This ICCV paper is the Open Access version, provided by the Computer Vision Foundation.
Except for this watermark, it is identical to the version available on IEEE Xplore.

Exploiting Multi-Grain Ranking Constraints for Precisely Searching Visually-similar Vehicles

Ke Yan[1,2], Yonghong Tian[1,2]*, Yaowei Wang[3], Wei Zeng[1,2], Tiejun Huang[1,2]
[1]National Engineering Laboratory for Video Technology, School of EE&CS,
Peking University, Beijing, China
[2]Cooperative Medianet Innovation Center, China
[3]Department of Electronic Engineering, Beijing Institute of Technology, China
{keyan, yhtian, weizeng, tjhuang}@pku.edu.cn;yaoweiwang@bit.edu.cn

Abstract

Precise search of visually-similar vehicles poses a great challenge in computer vision, which needs to find exactly the same vehicle among a massive vehicles with visually similar appearances for a given query image. In this paper, we model the relationship of vehicle images as multiple grains. Following this, we propose two approaches to alleviate the precise vehicle search problem by exploiting multi-grain ranking constraints. One is Generalized Pairwise Ranking, which generalizes the conventional pairwise from considering only binary similar/dissimilar relations to multiple relations. The other is Multi-Grain based List Ranking, which introduces permutation probability to score a permutation of a multi-grain list, and further optimizes the ranking by the likelihood loss function. We implement the two approaches with multi-attribute classification in a multi-task deep learning framework. To further facilitate the research on precise vehicle search, we also contribute two high-quality and well-annotated vehicle datasets, named VD1 and VD2, which are collected from two different cities with diverse annotated attributes. As two of the largest publicly available precise vehicle search datasets, they contain 1,097,649 and 807,260 vehicle images respectively. Experimental results show that our approaches achieve the state-of-the-art performance on both datasets.

1. Introduction

With the rapid development and popularization of security systems, there is a growing need for precise vehicle search from a huge number of images captured by various surveillance cameras. Given a query vehicle image, precise vehicle search aims at finding out all instances of that vehicle. Intuitively, license plate recognition can be used to solve this problem under the assumption that license plate number is unique and recognisable for each vehicle. Unfortunately, such an assumption does not hold in some practical situations as the license plate can be easily removed, occluded or faked [1]. Precise vehicle search based on visual features can thus serve as a significant alternative approach to steadily identify a vehicle by taking vehicle's visual appearances into consideration. However, different cars may have similar visual appearances, such as the same color and the same vehicle model (product year and type). An illustration of such a phenomenon is displayed in Fig.1, where the four vehicles in the second row have nearly the same ap-

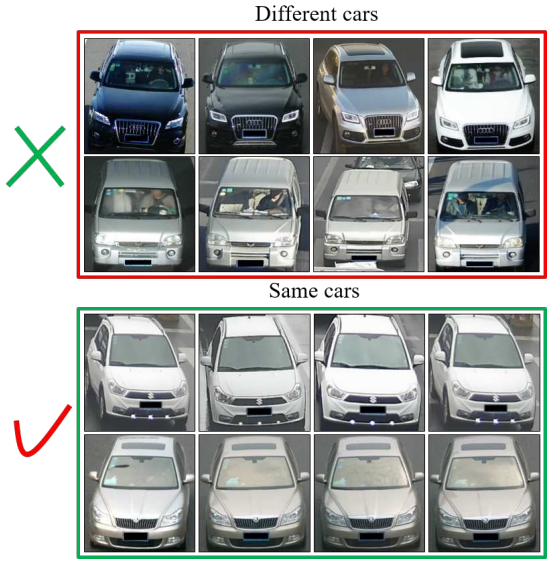


Figure 1. Some vehicle images. The first two rows exhibit vehicle images of different cars while the last two rows show some images of the same car.

*Corresponding author: Yonghong Tian

pearance except for some minor differences at windscreens. These differences can only be discovered through carefully observations even for humans. Although some existing works attempt to address this problem, their performances in practical applications are still limited. Lack of large labeled training set is an obvious reason. Collecting a large-scale vehicle search dataset containing all popular vehicle models from real scenes with meticulously annotated contents is a notoriously difficult task. As far as we know, most published vehicle datasets [2, 3, 4] are constructed for vehicle attribute recognition, with annotations on several attributes such as type, make and color. There exist two small datasets related with vehicle search, VeRi [5] and VehicleID [1]. VeRi dataset contains only 50,000 images from 776 vehicles. The scale of this dataset is small, thus only a few visually-similar vehicles are included. Relatively, VehicleID dataset [1] is more suitable for precise vehicle search, in which each vehicle is assigned with a unique ID. However, limitations also exist in VehicleID as it 1) collects only about 221,763 images from 26,267 vehicles and 2) annotates only 10,319 vehicles (90,196 images) with vehicle model information. Totally, 250 vehicle models appear in this dataset, which is far from enough since thousands of popular vehicle models could appear in a real world searching task. In this paper, we contribute two large-scale vehicle datasets (VD1 and VD2) captured from real world surveillance cameras and videos in two cities. With 1,097,649 and 807,260 images being collected and carefully annotated, the datasets contain almost all popular vehicle models and colors. To our knowledge, VD1 and VD2 are the largest high-quality annotated vehicle datasets published so far.

