes such as color and type, Liu et al. [33] designed a progressive searching scheme which employed the appearance attributes of the vehicle for coarse filtering. Li et al. [34] designed a unified vehicle Re-ID framework combining identification, attribute recognition, verification, and triplet tasks. Zhou et al. [35] proposed a conditional generative network to generate cross viewpoint vehicle images and combine them with input vehicle images to improve the performance of vehicle re-identification.

Multi-View Learning

Multi-view sparse ranking, which learns the shared latent representation for multi-view data, is one of the active research topics in multi-view learning and has been widely explored in the past decade. Jia et al. [36] exploit multi-view learning with structured sparsity to address human pose estimation. Liu et al. [37] proposed multi-view Hessian discriminative sparse ranking to improve the image annotation performance. Han et al. [38] designed a framework of sparse unsupervised dimensionality reduction to find a low-dimensional optimal consensus representation for image classification. Yu et al. [39] used sparse ranking to choose as few basic images as possible from the codebook and described similar Web images by fully distinct sparse codes to obtain completeness results. Wu et al. [40] proposed a sparse multi-modal hashing approach, which can jointly learn multi-modal dictionary. Lan et al. [41] considered sparse representation to fuse feature for multi-cue visual tracking.

Meanwhile, many researchers tried to use multi-view learning to solve 3D or other visual recognition tasks. Yang et al. [5] proposed a novel network structure combining multi-view networks to solve the viewpoint-based 3D shape recognition task. Chen et al. [42] designed a deep fusion scheme to combine region-wise features from multiple views for 3D object detection. Rubino et al. [43] presented a novel approach to combine dual space fitting and non-linear optimization to recover objects 3D position.

The Proposed Approach

Overview

Given a probe vehicle image, the proposed approach regarding the vehicle Re-ID consists of the following three steps as shown in Fig. 1.

(a) Multi-view deep feature learning: We design multiple deep learning–based subnetworks to extract the multi-view features.
(b) Feature fusion via Laplacian-regularized correlative sparse ranking: We propose to explore the correlation between the multi-view feature spaces via Laplacian-regularized correlative sparse ranking, which enforces the consistency between the sparse coefficients of the multi-view features and preserves the local manifold structure. The original ranking results are obtained according to the reconstruction coefficients between the probe and gallery.
(c) ECN-based re-ranking: The final ranking results are achieved via ECN based on the re-ranking technique.

We shall elaborate the procedure in the following three subsections.

Multi-View Deep Feature Learning

Inspired by the human visual system, deep learning builds hierarchical layers of visual representation to extract the high-level features of an image. We exploit K arbitrary deep learning network to generate the multi-view features for vehicle Re-ID as shown in Fig. 1a.

Furthermore, we extract K feature vectors X^1, X^2, and X^K. For the gallery with N images, K feature matrices $\boldsymbol{U^1} = [u_1^1, \cdots, u_N^1]$, $\boldsymbol{U^2} = [u_1^2, \cdots, u_N^2]$, and $\boldsymbol{U^K} = [u_1^K, \cdots, u_N^K]$ are generated in the same manner, where u_i^1, u_i^2, and u_i^K represent the feature vector of the ith gallery image h_i from the kth subnetworks respectively.

In this paper, we design three subnetworks to generate multi-view deep features: ResNet-50-based attribute aggregated subnetwork (R_{attr}), GoogleNet-based attribute

