# MULTI-MODAL METRIC LEARNING FOR VEHICLE RE-IDENTIFICATION IN TRAFFIC SURVEILLANCE ENVIRONMENT

*Yi Tang, Di Wu, Zhi Jin, Wenbin Zou*, Xia Li*

Shenzhen Key Lab of Advanced Telecommunication and Information Processing
College of Information Engineering, Shenzhen University

## ABSTRACT

Vehicle re-identification (Re-Id) aims to retrieve the same vehicle captured by disjoint cameras at different time instants from different locations, and is a challenging task mainly due to the high similarity among the captured vehicle images in surveillance environment. With the rapid development of Convolutional Neural Network (CNN), learning-based deep features have been adopted to combine with hand-crafted features to re-identify vehicles in traffic surveillance environment. However, the two kinds of features are in different feature space, and if they are fused directly together, their complementary correlation is not able to be fully explored. To address such an issue, this paper proposes a multi-modal metric learning architecture to fuse deep features and hand-crafted ones in an end-to-end optimization network, which achieves a more robust and discriminative feature representation for vehicle re-identification. The extensive experiments on a large-scale traffic surveillance vehicle dataset demonstrate that our proposed approach substantially outperforms the state-of-the-art methods on vehicle Re-Id.

***Index Terms***— Vehicle Re-identification, Convolutional Neural Network, Multi-Modal Architecture, Feature fusion

## 1. INTRODUCTION

Vehicle Re-Id is to find the target vehicle from multiple cameras captured images. Based on some discriminative features from the given vehicle image, vehicle Re-Id algorithm is used for comparing the matched features among the whole database and to detect the suspect vehicles. Hence, it is meaningful to intelligent transportation [1], urban computing [2], arterial travel time estimation [3] and even security assurance.

Different from other vehicle related works, such as vehicle classification [4], detection [5] and tracking [6] which have attracted extensive attention for many years, Vehicle Re-Id, especially in surveillance system, still faces several significant challenges. Firstly, it is the lack of proper vehicle image dataset captured from surveillance environment. Since the video surveillance data relates to daily life security, it is hard to access surveillance system freely. Although some car datasets have been proposed, for example CompCars [4], whose vehicle images mainly come from the internet and partly from surveillance system, they are mainly for fine-grained classification and attribute prediction and not suitable for vehicle Re-Id task. Secondly, the captured images/videos from the surveillance system has limited resolution and suffer from blurring and variation of illumination. Finally, in real surveillance system vehicles could be captured in various viewpoints and distances. With the emergence of the VeRi dataset [7], which is totally collected from the real traffic surveillance scenes, the first difficulty is somehow remitted and more works begin to focus on this field in recent years. However, these kinds of difficulties still make the traditional hand-crafted features, such as appearance feature, Local Binary Pattern (LBP) and scale-invariant feature transform (SIFT), not discriminative enough for vehicle Re-Id.

Currently, with the resurgence of Convolutional Neural Network (CNN), the learning-based deep features have been employed, combining with the hand-crafted features for vehicle Re-Id [7, 8]. However, their late fusion strategy, which combines the similarity scores from different features, does not fully explore the complementary correlation among different types of features. Therefore, a multi-modal architecture is proposed to address this problem and this architecture achieves good performance on video classification [9], object recognition [10], and gesture recognition [11, 12, 13].

Motivated by these works, and aiming to obtain more robust and discriminative fusing features, this paper proposes a multi-modal metric learning architecture to integrate the hand-crafted and learning-based deep features. Our method integrates two kinds of hand-crafted features, a LBP feature map and a Bag-of-Word-based Color Name (BoW-CN) feature. The former binding with the raw RGB image converts to a four-channel image, and is fed as an input of the multi-modal network. The BoW-CN is a high-dimensional feature vector by using the codebook from [14] and has been proved as one of the best hand-crafted features in some re-identification works [7, 14]. In our proposed architecture, BoW-CN is used as the other input of the network. Through the training of multi-modal network, we can obtain the fusing feature representations, which can retain complementary correlation and are more robust as well as discriminative for