Apart from lacking sufficient training data, model design also limits the searching performance. Some researchers consider that person re-identification (Re-ID) related approaches [6] can be directly used to solve precise vehicle search problem [1]. Specifically, the color and model annotations in these methods are used as supervised labels for classification whilst the ID information is used to generate pairwise or triplet (including some variation) training samples only. The performance of these Re-ID methods are still not good enough because of two reasons. First, they use ID information just to form image pairs/triplets, without considering the fact that ID is also a natural and strong attribute. Second, they only consider binary relationship (i.e., 'similar' and 'dissimilar') between vehicle images. We argue that in order to perform the accurate search on vehicle data, image annotations and relationships must be fully explored and utilised. To verify this claim, we propose a multi-task deep learning framework that can be trained in attribute classification task and multi-grain based ranking task simultaneously. Different from existing models, we treat ID as an attribute of vehicle images and perform classification task on both ID and conventional vehicle attributes such as model and color. Instead of using binary relationship only, we summarize the relationship of vehicle images as multiple grains. Specifically, in the first grain two vehicle images belong to the same vehicle. The second grain is that they belong to different vehicles but having the same model and color. The rest can be defined in the same manner. The most farthest grain is that they belong to different models and colors. By introducing multi-grain relationships, we force the deep model to learn the more discriminative feature between different grains over a great deal of images. As the result, multi-grain relations are embedded in feature space so as to improve the searching performance.

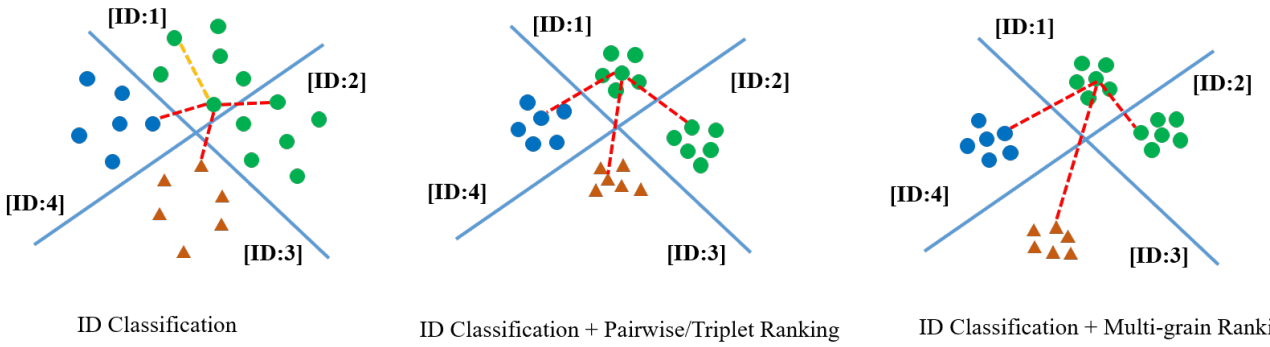
Based on the multi-grain relations, we propose two ranking approaches. The first approach is generalized pairwise ranking (GPR). Conventional pairwise ranking methods directly transform ranking problem to binary classification problem after fusing the feature of anchor images and reference images. Similarly, we generalize the formulation from binary (0/1) to multiple relations (0/1/2/3...n) by embedding the multi-grain constraints. Experimental results show that our approach outperforms the conventional pairwise approaches.

Considering the hierarchical structure of multi-grain constraints, we also propose a multi-grain based list ranking (MGLR) approach. In this model, a list of multi-grain images is maintained during training, in which the ranking is explicitly corresponding to the multi-grain relationship. We assume that any permutation for a list of images is possible after using the ranking function, but different permutations may have different likelihood values. To satisfy those requirements, we introduce the permutation probability [7, 8], which is an important model in document retrieval. It has desirable properties for representing the likelihood of a permutation. To this end, ranking of a list of images including different grains can be evaluated by the likelihood value. The likelihood loss is adopted as loss function for the multi-grain list ranking approach. Experimental results demonstrate this approach achieves the state-of-the-art performance in precise vehicle search.

The rest of this paper is organized as follows. Related works are presented at section 2. In section 3, we introduce our proposed two multi-grain based ranking approaches in detail. Section 4 gives the description of the two large-scale vehicle datasets including data collection, annotation and evaluation. Experiments are shown in section 5.

2. Related Works

Image search Traditional image search methods especially for object search focus on both feature representation and similarity learning. Generally, after representing an image with hand-crafted features (e.g., [9, 10]), a distance metric learning method (e.g., [11, 12]) is adopted to learn an optimal metric which minimizes the distance be-



ID Classification | ID Classification + Pairwise/Triplet Ranking | ID Classification + Multi-grain Ranking

Figure 2. The distribution of vehicles with different constraints in feature space. Each marker represents the feature of a vehicle image. For example, images of ID1 and ID2 come from different vehicles, but they have the same model and color; Images of ID1 and ID4 have different colors and the same model.

tween similar images and maximizes the distance between dissimilar images simultaneously. For a long time, the performance is limited as they are optimized in separate steps [13].