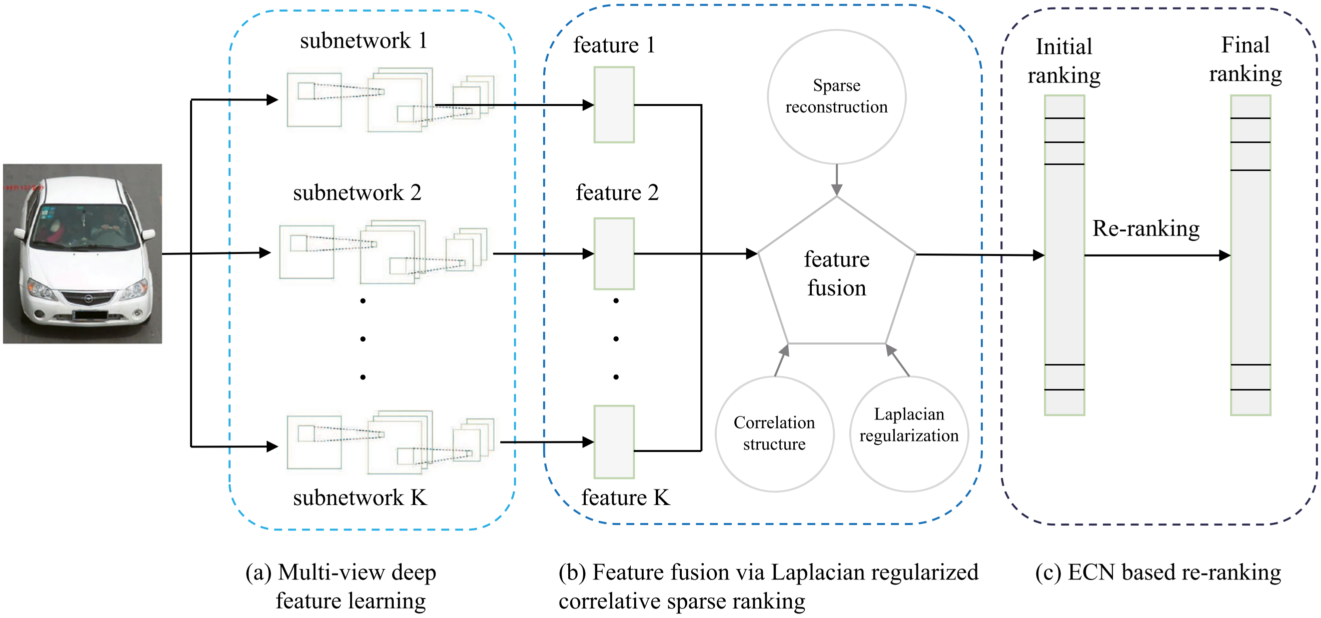


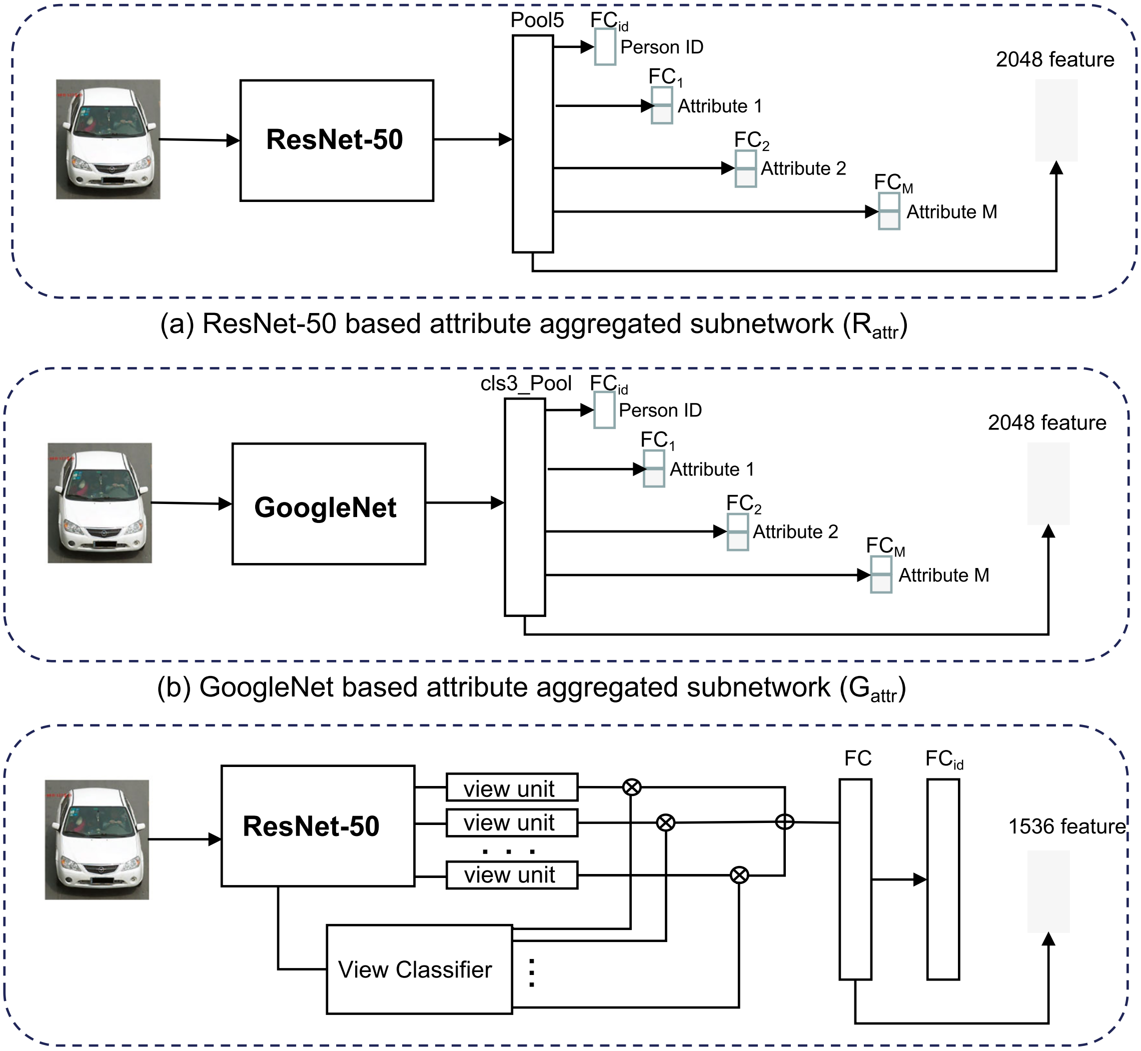
Fig. 1 Overall architecture of the proposed method

aggregated subnetwork (G_{attr}), and ResNet-50-based viewpoint embedded subnetwork (R_{view}). We shall elaborate the architectures of the three subnetworks as follows.

ResNet-50-Based Attribute Aggregated Subnetwork (R_{attr})

Inspired by the attribute aggregated network in person Re-ID [44], we encode ten color attributes (yellow, orange, green, gray, red, blue, white, golden, brown, black) and nine type attributes (sedan, suv, van, hatchback, mpv, pickup, bus, truck, estate) into the ResNet-50 deep framework, as shown in Fig. 2a. Color and type are the most recognizable appearance information. For the sake of attribute recognition, here we attach $M + 1$ fully connected (FC) layers in the end, including a ID classification and M attributes, where M is the sum of the number of attributes (colors and types). Specifically, for the FC layer for ID classification, the number of output nodes equals the number of training vehicle identities C; while each of the FC layers for one attribute (color or type) links B output nodes corresponding to the B discriminant results.

For loss computation, we use the softmax classification loss function to optimize vehicle identity discrimination in vehicle ID classification branch, and the total vehicle ID loss is calculated by cross entropy loss function as: $L_{ID} = -\sum_{c=1}^{C} f(c)log(p(c))$, where C is the vehicle identity and $f(c)$ represents the vehicle ID ground truth. The attribute probability can be predicted in the same manner with cross entropy loss function $L_{Att_j} = -\sum_{j_b=1}^{B} f(j_b)log(p(j_b))$, where $f(j_b)$ denotes the ground truth of vehicle attribute, and $B = 2$ indicates the binary discriminant results (“yes” or “no”) for attribute j. Finally,



(a) ResNet-50 based attribute aggregated subnetwork (R_{attr})

(b) GoogleNet based attribute aggregated subnetwork (G_{attr})

(c) ResNet-50 based viewpoint embedding subnetwork (R_{view})

Fig. 2 The architecture of the subnetworks

we extract 2048 dimensional features from the pool layers of the subnetwork.

GoogleNet-Based Attribute Aggregated Subnetwork (G_{attr})

GoogleNet-based attribute aggregated subnetwork (G_{attr}) is constructed in the same manner as R_{attr} shown in Fig. 2b. The only difference is that we replace the ResNet-50 in R_{attr} with GoogleNet.

ResNet-50-Based Viewpoint Embedding Subnetwork (R_{view})

As shown in Fig. 2c, we encode five viewpoint information (front, left front side, left side, left side, left rear side, rear) of the vehicle into the ResNet-50 deep framework [18]. The backbone of this subnetwork is the first three blocks of ResNet-50, we split the view classifier branch with three convolutional layers of 5 × 5, 3 × 3, and 5 × 5 respectively and one fully connected layer from the first block of ResNet-50. The classifier is used to predict a probability distribution over the corresponding view values. And then we copy the fourth block of ResNet-50 five times as the equivalent view units to extract high-level features. We shall weigh the five view softmax prediction scores of the view classifiers to the corresponding units with the element-wise multiplication. Then, we use the manner of element-wise sum to fuse these high-level features. Finally, we exploit two fully connected layers to embed the fused feature. And the softmax classification loss function is used to classify the vehicle ID by the cross entropy loss function $L_{ID} = -\sum_{c=1}^{C} f(c) log(p(c))$ for model training, where C is the number of the vehicle images and $f(c)$ represents the ground truth of vehicle ID.

It is worth mentioning that the VehicleID dataset contains incomplete attribute information and no view labels, so we only select R_{attr} and G_{attr} without attribute aggregation during implementation. As for R_{view}, we train the view predictor on the VeRi-776 dataset with view annotation then directly apply to the VehicleID dataset.

Feature Fusion via Laplacian-Regularized Correlative Sparse Ranking

After obtaining the multi-view deep features of the vehicle, we propose a Laplacian-regularized correlative sparse representation to bridge the multi-view features generated from three subnetworks as shown in Fig. 1b.

Model Formulation

In this subsection, based on their close latent correlations, we shall present the detailed formulation of the multi-view feature fusion problem via a sparse ranking framework due to its robustness to noise.

Sparse Reconstruction The main idea of sparse ranking is to represent an input vector approximately as the weighted linear combination of a small number of basis vectors from the dictionary. These basis vectors thus capture high-level patterns in the input data, while the coefficients consist of the sparse representation of the input data. According to this principle, for each query sample, we calculate sparse representation α^k under the kth channel, where $X^k \approx \boldsymbol{U}^k \alpha^k$, for $k = 1, \cdots, K$, where K is the number of the views and $K = 3$ in this paper. The reason why we only evaluated on three networks (view) is that more views will introduce higher computational complexity without distinct improvement in accuracy. Therefore, three views can keep the balance between accuracy and efficiency. The process above can be converted into a ℓ_1-norm sparsity constraint regularized least squares problem:

$$\min_{\alpha^k} \|X^k - \boldsymbol{U}^k \alpha^k\|_2^2 + \lambda^k \|\alpha^k\|_1, \tag{1}$$

where λ^k controls the trade-off between the ℓ_2-norm reconstruction error and the ℓ_1-norm sparsity constraint of the coefficients under the kth view. And $\boldsymbol{U}^k$ is the feature matrix of the kth view.