Recent advances in image understanding [14, 15, 16] have been driven by the success of deep convolutional neural networks (CNNs). With proper network architecture, feature representation and similarity learning can be incorporated in a unified deep framework [17]. Several similarity constraints have been proposed for feature representation in deep networks. For instance, Siamese networks [18] organize input images as similar and dissimilar pairs, after that, a two branch network with shared parameters across branches are trained to minimize a pairwise contrastive loss. Such a pairwise ranking method and its variants have achieved impressing performances for various tasks, such as face verification [19] and person Re-ID [20, 21]. Additionally, Sun et.al [22] and Yi et.al [23] propose the improved method that utilizes multi-task learning to jointly optimize classification and pairwise ranking losses. The multi-task strategy is also adopted in our methods to augment feature representation.

Apart from the pairwise related works, some researchers utilize triplet constraints to learn the similarity ranking in a CNN framework [24, 25, 26]. Triplet ranking methods can be used to address face recognition [24], cross-domain image retrieval [25] and person Re-ID [26]. They have achieved promising performances since triplet can preserve the intra-class variation well.

Vehicle search Precise vehicle search aims to search out all instances for a given object image. Its main challenge is that the inter-class variance may be as small as the intra-class variance since there are many vehicles having the same model and color. Although there are some existing works on vehicle model classification [2, 27, 28, 29] and vehicle model verification [3], only a few published works are related to vehicle search [30, 5, 1]. Zhang et.al [30] propose to model the label structures of vehicles and seamlessly embed it into a deep framework by minimizing a generalized multi-level triplet loss. Different from other triplet based methods, they manually assign different margins for each level to build triplet samples. Although this method can achieve good performance for vehicle search, it only consider the model level retrieval rather than the much harder instance level retrieval addressed in this paper. Recently Liu et.al [5] propose an approach to alleviate the vehicle re-identification problem. They focus on filtering negative images by license plate verification and spatiotemporal relations. As a result, the method works well only when these information is available. Liu et.al [1] also aims to address precise vehicle search in images from surveillance cameras. Considering that the distribution of positive samples maybe disperse, it proposes a ranking loss named coupled cluster loss to make positive samples clustered so as to improve the retrieval performance. However, it only considers binary relationship of vehicle images like most methods mentioned above. In addition, it does not take into account the importance of vehicle attributes especially the ID label.

From above related works, we find that many intrinsic and special information about vehicle is not be fully exploited. In this paper we summarize the relationship of vehicle images as multiple grains and propose two multi-grain based ranking approaches integrated with multi-attribute classification in a multi-task CNN framework.

3. Methodology

3.1. Multi-grain Relationship

Given multiple attributes, the relationship between vehicle images is abstracted to multiple grains. To model the multi-grain labels, we first introduce the concept of a multi-grain list (MGL). Specifically, each MGL $(a, R^1, R^2, \cdots, R^n)$ consists of one anchor image a and N reference images from n grain levels corresponding to the anchor image a, where $R^k = \{r_1^k, \cdots, r_C^k\}, C > 0$ rep-

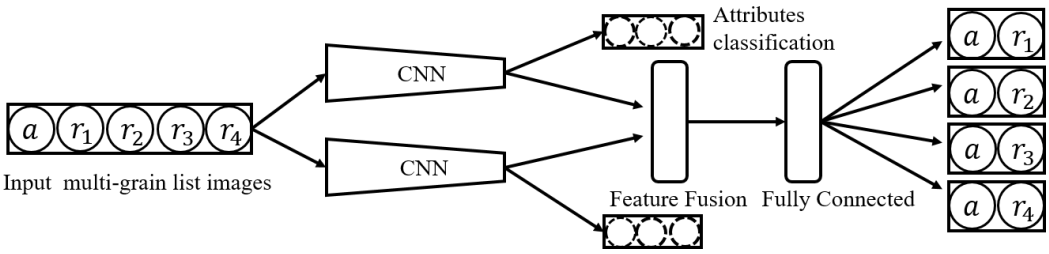


Figure 3. The framework of a multi-task learning integrating GPR with multi-attribute classification. We only draw up four grains to illustrate the main flow path of our method. All the images of a multi-grain list are extracted features in a shared CNN architecture. After fusing the anchor feature with each reference feature, the fused feature is fed into a softmax classifier to conduct grain classification.

resents the reference image set from k-th grain level. For example, r_c^1 corresponds to the c-th reference image that has the closest relationship to a (i.e., belonging to the same vehicle with a) in the list. To constrain distinct distances for different grains, we formulate multi-grain based ranking problem as:

$$\begin{aligned} &\mathcal{D}(a, R^k) + m_k < \mathcal{D}(a, R^{k+1}), \\ &m_1 > 0, m_2 > 0, \cdots, m_k > 0 \end{aligned} \tag{1}$$

where $\mathcal{D}(a, R^k)$ represents the distance between the anchor image and any reference image from grain k. There are k hyper-parameters to control the distance margins across multiple grains. It is worth noting that the distance constraints across multiple grains can be transmitted in Eq.1. Triplet [24, 26, 25] and quadruplet [30] are special cases of this formulation corresponding to $n = 2$ and $n = 3$ respectively. Evidently, it properly reflects multi-grain constraints of distances in feature space.

Treating image relationship as multi-grain is more reasonable when multiple attributes are available for each training image. A toy example is illustrated in Fig.2. The distance of images belonging to the same vehicle maybe close to the distance of images belonging to different vehicles under ID constrained classification, as shown in the left image of Fig.2. Cooperating with conventional ranking methods (i.e., pairwise or triplet ranking) can alleviate the problem and achieve some gains. However, the pairwise and triplet ranking only distinguish whether the images come from the same vehicle as illustrated in the middle image of Fig.2. With multi-grain constraints, we can suitably reflect more accurate relationships in feature space as illustrated in the right image of Fig.2.

Although Eq.1 reveals the objective of multi-grain constraints based ranking and shows the generalized property on some ranking models, it is difficult to directly optimize it under so strong constraints. To optimize the objective, we propose two approaches, generalized pairwise ranking and multi-grain based list ranking.

3.2. Generalized Pairwise Ranking

Conventional pairwise ranking method splits training data into positive (similar) pair set and negative (dissimilar) pair set. Suppose that $\{(x_i, x_j, y_{ij})\}$ is a pair of training data, where x_i, x_j are two images and $y_{ij} \in \{0, 1\}$ indicates their relationship. Generally, image pair x_i, x_j are first fed into CNNs to extract high-level features, and then their features are fused through some strategies (e.g., concatenation or element-wise substraction), followed by some fully-connected layers. Finally, contrastive loss or softmax loss can be adopt to minimize the distance between a positive pair and penalize the negative pair distance. It has achieved promising performances in some tasks especially for person Re-ID [21, 20]. However, it is not the best choice for precise vehicle search as it only takes similar and dissimilar relations of images into account.

Given these limitations, we propose the generalized pairwise ranking approach. Fig.3 illustrates its overall framework. Specifically, input images are organised as a MGL. After extracting deep features for each image in the MGL, the feature of each reference image is fused with the anchor feature respectively. Finally, a softmax classifier is used to estimate which grain level each pair of input images belongs to. The loss function of the GPR for a MGL can be formulated as

$$L_{GPR} = -\sum_{i=1}^{N} \log \frac{e^{p(i,g(i))}}{\sum_{k=1}^{n} e^{p(i,k)}}, \tag{2}$$

where $p(i, k)$ represents the grain prediction value of i-th image pair on k-th grain level and $g(i)$ is the ground truth grain of this pair. N is the number of image pairs in a MGL. Despite the generalized pairwise ranking loss, we also integrate multi-attribute classification to form a multi-task learning framework.

In the whole, our method can be regarded as a joint optimization problem. The overall loss function for a MGL can be formulated by

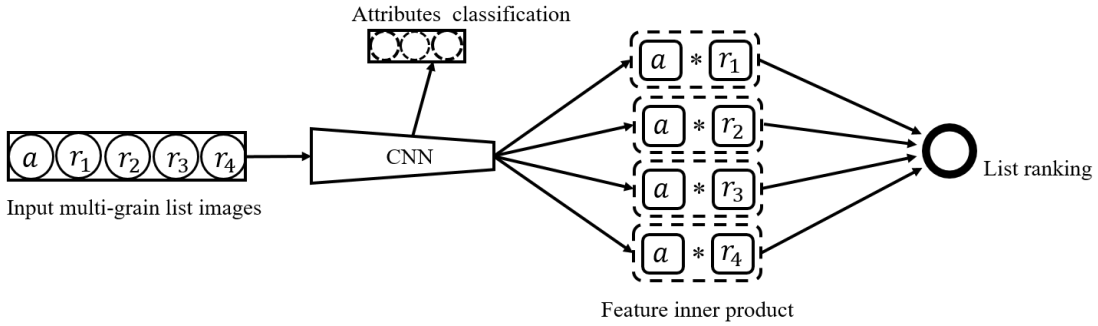


Figure 4. The framework of a multi-task learning integrating MGLR with multi-attribute classification. All images are extracted features in a shared CNN architecture supervised by multi-attribute labels. After calculating the similarity score by inner product for each pair, they are fed into the final layer to conduct list ranking learning.

$$L_{ALL} = -\sum_{y=1}^{h} \lambda_y \sum_{x=1}^{N+1} \log \frac{e^{p_y(x,a(x))}}{\sum_{j=1}^{t_y} e^{p_y(x,j)}} + \lambda L_{GPR}, \quad (3)$$

where $p_y(x, j)$ represents the attribute prediction value of x-th image in a MGL on j-th category of y-th attribute and $a(x)$ is the ground truth attribute of this image. t_y is the number of categories of y-th attribute. Here, we use three types of attributes, namely vehicle ID, model and color. In addition, λ is a hyper parameter to control the balance between similarity learning and attributes learning and λ_y is used to assign weights to different attributes. As the number of image pairs in a MGL is N in Eq.2, the total number of images in a MGL is $N + 1$ by adding the anchor image.