Multi-View Correlation To explore the correlations of multi-view features, it is natural to punish the diversity between sparse coefficients from arbitrary two corresponding views, that is minimizing the Euclidean distance $\|\alpha^k - \alpha^l\|_2^2$ for $k, l \in 1, \ldots, K$ to find the collaborative representation from multi-views of the same vehicle. Thus, the correlative sparse ranking model can be formulated as:

$$\begin{aligned} \min_{\alpha^k} \sum_{k=1}^{K} \{\|X^k - \boldsymbol{U}^k \alpha^k\|_2^2 + \lambda^k \|\alpha^k\|_1\} \\ +\mu \sum_{k,l} \|\alpha^k - \alpha^l\|_2^2, \quad \text{s.t.} \forall \alpha^1, \alpha^2, \cdots, \alpha^K \succeq 0, \end{aligned} \tag{2}$$

where μ is the trade-off parameter to balance the sparse reconstruction error and the pairwise correlation constraints.

Laplacian Regularization Note that the local geometrical structure of the data plays an important role in data analysis [45–47], We further enforce Laplacian regularization as a smooth operator to preserve the local manifold structure. That is minimizing $\beta^k \sum_{i,j} \left(\alpha_i^k - \alpha_j^k\right)^2 W_{ij}^k$ for

$k\in 1,\ldots,K$. Thus, the Laplacian-regularized correlative sparse ranking model can be formulated as:

$$\min_{\alpha^k}\sum_{k=1}^{K}\{\underbrace{\|X^k-\boldsymbol{U}^k\alpha^k\|_2^2+\lambda^k\|\alpha^k\|_1}_{\text{sparse reconstruction}}+\underbrace{\beta^k\sum_{i,j}\left(\alpha_i^k-\alpha_j^k\right)^2W_{ij}^k\}}_{\text{Laplacian regularization}}+ \underbrace{\mu\sum_{k,l}\|\alpha^k-\alpha^l\|_2^2}_{\text{Multi-view correlation}}\ \text{s.t.}\forall\alpha^1,\alpha^2,\cdots,\alpha^K\geq 0, \tag{3}$$

where β^k is the trade-off parameter to balance the sparse reconstruction error and the Laplacian regularization constraints; W_{ij}^k is the distance of ith and jth vehicle features of the kth view.

At last, the final Laplacian-regularized correlative sparse representation vector for one query to all gallery images is expressed as:
$\alpha=\sum_{k=1}^{K}\alpha^k$.

Model Optimization

Due to the non-negativeness of α^k, Eq. 3 can be written as follows:

$$\min_{\alpha^k}\sum_{k=1}^{K}\left\{\|X^k-\boldsymbol{U}^k\alpha^k\|_2^2+\lambda^k\alpha^k\mathbf{1}+\beta^k\sum_{i,j}\left(\alpha_i^k-\alpha_j^k\right)^2W_{ij}^k\right\}$$
$$+\mu\sum_{k,l}\|\alpha^k-\alpha^l\|_2^2,\quad \text{s.t. }\forall\alpha^1,\alpha^2,\cdots,\alpha^K\geq 0, \tag{4}$$

where $\mathbf{1}$ denotes the vector with all elements as 1. To solve Eq. 4, we convert it to an unconstrained form as:

$$\min_{\alpha^k}\sum_{k=1}^{K}\left\{\|X^k-\boldsymbol{U}^k\alpha^k\|_2^2+\lambda^k\alpha^k\mathbf{1}+\beta^k\sum_{i,j}\left(\alpha_i^k-\alpha_j^k\right)^2W_{ij}^k\right\}$$
$$+\mu\sum_{k,l}\|\alpha^k-\alpha^l\|_2^2+\psi(\alpha), \tag{5}$$

where $\psi(\alpha_i^k)$ equals 1 if $\alpha_i^k\geq 0$, and 0 otherwise. α_i^k denotes the representation coefficient of gallery image h_i to the query sample from the kth subnetwork. In this paper, we utilize the accelerated proximal gradient (APG) approach [48] to optimize efficiently. We denote:

$$\begin{aligned}F&=\min_{\alpha^k}\|X^k-\boldsymbol{U}^k\alpha^k\|_2^2+\lambda^k\alpha^k\mathbf{1}\\&\quad+\beta^k\sum_{i,j}\left(\alpha_i^k-\alpha_j^k\right)^2W_{ij}^k+\mu\sum_{k,l}\|\alpha^k-\alpha^l\|_2^2,\\J&=\psi(\alpha).\end{aligned} \tag{6}$$

Obviously, F is a differentiable convex function and J is a nonsmooth convex function. Therefore, according to the APG method, we obtain:

$$\alpha^k(r+1)=\min_{\alpha^k}\frac{\xi}{2}\|\alpha^k-\Omega^k(r+1)+\frac{\nabla F(\Omega^k(r+1))}{\xi}\|_2^2+J(\alpha^k), \tag{7}$$

where ξ is the Lipschitz constant, r indicates the current iteration, and $\alpha^k(r+1)$ denotes the sparse representation coefficients of the query image at the $(r+1)$-th iteration based on the kth subnetworks. $\Omega^k(r+1)=\alpha^k(r)+\frac{\rho(r-1)-1}{\rho(r)}(\alpha^k(r)-\alpha^k(r-1))$, where $\rho(r)$ is a positive sequence with $\rho(0)=\rho(1)=1$. Equation 7 can be solved by:

$$\alpha^k(r+1)=\max(0,\Omega^k(r+1)-\frac{\nabla F(\Omega^k(r+1))}{\xi}). \tag{8}$$

where $\nabla F(\Omega^k(r+1))$ can be calculated as:

$$\begin{aligned}\nabla F(\Omega^k(r+1))&=2\boldsymbol{U}^{k^T}\boldsymbol{U}^k\Omega^k(r+1)-2\boldsymbol{U}^{k^T}X^k\\&\quad+\lambda^k\mathbf{1}+2\beta^k\Omega^k(r+1)(D-W)^T\\&\quad+4\mu(K\Omega^k(r+1)-\sum_{l}^{K}\Omega^l(r+1)).\end{aligned} \tag{9}$$

Algorithm 1 summarizes the whole optimization procedure. Figure 3 shows the convergence curve of our LCSR method.

Algorithm 1 Optimization algorithm to Eq. 5.

```
Input: The feature vector $X^k$ of query image $q$, the gallery feature matrix $\boldsymbol{U}^k$, $k=1,\cdots,K$, $l=1,\cdots,K$, the parameters $\lambda,\mu,\beta$; Set $\xi=2.7\times 10^5$, $\rho(0)=\rho(1)=1$, $\varepsilon=10^{-4}$, $maxIter=200$, $r=1$.
Output: $\alpha^k$
while not converged do
    Update $\Omega^k(r+1)$ by $\Omega^k(r+1)=\alpha^k(r)+\frac{\rho(r-1)-1}{\rho(r)}(\alpha^k(r)-\alpha^k(r-1))$, where $\rho_r$ is a positive sequence;
    Update $\alpha^k(r+1)$ by Eq. 8;
    Update $\rho(r+1)=\frac{1+\sqrt{1+4\rho^2(r)}}{2}$;
    Update $r$ by $r=r+1$;
    Check the convergence condition: the maximum element change of $\alpha^k$ between two consecutive iterations is less than $\varepsilon$ or maximum number of iterations reaches $maxIter$.
end while
return $\alpha^k$
```

After obtaining the Laplacian-regularized correlative sparse representations α for each query image, we exploit $\alpha=\sum_{k=1}^{K}\alpha^k$ to aggregate them as a representation coefficients matrix $\mathbf{A}\in\mathcal{R}^{Q\times N}$, where Q represents the number of query image and N represents the number