Generalized pairwise ranking adequately exploits multi-grain information and transforms the rank learning problem to a multi-class classification problem. Moreover, it is effective and easy to implement. Implementation details are described in the experimental section.

3.3. Multi-grain based List Ranking

Except for generalized pairwise ranking, we also find that the multi-grain constraints based ranking problem can be formulated as a listwise learning-to-rank problem [7, 8], which is an effective ranking method in document retrieval. In this paper, we design a multi-task learning framework to integrate list ranking with multi-attribute classification.

Fig.4 shows the pipeline of processing a list of multi-grain images. After extracting features for all images, the similarity for each image pair consisting of an anchor image and a reference image can be computed as

$$s(f_a, f_{r_i}) = \frac{1}{2}[1 + (\frac{f_a}{\|f_a\|_2})^T \frac{f_{r_i}}{\|f_{r_i}\|_2}], \quad (4)$$

where $s(f_a, f_{r_i}) \in [0, 1]$ abbreviated as s_i is the similarity score based on the feature of anchor image f_a and the feature of reference image f_{r_i}, which are extracted from the CNN framework.

As for listwise ranking problem, we first introduce the formal definition of a permutation. Suppose that the set of reference images to be ranked are identified with the numbers $1, 2, \cdots, N$. A permutation π of reference images is defined as a bijection from $\{1, 2, \cdots, N\}$ to itself. We denote the permutation as $\pi =< \pi(1), \pi(2), \cdots, \pi(N) >$ in which $\pi(i)$ represents the reference image at ranking position i. The number of all possible permutations for N reference images is $N!$. We assume that any permutation is possible after sorting as the similarity scores between the anchor image and each reference image in descending order. However, different permutations should have different likelihood values according to its consistence with the ranking result of similarity scores. To quantitatively evaluate the likelihood of a permutation, we utilize the permutation probability [7] as its nice properties for representing the likelihood of a ranking list, which can be calculated as

$$P_s(\pi) = \prod_{j=1}^{N} \frac{\phi(s_{\pi(j)})}{\sum_{i=j}^{N} \phi(s_{\pi(i)})}, \quad (5)$$

where $\phi(\cdot)$ is an increasing and strictly positive function and $s_{\pi(j)}$ denotes the similarity score between the anchor image and the reference image ranking at position j in permutation π. Based on the similarity scores of a MGL, each permutation has a probability value and the permutation can obtain the highest likelihood value only if it is the descent sorting sequence as similarity scores. More properties and theorem proving are demonstrated in [7].

Based on the likelihood value of a permutation, a key problem is to learn better features that can derive the ground truth permutation based on similarity scores. Inspired by the analysis of different listwise approaches in [8], we adopt the listMLE method which is an effective method and easy to implement. It employs the negative log likelihood of the ground truth permutation as the loss function (i.e., the likelihood loss). ListMLE is suitable to optimize the multi-grain

constraints based list ranking since the multi-grain relationship is fixed for a MGL and can generate ground truth permutation easily. We formulate the loss function of a list of images as

$$L_{list}(\pi_{gt}) = -\log \prod_{j=1}^{N} \frac{\exp(s_{\pi_{gt}(j)})}{\sum_{i=j}^{N} \exp(s_{\pi_{gt}(i)})}, \qquad (6)$$

where π_{gt} is the ground truth permutation and we adopt the exp form for function $\phi(\cdot)$ in Eq.5. Stochastic gradient descent is used to conduct the minimization. Fig.4 illustrates the multi-grain based list ranking method in a multi-task deep learning framework. Similar to generalized pairwise ranking method, multi-attribute classification also plays an important role here.

4. Large-scale Vehicle Datasets

4.1. Overall Description

In this paper, we construct two large-scale vehicle datasets[1] (i.e., VD1 and VD2) based on real-world unconstrained scenes from two cities respectively. The images in VD1 are obtained from high resolution traffic cameras, and images in VD2 are captured from surveillance videos. We perform vehicle detection on the raw data to make sure that each image only contains one vehicle. The region of plate number has been covered by black color due to privacy protection. All vehicle images are captured from the front view. Some example images are shown in Fig.5.

We provide diverse attribute annotations for each image in both two datasets, including identity number, precise vehicle model and vehicle color. Specifically, identity number (ID) is unique and all images belong to the same vehicle have the same ID (we make sure that there are at least two images in the dataset for each vehicle ID). We provide the most precise model type with detailed vehicle type and different produced years. For example, *Audi-A6L-2012_2015, Audi-A6-2004, Audi-A4-2006_2008 and Audi-A4-2004_2005* are four different vehicle models in our datasets. As for color information, 11 common colors are annotated in our datasets. We carefully check all annotations to ensure the consistency of labels so that all the images belonging to the same vehicle ID are annotated with the same vehicle model and color. To keep the datasets generalized for fine-grained classification tasks, we also ensure that at least two vehicles exist for each precise vehicle model.

4.2. Data Statistics and Split

VD1: There are total 1,097,649 images in the dataset. We label 1,232 vehicle models and 11 colors. After subtracting those improper images (e.g., models containing only one vehicle or images captured from the back), 846,358 images of 141,756 vehicles are remained in VD1.

[1]Available at http://pkuml.org/resources/pku-vds.html

Figure 5. VD1 and VD2 datasets. Left images come from VD1, and right images come from VD2. Images in the same row belong to the same vehicle.

To generalize to other vehicle related tasks, we split dataset into training set and testing set. Specifically, we randomly choose nearly half of vehicles from each vehicle model to construct the training set. The remaining vehicles constitute the testing set for classification task. The detail information is shown in Table.1. For the vehicle search task, we randomly select 2,000 vehicle IDs in the testing set, then randomly select one image from each vehicle to form the query list. Similar to VehicleID [1] dataset, we form reference sets in three scales (i.e., small, medium and large) as shown in Table.2. However, the number of images for each scale is significantly increasing compared with VehicleID. For the largest reference set, both improper images and images of training set are included in order to augment the search scale. Note that all the query vehicles have no overlapping with training set.

VD2: There are total 807,260 images in the dataset. We label 1,112 vehicle models and 11 colors. The split strategy for this dataset is the same as VD1's. The split for training set and testing set are shown in Table.1 and the split for different scale reference sets are shown in Table.2.

Table 1. Data split for training set and testing set.

		Training	Testing
VD1	No. of vehicle	70,591	71,165
	No. of model	1,232	1,232
	No. of color	11	11
	No. of image	422,326	424,032
VD2	No. of vehicle	39,619	40,144
	No. of model	1,112	1,112
	No. of color	11	11
	No. of image	342,608	347,910

Table 2. Reference sets in different scales.

dataset	Small	Medium	Large
VD1	106,887	604,032	1,097,649
VD2	105,550	457,910	807,260

5. Experiments

We conduct two parts of experiments. The first part is to validate the effectiveness of multi-attribute classification for precise vehicle search. Then, we conduct experiments to prove the effectiveness of the proposed two multi-grain constraints based ranking methods.

5.1. Multi-attribute Classification

Multi-attribute of a vehicle can naturally provide strong constraints to learn discriminative feature for vehicle search. However, they, especially the ID information, are ignored for a long time. To validate the the efficiency of multi-attribute classification, we train several deep models with the softmax classifier.

Experiment settings: Experiments are conducted based on the widely-used *Caffe* [31] deep learning framework. We add a softmax classification layer after *fc7* layer to train attributes classification in *VGG_CNN_M_1024* network [32]. Note that the network is fine-tuned on the weights pre-trained with the ImageNet [33] dataset rather than trained from scratch. All the settings are same with [1]. In addition, we also follow the data split strategy in VehicleID. The loss weights of ID, model and color are set to 1.0, 1.0, 1.0 respectively. In test phase, we extract the features in *fc7* layer of test images, then directly utilize L2 distance to measure the similarity between a query image and a reference image. Mean average precision(mAP) is used to evaluate the performance of different strategies.

Table 3. Attributes classification on VehicleID dataset.

Mehotd	Small	Medium	Large
CCL (CVPR16 [1])	0.546	0.481	0.455
ID(VGG)	0.597	0.598	0.552
ATTs(VGG)	0.625	0.623	0.575
ATTs(GoogLeNet)	0.628	0.623	0.586

Results and analysis: Experimental results are shown in Table 3. ATTs in Table 3 represents multi-attribute classification. In all scale test sets, the performance of ID supervised classification significantly outperforms CCL which is the state-of-the-art method in VehicleID. Note that attributes classification including only vehicle model is also used in [1] to corporate with the ranking method. Unfortunately, ID label, which can also be treated as a special attribute, is ignored. The result in second row demonstrates its effectiveness for vehicle search.

To learn the more robust feature representation, model and color information are added to constitute a multi-attribute classification task. The experimental result in the third row of Table.3 shows that multi-attribute supervised classification further boosts the performance in all scales. Benefited from the model and color information, ID can better separate different vehicles especially for the case that vehicles have similar appearances but different models or colors. The results prove that multi-task learning is an authentically effective strategy to strengthen the feature representation with multiple constraints compared with single task learning.

In order to validate the performance of different network architectures, we also conduct the experiment of multi-attribute classification on GoogLeNet [14] which has more layers. The result in the fourth row of Table.3 shows that GoogLeNet achieves the comparable performance with VGG in small and medium sets, but outperforms VGG in the large set. Considering that the deeper network can learn better representation in a large-scale training set, we adopt GoogLeNet as our basic network in the rest of experiments.

5.2. Multi-grain Constraints based Ranking

Based on the diverse attributes of a vehicle, we leverage four grains to represent relations between an anchor image and reference images. Specifically, a reference image must belong to one of the following categories: 1) being the same vehicle with the anchor image (i.e., having the same ID). 2) being the same model and color with the anchor image, but belonging to different vehicles. 3) being the same model but the different color with the anchor image. 4) others.