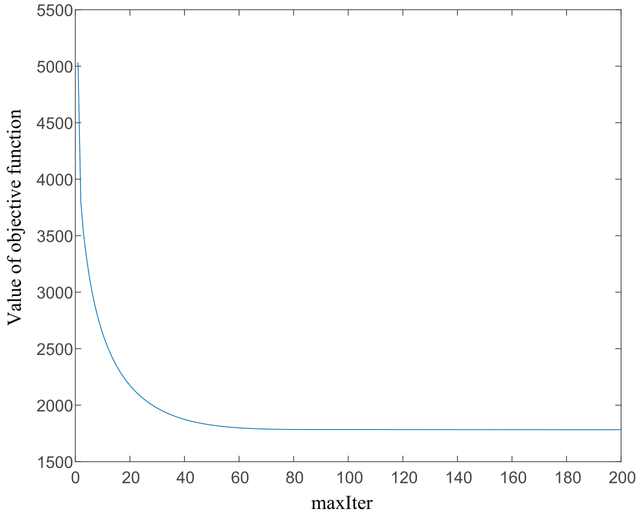


Fig. 3 Convergence curve of our LCSR method

of gallery image. The entry $\alpha_{q,i}$ in $\mathbf{A}$ denotes the representation coefficient of a gallery image h_i to the query image q. Then, the original distance between two vehicle images q and h_i can be calculated by $d(q, h_i) = 1/\alpha_{q,i}$. Therefore, the initial ranking to query sample q is $M(q, N) = \{h_1^q, h_2^q, \cdots, h_N^q\}$, where $d(q, h_i^q) < d(q, h_{i+1}^q)$.

ECN-Based Re-Ranking

The initial ranking directly compares the distance between the two images, and ignores the correlations among similar images. In order to enhance retrieval performance, here we calculate the distance by averaging the expanded neighbors of probe and gallery image pairs, that is the ECN distance, as shown in Fig. 1c.

Formally, given a probe image q and a gallery set G with N images $G = \{h_1, h_2, \cdots, h_N\}$, we can acquire the initial ranking $M(q, N) = \{h_1^q, h_2^q, \cdots, h_N^q\}$. We first define the top l samples of the query q as $M(q, l)$:

$$M(q, l) = \{h_i^q | i = 1, 2, \cdots, l\}, l \leq N. \tag{10}$$

Then, $M(h_i^q, p)$ contains the top p neighbors of each element in set $M(q, l)$:

$$M(h_i^q, p) = \{M(h_1^q, p), \cdots, M(h_l^q, p)\}. \tag{11}$$

The final expanded neighbors of q are defined as the multi-set $E(q, R)$:

$$E(q, R) = \{M(q, l), M(h_i^q, p)\}, R = l + l \times p. \tag{12}$$

In the same manner, we can obtain the expanded neighbors of each gallery image as $E(h_i, R)$. Finally, the ECN [18] distance between the query image q and any gallery image h_i is calculated as:

$$ECN(q, h_i) = \frac{1}{2R} \sum_{j=1}^{R} \{d(E(q, R)\{j\}, h_i) + d(E(h_i, R)\{j\}, q)\}, \tag{13}$$

where $E(q, R)\{j\}$ and $E(h_i, R)\{j\}$ indicate the jth expanded neighbor of query q and the ith image h_i respectively. We exploit the $ECN(q, h_i)$ to acquire the final ranking. In practice, $l = 4$, $p = 12$ as discussed in the section "Parameter Analysis."

Experiments

We evaluate our method on the two recent public benchmark datasets VeRi-776 [23] and VehicleID [24] for vehicle re-identification, and compared these with twelve state-of-the-art methods. Besides, we have annotated the view information for both VeRi-776 and VehicleID datasets.

Experiment Setting

Parameters

During the deep feature extraction, we resize all training images into 256×256 pixels and extract randomly 224×224 patches to data augmentation. We train our models using stochastic gradient descent (SGD) with a batch size of 16, momentum of 0.9, and weight decay of $\delta = 0.0001$. The learning rate is set to 0.1 at the beginning and changed to 0.01 in the last few epochs. $\lambda = 0.118$, $\beta = 0.048$, and $\mu = 0.5$ in Eq. 3.

Evaluation Metric

Following the evaluation protocol of re-identification work [31, 33], the mean average precision (mAP), and Rank1 and Rank5 accuracies are utilized to evaluate the performance of re-identification in the camera network.

Compared State-of-the-Art Methods

The specifications of the compared methods are described as follows:

1. **.LOMO** [49]. Local Maximal Occurrence Representation (LOMO) is a local feature descriptor coping with illumination variations and viewpoint changes.
2. **BOW-CN** [50]. Bag-of-Word-based hand-crafted features for vehicle Re-ID.
3. **GoogLeNet** [51]. Pre-trained on ImageNet [52] and then fine-tuned on the CompCars dataset for semantic feature representation of vehicles.
4. **FACT** [23]. Fused Appearance features including color, texture and shape.
5. **FACT+Plate-SNN+STR** [33]. **FACT** [23] with additional plate verification and spatio-temporal relations (STR) based on Siamese Neural Network (SNN).
6. **NuFACT** [53]. The null space-based **FACT** [23] to integrate the multi-level appearance features of vehicles and high-level attribute features.
7. **Siamese-Visual** [31]. Siamese-CNN with only visual information.
8. **Siamese+Path-LSTM** [31]. Siamese-CNN together with Path LSTM with visual-spatio-temporal path information.
9. **VAMI** [35]. Generating multi-viewpoint features of a vehicle from single-viewpoint feature.
10. **C2F-Rank** [26]. Exploring structured feature embedding and a novel coarse-to-fine ranking loss to boost the performance of vehicle re-identification.
11. **CLVR** [29]. The structured information of vehicle identity and vehicle model class to construct Cross-Level Vehicle Recognition method.
12. **SDC-CNN** [30]. The short and dense connection mechanism which can improve the ability of feature embedding.

Evaluation on the VeRi-776 Dataset

VeRi-776 [23] dataset contains 51,035 images of 776 vehicles captured by 20 cameras in real-world traffic surveillance environment. Due to different viewpoints, illuminations, resolutions, and occlusions, VeRi-776 is a challenging dataset for vehicle Re-ID. Specifically, there are 37,778 images of 576 vehicles for training, 11,579 images of 200 vehicles for testing, and 1678 for query. Each vehicle is captured by 2–18 cameras along a circular road. Furthermore, each vehicle image is annotated with corresponding attributes, e.g., type and color.

Quantitative Result

We evaluate the performance of the proposed method when compared with the state-of-the-art methods on VeRi-776 dataset and report the quantitative results in Table 1. Although the mAP of Siamese+Path-LSTM [31] is comparative with our LCSR, it is worth noting that it has utilized additional spatio-temporal path information, even though the Rank1 and Rank5 accuracies of our LCSR are significantly higher (without any path information). Furthermore, the ECN re-ranking technique can further boost the performance, especially on mAP and Rank1. Note that Rank5 slightly declines. The reason might be, on VeRi-776 dataset, the high accuracy in Rank1 (91.29%) implies that most of the top-1 matching is the right hit which has less probability to remove the right hit out of top-5 matchings after re-ranking, which will not change the accuracy of Rank5. However, when the top-1 matching is the wrong hit,

Table 1 The mAP, Rank1, and Rank5 comparisons on VeRi-776 dataset (in %)

Method	mAP	Rank1	Rank5	Reference
(1) LOMO [49]	9.64	25.33	46.48	CVPR2015
(2) BOW-CN [50]	12.20	33.91	53.69	ICCV2015
(3) GoogLeNet [51]	17.89	52.32	72.17	CVPR2015
(4) FACT [23]	18.49	50.95	73.48	ICME2016
(5) FACT+Plate-SNN+STR [33]	27.70	61.44	78.78	ECCV2016
(7) Siamese-Visual [31]	29.48	41.12	60.31	ICCV2017
(8) Siamese+Path-LSTM [31]	58.27	83.49	90.04	ICCV2017
(6) NuFACT [53]	48.47	76.76	91.42	TMM2018
(9) VAMI [35]	50.13	77.03	90.82	CVPR2018
(12) SDC-CNN [30]	53.45	83.49	92.55	ICPR2018
LCSR	59.58	91.29	95.47	Ours
LCSR + ECN	63.66	92.61	94.00	Ours

The top three results are highlighted in red, green, and blue, respectively the right hit, which has high probability (95.47%) within top-5 matching, may be removed from the top-5 matchings which will decline the accuracy of Rank5.