We first utilize multi-attribute to train initial weights for GoogLeNet, which is prepared to conduct ranking learning. As our target is to learn the discriminative feature to facilitate precise vehicle search with abundant training data, we attempt to train the GoogLeNet from scratch instead of using weights pre-trained in ImageNet. The vehicle model label is used to constrain the two auxiliary classification losses in GoogLeNet. After the initial weights are trained, all the experiments about ranking learning are fine-tuned on those weights.

Experiment settings: For the proposed two methods, the smallest input unit is a MGL containing five images including one anchor image and four reference images from four grains respectively. For GPR, we adopt feature concatenation as the fusion strategy. The batch size of the network is set to 75. The initial learning rate is set to 0.01 for new layers of ranking learning and 0.001 for others. The learning rate decay factor is 0.96 for every 4,000 iterations. The weight decay factor is set to 0.0002. As for MGLR, all settings are the same with GPR except for the batch size, which is set to 90.

To compare with existing methods, we implement two ranking methods in our learning framework, including general pairwise ranking and triplet ranking. We also attempt

to implement the method [30] which embeds hierarchical relations on feature learning, but it is not converged as the strong constraints are hard to optimize directly. Note that all the procedures are the same with our methods except for the ranking method. In the test phase, we extract features from the *pool5_7x7_s1* layer. We evaluate the performance with mAP in all experiments.

Results and analysis: The performances of all methods on VD1 and VD2 are shown in Table.4 and Table.5, respectively. Several conclusions can be drawn from the results. First, considering the first and second rows in the two tables which are results of multi-attribute classification, the performance of the model trained from scratch significantly outperforms the model pre-trained in ImageNet. The results conflict with conventional experience as we have a mass of training data (422,326 images in VD1 and 342,608 images in VD2) enough to learn vast parameters from zero. On the contrary, the initial parameters learned from a large universal dataset may influence the performance since a large domain gap exists between ImageNet and vehicle dataset.

Table 4. The performance of precise vehicle search on VD1 dataset.

Methods	Small	Medium	Large
Fine-tune for ATTs	0.492	0.285	0.239
New model for ATTs	0.734	0.532	0.461
ATTs + Pairwise [22]	0.747	0.546	0.474
ATTs + Triplet [26]	0.759	0.556	0.482
ATTs + GPR	0.776	0.575	0.501
ATTs + MGLR	**0.791**	**0.583**	**0.511**

Table 5. The performance of precise vehicle search on VD2 dataset.

Methods	Small	Medium	Large
Fine-tune for ATTs	0.553	0.379	0.317
New model for ATTs	0.685	0.544	0.492
ATTs + Pairwise [22]	0.692	0.567	0.517
ATTs + Triplet [26]	0.710	0.575	0.523
ATTs + GPR	0.717	0.588	0.537
ATTs + MGLR	**0.747**	**0.606**	**0.553**

Second, on the basis of powerful representation from multi-attribute classification, pairwise and triplet ranking methods can achieve some improvements (e.g., 0.01 for pairwise and 0.02 for triplet in terms of mAP) in precise vehicle search, which indicates the effectiveness of multi-task learning integrating attributes classification and ranking learning. The results also demonstrate that ranking methods indeed facilitate feature representation in CNNs. Additionally, triplet ranking achieves slightly better performance compared to pairwise ranking.

Third, our proposed two methods achieve promising performances. From the third and fifth rows in the two tables, the GPR method surpasses standard pairwise ranking method by about 0.03 in VD1 and 0.02 in VD2 in terms of mAP. It also outperforms the triplet ranking method in all scales of reference sets. These results strongly suggest that multi-grain constraints can effectively facilitate similarity learning for vehicle search. Furthermore, MGLR method achieves the state-of-the-art performance in both two datasets. It shows that the permutation probability model based ranking method is more effective for precise vehicle search. The reason may be that it can directly optimize the permutation of a multi-grain list images well.

6. Conclusion

In this paper, we focus on the problem of precise vehicle search, which aims at finding out the images belonging to exactly the same vehicle with the query image. To address the problem, we first summarize the relationship between different vehicle images as multiple grains by using diverse attributes of vehicles. Based on the multi-grain constraints, we further propose two ranking methods, generalized pairwise ranking and multi-grain based list ranking, which are incorporated with multi-attribute classification in a unified deep learning framework. To further facilitate the research on this problem, we contribute two high-quality and well-annotated vehicle datasets, which are the largest vehicle datasets so far. Experimental results show that our methods achieve promising performances on the new datasets.

Acknowledgments

This work is partially supported by grants from the National Basic Research Program of China under grant 2015CB351806, and the National Natural Science Foundation of China under contract No. U1611461, No. 61425025, No. 61633002, and No. 61471042.

References

[1] H. Liu, Y. Tian, Y. Yang, L. Pang, and T. Huang. Deep relative distance learning: Tell the difference between similar vehicles. In *CVPR*, 2016.