Qualitative Result

An example of the qualitative results of our method is demonstrated in Fig. 4, where Fig. 4a, b, and c demonstrate the ranking results with GoogleNet-based attribute aggregated subnetwork G_{attr}, ResNet-50-based attribute aggregated subnetwork R_{attr}, and ResNet-50-based viewpoint embedding subnetwork R_{view}, respectively. Figure 4d is the ranking results of our Laplacian-regularized correlative sparse ranking framework from which we can observe that, by correlatively learning the multi-view deep features, our method can improve the ranking to the single-view method.

Evaluation on the VehicleID Dataset

VehicleID [24] dataset contains 221,763 images of a total of 26,267 vehicles captured in real-world traffic surveillance environment. Due to illuminations, resolutions, and same appearance of different vehicle identities, VehicleID is also a challenging dataset for vehicle Re-ID. It is divided into the training set with 110,178 images of 13,134 vehicles, and the testing set with 111,585 images of 13,133 vehicles specifically. Following the protocols, we use three different testing sets for re-identification task, which contains 800, 1600, and 2400 vehicles.

It is worth mentioning that the VehicleID dataset contains incomplete attribute information and no view labels, besides each image only has the front or the back viewpoint; therefore we only select R_{attr} and G_{attr} without attribute aggregation during implementation. As for R_{view}, we train the view predictor on VeRi-776 dataset with view annotation then directly apply to the VehicleID dataset.

Quantitative Result

The quantitative results of the proposed method on VehicleID dataset when compared with the state-of-the-art methods are reported in Table 2. The results consistently demonstrate the promising performance of the proposed method. The Rank5 accuracies of our method are worse than those of SDC-CNN [30], but with significantly higher mAP and Rank1, which are more competitive metrics with the meaning of more true hits in the front ranks. Again,

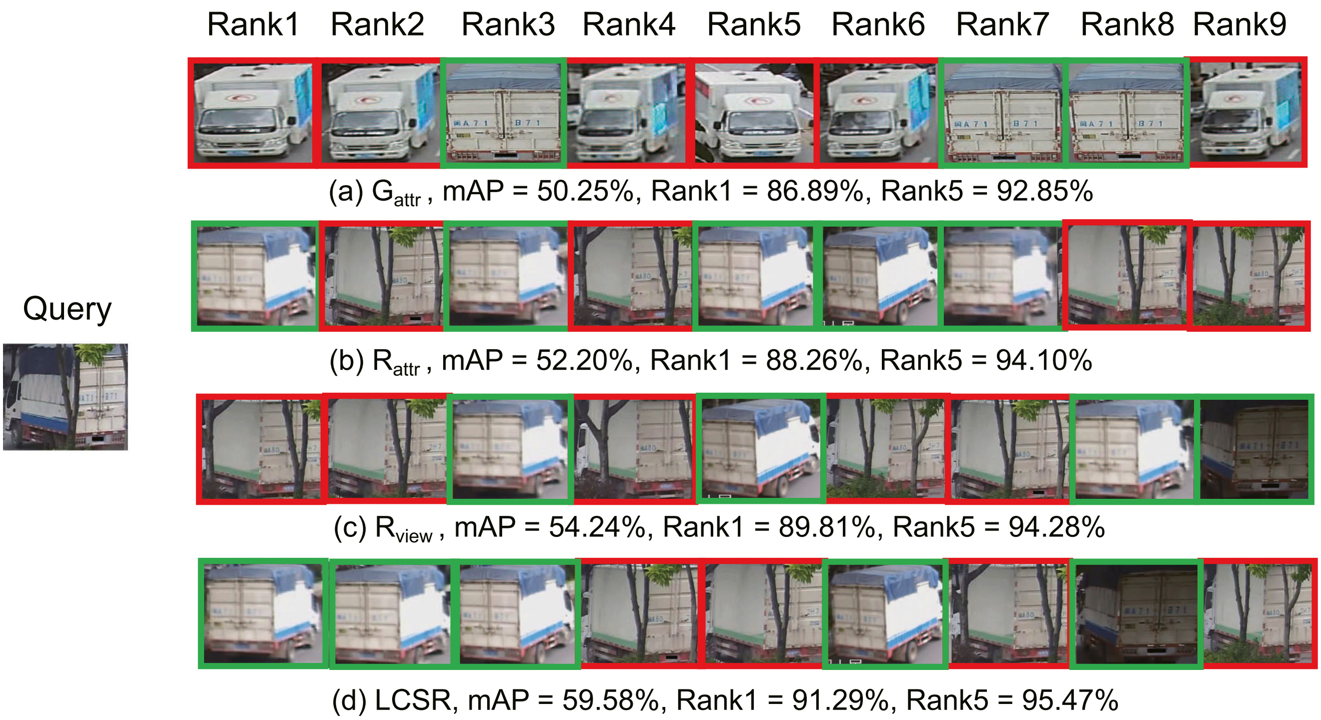

Rank1
Rank2
Rank3
Rank4
Rank5
Rank6
Rank7
Rank8
Rank9
Query
(a) G_{attr} , mAP = 50.25%, Rank1 = 86.89%, Rank5 = 92.85%
(b) R_{attr} , mAP = 52.20%, Rank1 = 88.26%, Rank5 = 94.10%
(c) R_{view} , mAP = 54.24%, Rank1 = 89.81%, Rank5 = 94.28%
(d) LCSR, mAP = 59.58%, Rank1 = 91.29%, Rank5 = 95.47%

Fig. 4 Example of our method (LCSR) on VeRi-776 dataset. The green and red boxes indicate the right hits and the wrong hits respectively

Table 2 The mAP, Rank1, and Rank5 comparisons on VehicleID dataset (in %)