[2] J. Krause, M. Stark, J. Deng, and L. Fei-Fei. 3d object representations for fine-grained categorization. In *ICCV workshop*, 2013.

[3] L. Yang, P. Luo, Change L. C., and X. Tang. A large-scale car dataset for fine-grained categorization and verification. In *CVPR*, 2015.

[4] J. Sochor, A. Herout, and J. Havel. Boxcars: 3d boxes as cnn input for improved fine-grained vehicle recognition. In *CVPR*, 2016.

[5] X. Liu, W. Liu, T. Mei, and H. Ma. A deep learning-based approach to progressive vehicle re-identification for urban surveillance. In *ECCV*. Springer.

[6] L. Zheng, Y. Yang, and A. G. Hauptmann. Person re-identification: Past, present and future. *arXiv preprint arXiv:1610.02984*, 2016.

[7] Z. Cao, T. Qin, T. Liu, M. Tsai, and H. Li. Learning to rank: from pairwise approach to listwise approach. In *ICML*. ACM, 2007.

[8] F. Xia, T. Liu, J. Wang, W. Zhang, and H. Li. Listwise approach to learning to rank: theory and algorithm. In *ICML*. ACM, 2008.

[9] H. Jégou, M. Douze, C. Schmid, and P. Pérez. Aggregating local descriptors into a compact image representation. In *CVPR*. IEEE, 2010.

[10] F. Perronnin, J. Sánchez, and T. Mensink. Improving the fisher kernel for large-scale image classification. In *ECCV*. Springer, 2010.

[11] G. Chechik, V. Sharma, U. Shalit, and S. Bengio. Large scale online learning of image similarity through ranking. *JMLR*, 2010.

[12] K. Q. Weinberger and L. K. Saul. Distance metric learning for large margin nearest neighbor classification. *JMLR*, 2009.

[13] J. Wan, D. Wang, S. C. H. Hoi, P. Wu, J. Zhu, Y. Zhang, and J. Li. Deep learning for content-based image retrieval: A comprehensive study. In *MM*. ACM, 2014.

[14] C. Szegedy, W. Liu, Y. Jia, P. Sermanet, S. Reed, D. Anguelov, D. Erhan, V. Vanhoucke, and A. Rabinovich. Going deeper with convolutions. In *CVPR*, 2015.

[15] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. In *NIPS*, 2015.

[16] J. Long, E. Shelhamer, and T. Darrell. Fully convolutional networks for semantic segmentation. In *CVPR*, 2015.

[17] J. Wang, Y. Song, T. Leung, C. Rosenberg, J. Wang, J. Philbin, B. Chen, and Y. Wu. Learning fine-grained image similarity with deep ranking. In *CVPR*, 2014.

[18] S. Chopra, R. Hadsell, and Y. LeCun. Learning a similarity metric discriminatively, with application to face verification. In *CVPR*. IEEE, 2005.

[19] Y. Taigman, M. Yang, M. A. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[20] E. Ahmed, M. Jones, and T. K. Marks. An improved deep learning architecture for person re-identification. In *CVPR*, 2015.

[21] W. Li, R. Zhao, T. Xiao, and X. Wang. Deepreid: Deep filter pairing neural network for person re-identification. In *CVPR*, 2014.

[22] Y. Sun, Y Chen, X. Wang, and X. Tang. Deep learning face representation by joint identification-verification. In *NIPS*, 2014.

[23] D. Yi, Z. Lei, S. Liao, and S. Z. Li. Learning face representation from scratch. *arXiv preprint arXiv:1411.7923*, 2014.

[24] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[25] J. Huang, R. S. Feris, Q. Chen, and S. Yan. Cross-domain image retrieval with a dual attribute-aware ranking network. In *ICCV*, 2015.

[26] S. Ding, L Lin, G. Wang, and H. Chao. Deep feature learning with relative distance comparison for person re-identification. *Pattern Recognition*, 2015.

[27] Y. Lin, V. I. Morariu, W. H. Hsu, and L. S. Davis. Jointly optimizing 3d model fitting and fine-grained classification. In *ECCV*. Springer, 2014.

[28] K. Ramnath, S. N. Sinha, R. Szeliski, and E. Hsiao. Car make and model recognition using 3d curve alignment. In *WAVC*. IEEE, 2014.

[29] Z. Zhang, T. Tan, K. Huang, and Y. Wang. Three-dimensional deformable-model-based localization and recognition of road vehicles. *TIP*, 2012.

[30] X. Zhang, F. Zhou, Y. Lin, and S. Zhang. Embedding label structures for fine-grained feature representation. In *CVPR*, 2016.

[31] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *MM*. ACM, 2014.

[32] K. Chatfield, K. Simonyan, A. Vedaldi, and A. Zisserman. Return of the devil in the details: Delving deep into convolutional nets. *BMVC*, 2014.

[33] J. Deng, W. Dong, R. Socher, L. Li, K. Li, and L. Fei-Fei. Imagenet: A large-scale hierarchical image database. In *CVPR*. IEEE, 2009.