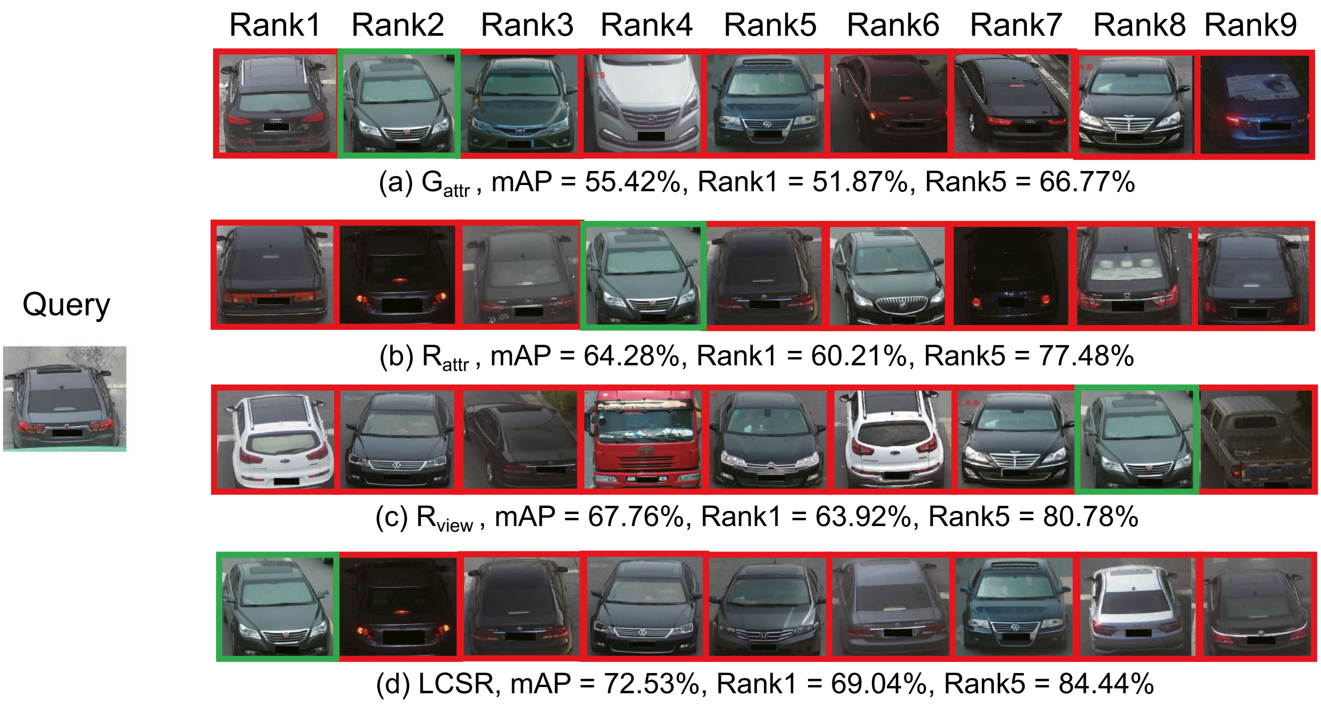
Test size	800			1600			2400			Reference
Method	mAP	Rank1	Rank5	mAP	Rank1	Rank5	mAP	Rank1	Rank5	
(1)LOMO [49]	–	19.76	32.01	–	18.85	29.18	–	15.32	25.29	CVPR2015
(2)BOW-CN [50]	–	13.14	22.69	–	12.94	21.09	–	10.20	17.89	ICCV2015
(3)GoogLeNet [51]	46.20	47.88	67.18	44.00	43.40	63.86	38.10	38.27	59.39	CVPR2015
(4)FACT [23]	–	49.53	68.07	–	44.59	64.57	–	39.92	60.32	ICME2016
(11)CLVR [29]	–	62.00	76.00	–	56.10	71.80	–	50.60	68.00	BMVC2017
(6)NuFACT [53]	–	48.90	69.51	–	43.64	65.34	–	38.63	60.72	TMM2018
(9)VAMI [35]	–	63.12	83.25	–	52.87	75.12	–	47.34	70.29	CVPR2018
(10)C2F-Rank [26]	63.50	61.10	81.70	60.00	56.20	76.20	53.00	51.40	72.20	AAAI2018
(12)SDC-CNN [30]	63.52	56.98	86.90	57.07	50.57	80.05	49.68	42.92	73.44	ICPR2018
LCSR	72.53	69.04	84.44	69.68	66.40	80.41	65.65	62.31	75.89	Ours
LCSR + ECN	68.70	64.85	82.61	55.94	53.35	64.88	52.05	49.35	61.46	Ours

The top three results are highlighted in red, green, and blue, respectively

by correlatively learning the multi-view deep features, our method can significantly improve the ranking to the single-view method. Since there is only one ground truth in the gallery for each query, and the overall accuracy is relatively low on the VehicleID dataset, the ECN may introduce more wrong hits to the top matchings, so the results after ECN tend to decline.

Qualitative Result

An example of the qualitative results on VehicleID dataset is demonstrated in Fig. 5 in the same manner as on the VeRi-776 dataset from which we can observe that the multi-view fusion result of our LCSR can significantly shift forward the true hit. Note that there is only one



(a) G_{attr} , mAP = 55.42%, Rank1 = 51.87%, Rank5 = 66.77%

(b) R_{attr} , mAP = 64.28%, Rank1 = 60.21%, Rank5 = 77.48%

(c) R_{view} , mAP = 67.76%, Rank1 = 63.92%, Rank5 = 80.78%

(d) LCSR, mAP = 72.53%, Rank1 = 69.04%, Rank5 = 84.44%

Fig. 5 Example of our method (LCSR) on VehicleID dataset (test size = 800). The green and red boxes indicate the right hits and the wrong hits respectively

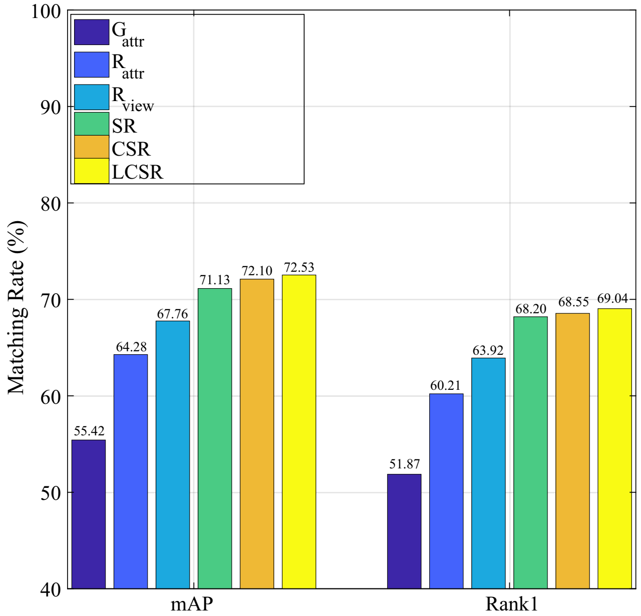


Fig. 6 Component analysis of our method (LCSR) on VehicleID dataset (test size = 800). G_{attr}, R_{attr}, and R_{view} represent the results of the GoogleNet subnetwork, ResNet-50 subnetwork, and ResNet-50-based viewpoint embedding subnetwork, respectively. SR, CSR, and LCSR indicate the results of fusing the three subnetworks G_{attr}, R_{attr}, and R_{view} via sparse ranking (SR), correlative sparse ranking [19], and our Laplacian-regularized correlative sparse ranking (LCSR) respectively

ground truth vehicle image in gallery set in the VehicleID dataset.

Component Analysis

In order to validate the component contribution of our Laplacian-regularized correlative sparse ranking (LCSR) framework, we further evaluate the components of the proposed method with its variants on VehicleID dataset (test size = 800) as shown in Fig. 6, from which we can see the following: (1) By fusing the three multi-view subnetworks G_{attr}, R_{attr}, and R_{view}, sparse ranking (SR) (comparing SR with G_{attr}, R_{attr}, or R_{view}), it can improve the performance of the single subnetwork. (2) By introducing the multi-view correlation to the original sparse ranking model (comparing CSR to SR), it can further boost the performance. (3) By introducing the Laplacian regularization constraint to CSR, our final model LCSR achieves the best performance.

Evaluation on Other Subnetworks

In order to validate the generality of our LCSR, we further evaluate it on the other three subnetworks, the conventional MobileNet [54], GoogleNet [55], and ResNet-50 [56] without any attribute aggregation on VeRi-776 dataset and VehicleID dataset (test size = 800). Table 3 reports the results of our LCSR, from which we can see that (1) all these three subnetworks achieve satisfactory performance. (2) Our LCSR framework further boosts the performance by fusing the multi-view information from these three networks, which implies that one can use our method to fuse any subnetworks in potential applications.

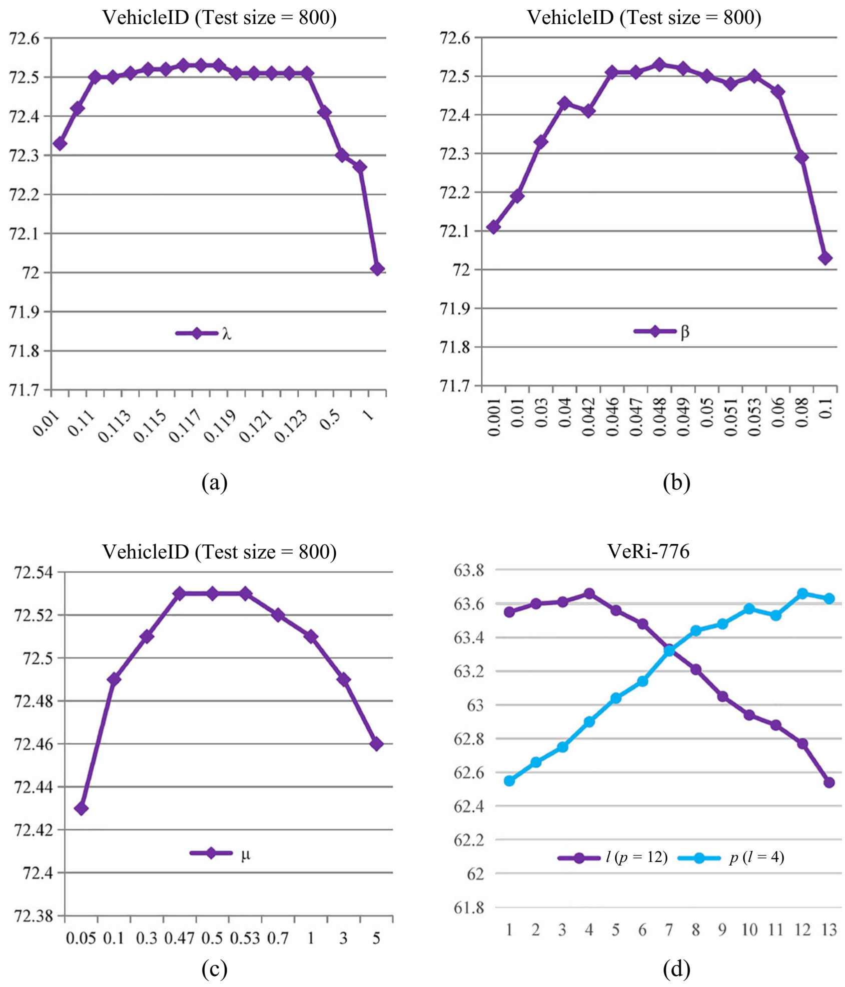
Parameter Analysis

There are three key parameters in the process of Laplacian-regularized correlative sparse ranking feature fusion. λ balances the ℓ_2-norm reconstruction error and the ℓ_1-norm sparsity constraint of the coefficients under the kth view. β balances the sparse reconstruction error and the Laplacian regularization constraints. μ balances the Laplacian sparse reconstruction error and the pairwise correlation constraints. In this paper, we empirically set $\lambda = 0.118$, $\beta = 0.048$, and $\mu = 0.5$ to obtain the best performance. We conduct an experiment on the VehicleID dataset (test size = 800) with different parameters and report the performance in Fig. 7a, b, and c. From Fig. 7, we can see that the accuracy does not change significantly by varying μ and λ in a hundred magnifications while β in a thousand magnifications, which demonstrates our model is not sensitive to these three parameters.

Table 3 Evaluation of mAP and Rank1 on the proposed LCSR on other networks (in %)

	VeRi-776		VehicleID (800)	
Method	mAP	Rank1	mAP	Rank1
MobileNet [54]	48.37	87.30	31.12	23.51
GoogleNet [55]	49.39	83.90	55.42	51.87
ResNet-50 [56]	47.39	86.05	64.28	60.21
LCSR	58.01	88.97	68.82	65.06

Fig. 7 Parameter analysis of our method (in %)



(a) (b) (c) (d)

In addition, there are two key parameters in ECN re-ranking, where l represents the number of the top samples to the query q, and p indicates the nearest neighbors to each top samples. We evaluate these two parameters on VeRi-776 dataset and report the performance in Fig. 7d, which consistently demonstrates that ECN is not sensitive to these two parameters.

Conclusion

In this paper, inspired by multi-channel and sparse representation visual cognition of the human eyes, we discover the correlation between multi-view deep sparse features for vehicle Re-ID. In deep feature extraction, three CNN re-identification networks are trained to generate the multi-view deep features. Then, we propose the Laplacian-regularized correlative sparse ranking method to jointly learn the sparse coefficients for multi-view features. Furthermore, we re-rank the initial ranking via ECN distance to boost the recognition accuracy. Experimental results on benchmark datasets VeRi-776 and VehicleID demonstrate the promising performance of our method. In the future, we shall further integrate the path and plate information for vehicle Re-ID.

Funding Information This research is supported in part by the National Natural Science Foundation of China (61976002, 61602001, 61671018 and 61860206004), the Open Project Program of the National Laboratory of Pattern Recognition (NLPR) (201900046), and the Natural Science Foundation of Anhui Higher Education Institutions of China (KJ2019A0033); Open fund for Discipline Construction, Institute of Physical Science and Information Technology, Anhui University.

Compliance with Ethical Standards

Conflict of interest The authors declare that they have no conflict of interest.

Ethical approval This article does not contain any studies with human participants or animals performed by any of the authors.

References

1. Yan Y, Ren J, Zhao H, Sun G, Wang Z, Zheng J, Marshall S, Soraghan J. Cognitive fusion of thermal and visible imagery for effective detection and tracking of pedestrians in videos. Cogn Comput. 2017:1–11.
2. Zhao C, Li X, Ren J, Marshall S. Improved sparse representation using adaptive spatial support for effective target detection in hyperspectral imagery. Int J Remote Sens. 2013:8669–8684.
3. Gegenfurtner KR. Cortical mechanisms of colour vision. Nat Rev Neurosc. 2003:563.
4. Ravello CR, Perrinet LU, Escobar MJ, Palacios AG. Speed-selectivity in retinal ganglion cells is sharpened by broad spatial frequency, naturalistic stimuli. Scientific reports. 2019:456.
5. Yang ZX, Tang L, Zhang K, Wong PK. Multi-view cnn feature aggregation with elm auto-encoder for 3d shape recognition. Cogn Comput. 2018:1–14.
6. Xing J, Niu Z, Huang J, Hu W, Yan S. Towards robust and accurate multi-view and partially-occluded face alignment. IEEE Trans Pattern Anal Mach Intell. 2018:1–1.
7. Niu L, Li W, Xu D, Cai J. An exemplar-based multi-view domain generalization framework for visual recognition. IEEE Trans Neural Netw Learn Sys. 2018:259–272.
8. Tulsiani S, Efros AA, Malik J. Multi-view consistency as supervisory signal for learning shape and pose prediction. In: IEEE conference on computer vision and pattern recognition; 2018. p. 2897–2905.
9. You X, Xu J, Yuan W, Jing XY, Tao D, Zhang T. Multi-view common component discriminant analysis for cross-view classification. Pattern Recognit. 2019:1.
10. Zhang H, Patel VM. Convolutional sparse and low-rank coding-based image decomposition. IEEE Trans Image Process. 2018:1–1.
11. De K, Masilamani V. A no-reference image quality measure for blurred and compressed images using sparsity features. Cogn Comput. 2018:1–11.
12. Qi Y, Qin L, Zhang J, Zhang S, Huang Q, Yang MH. Structure-aware local sparse coding for visual tracking. IEEE Trans Image Process. 2018:1–1.
13. Zhang T, Xu C, Yang MH. Robust structural sparse tracking. IEEE Trans Pattern Anal Mach Intell. 2019:473–486.
14. Zeng N, Zhang H, Song B, Liu W, Li Y, Dobaie AM. Facial expression recognition via learning deep sparse autoencoders. Neurocomputing. 2018:643–649.
15. He R, Zheng WS, Hu BG, Kong XW. Two-stage nonnegative sparse representation for large-scale face recognition. IEEE Trans Neural Netw Learn Sys. 2013:35–46.
16. He R, Zheng WS, Tan T, Sun Z. Half-quadratic-based iterative minimization for robust sparse representation. IEEE Trans Pattern Anal Mach Intell. 2014:261–275.
17. Yao Y, Guo P, Xin X, Jiang Z. Image fusion by hierarchical joint sparse representation. Cogn Comput. 2014:281–292.
18. Sarfraz MS, Schumann A, Eberle A, Stiefelhagen R. A pose-sensitive embedding for person re-identification with expanded cross neighborhood re-ranking. arXiv:1711.10378. 2017.
19. Sun D, Liu L, Zheng A, Jiang B, Luo B. Visual cognition inspired vehicle re-identification via correlative sparse ranking with multi-view deep features. In: International conference on brain inspired cognitive systems; 2018. p. 54–63.
20. Chen YC, Zhu X, Zheng WS, Lai JH. Person re-identification by camera correlation aware feature augmentation. IEEE Trans Pattern Anal Mach Intell. 2018:392–408.
21. Li X, Wu A, Zheng WS. Adversarial open-world person re-identification. arXiv:1807.10482. 2018.
22. Zheng L, Yang Y, Hauptmann AG. Person re-identification: Past, present and future. arXiv:1610.02984. 2016.
23. Liu X, Liu W, Ma H, Fu H. Large-scale vehicle re-identification in urban surveillance videos. In: IEEE International Conference on Multimedia and Expo; 2016. p. 1–6.
24. Liu H, Tian Y, Yang Y, Pang L, Huang T. Deep relative distance learning: Tell the difference between similar vehicles. In: IEEE conference on computer vision and pattern recognition; 2016. p. 2167–2175.
25. Yang L, Luo P, Chen CL, Tang X. A large-scale car dataset for fine-grained categorization and verification. In: IEEE conference on computer vision and pattern recognition; 2015. p. 3973–3981.
26. Guo H, Zhao C, Liu Z, Wang J, Lu H. Learning coarse-to-fine structured feature embedding for vehicle re-identification. In: Association for the advancement of artificial intelligence; 2018. p. 1–8.
27. Zapletal D, Herout A. Vehicle re-identification for automatic video traffic surveillance. In: IEEE conference on computer vision and pattern recognition workshops; 2016. p. 25–31.
28. Zhang Y, Liu D, Zha ZJ. Improving triplet-wise training of convolutional neural network for vehicle re-identification. In: IEEE international conference on multimedia and expo; 2017. p. 1386–1391.
29. Kanacı A, Zhu X, Gong S. Vehicle reidentification by fine-grained cross-level deep learning. In: British machine vision conference; 2017. p. 1–6.
30. Zhu J, Du Y, Hu Y, Zheng L, Cai C. Vrsdnet: vehicle re-identification with a shortly and densely connected convolutional neural network. Multimedia Tools and Applications. 2018:1–15.
31. Shen Y, Xiao T, Li H, Yi S, Wang X. Learning deep neural networks for vehicle re-id with visual-spatio-temporal path proposals. In: IEEE international conference on computer vision; 2017. p. 1918–1927.
32. Wang Z, Tang L, Liu X, Yao Z, Yi S, Shao J, Yan J, Wang S, Li H, Wang X. Orientation invariant feature embedding and spatial temporal regularization for vehicle re-identification. In: IEEE conference on computer vision and pattern recognition; 2017. p. 379–387.
33. Liu X, Liu W, Mei T, Ma H. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. 2016.
34. Li Y, Li Y, Yan H, Liu J. Deep joint discriminative learning for vehicle re-identification and retrieval. In: IEEE international conference on image processing; 2017. p. 395–399.
35. Zhou Y, Shao L. Viewpoint-aware attentive multi-view inference for vehicle re-identification. In: Proceedings of the IEEE conference on computer vision and pattern recognition; 2018. p. 6489–6498.
36. Jia Y, Salzmann M, Darrell T. Factorized latent spaces with structured sparsity. Advances in Neural Information Processing Systems. 2010:982–990.
37. Liu W, Tao D, Cheng J, Tang Y. Multiview hessian discriminative sparse coding for image annotation. IEEE conference on computer vision and pattern recognition. 2014:50–60.
38. Han Y, Wu F, Tao D, Shao J, Zhuang Y, Jiang J. Sparse unsupervised dimensionality reduction for multiple view data. IEEE Trans. Circuits Syst. Video Techno. 2012:1485–1496.
39. Yu J, Rui Y, Tao D. Click prediction for web image reranking using multimodal sparse coding. IEEE Trans Image Process. 2014:2019–2032.
40. Wu F, Zhou Y, Yang Y, Tang S, Zhang Y, Zhuang Y. Sparse multi-modal hashing. IEEE Transactions on Multimedia. 2014:427–439.
41. Lan X, Ma AJ, Yuen PC. Multi-cue visual tracking using robust feature-level fusion based on joint sparse representation. In: IEEE conference on computer vision and pattern recognition; 2014. p. 1194–1201.
42. Chen X, Ma H, Wan J, Li B, Xia T. Multi-view 3d object detection network for autonomous driving. In: IEEE conference

on computer vision and pattern recognition; 2017. p. 6526–6534.
43. Rubino C, Crocco M, Bue AD. 3d object localisation from multi-view image detections. IEEE Trans Pattern Anal Mach Intell. 2017:1–1.
44. Lin Y, Zheng L, Zheng Z, Wu Y, Yang Y. Improving person re-identification by attribute and identity learning. arXiv:1703.07220. 2017.
45. Zheng M, Bu J, Chen C, Wang C, Zhang L, Qiu G, Cai D. Graph regularized sparse coding for image representation. IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society. 2011:1327.
46. Jiang B, Ding C, Tang J, Luo B. Image representation and learning with graph-laplacian tucker tensor decomposition. IEEE Trans Cybern. 2018:1–10.
47. Jiang B, Ding C, Luo B, Tang J. Graph-laplacian pca: Closed-form solution and robustness. In: IEEE conference on computer vision and pattern recognition; 2013. p. 3492–3498.
48. Parikh N, Boyd S, et al. Proximal algorithms. Foundations and Trends in Optimization 2014:127–239.
49. Liao S, Hu Y, Zhu X, Li SZ. Person re-identification by local maximal occurrence representation and metric learning. In: IEEE conference on computer vision and pattern recognition; 2015. p. 2197–2206.
50. Zheng L, Shen L, Tian L, Wang S, Wang J, Tian Q. Scalable person re-identification: a benchmark. In: IEEE international conference on computer vision; 2015. p. 1116–1124.
51. Yang L, Luo P, Change Loy C, Tang X. A large-scale car dataset for fine-grained categorization and verification. In: IEEE conference on computer vision and pattern recognition; 2015. p. 3973–3981.
52. Krizhevsky A, Sutskever I, Hinton GE. Imagenet classification with deep convolutional neural networks. In: Advances in neural information processing systems; 2012. p. 1097–1105.
53. Liu X, Liu W, Mei T, Ma H. Provid: Progressive and multimodal vehicle reidentification for large-scale urban surveillance. IEEE Transactions on Multimedia. 2018:645–658.
54. Howard AG, Zhu M, Chen B, Kalenichenko D, Wang W, Weyand T, Andreetto M, Adam H. Mobilenets: Efficient convolutional neural networks for mobile vision applications. arXiv:1704.04861. 2017.
55. Szegedy C, Ioffe S, Vanhoucke V, Alemi AA. Inception-v4, inception-resnet and the impact of residual connections on learning. In: Association for the advancement of artificial intelligence; 2017. p. 1.
56. He K, Zhang X, Ren S, Sun J. Deep residual learning for image recognition. In: IEEE conference on computer vision and pattern recognition; 2016. p. 770–778.