* Wenbin Zou is the corresponding author. E-Mail: zouszu@sina.com

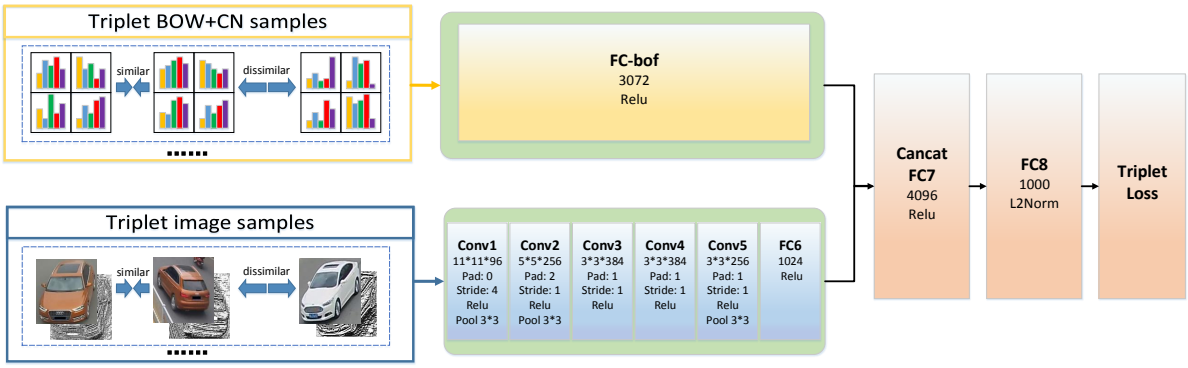


**Fig. 1**. The framework of our proposed multi-modal metric learning architecture for vehicle re-identification. At the input side, there are two types of input data, where the first one includes the BoW-CN features for the input of Multi-Layer Perceptron (MLP), and the second contains the four-channel images for the input of CNN. The middle shows the parameters of network structure. At the end, the triplet loss function is used for optimizing the parameters of the whole architecture.

vehicle Re-Id.

Meanwhile, different from traditional deep learning architecture, we exploit a metric learning network for the feature learning. Among the traditional structures of metric learning, Siamese work is shown to achieve good performance [15]. The recent triplet network [16, 17] is employed as our architecture, since it is able to decrease the intra-class distance and increase the inter-class distance simultaneously with its triplet samples.

In summary, this paper has three contributions as follows:

1) LBP feature map is introduced to complement RGB image to form multi-channel inputs for high-level semantic feature learning.

2) Furthermore, a multi-modal metric learning architecture is proposed to fuse deep features and hand-crafted ones in an end-to-end optimization network, which achieves a more robust and discriminative feature representation for vehicle re-identification.

3) Last, the comprehensive evaluations demonstrate that our proposed approach outperforms the state-of-the-art methods by a large margin on a large-scale vehicle dataset in real traffic surveillance.

## 2. THE PROPOSED METHOD

### 2.1. Overview

As shown in Fig. 1, our framework employs a deep learning-based multi-modal architecture to obtain the feature representations for vehicle re-identification. As for the raw images, the pre-processing is implemented in two steps: 1) We firstly extract the LBP feature map from the raw vehicle image. Then each RGB image is combined with its corresponding LBP feature map as multi-channel image, which is one of training data of multi-modal architecture. 2) By using the codebook from [14], the BoW-CN feature vectors are extracted from the raw images. These vectors are the other input of the network.

Our network applies the triplet loss as the optimization function. The input training data need to be converted into triple samples, and each sample contains a reference instance, a positive instance and a negative instance. The reference instance is similar with the positive one and dissimilar with the negative one. After pre-processing, the two kinds of input data can be trained by a Multi-Layer Perceptron (MLP) and a CNN respectively. We set a concatenation layer for integrating the dual-modal outputs from MLP and CNN, and use the triplet loss function to optimize the whole network. In the test stage, the testing vehicle images are fed into the network, and the output of the penultimate fully-connected layer is used as the feature representation. At last, the similarity between the features of vehicle images can be measured by Euclidean distance for re-identification.

### 2.2. CNN for image data

Different from only using RGB three channel images as training data of CNN model, our method utilizes four-channel images, including a channel of texture feature map and the RGB channels, as the input of CNN modality. Although texture features can be extracted in front convolutional layer of CNN model, the raw image combining with LBP feature map can improve the quality of texture feature in front layers of network and make the network learn more robust semantic features in later layers.

In our approach, the LBP feature are extracted as follows:

$$\begin{aligned} LBP(x_c, y_c) &= \sum_{j=0}^{N-1} 2^j * h(j_i - j_c) \\ s.t. \quad h(x) &= \begin{cases} 1 & if \quad x \geq 0 \\ 0 & otherwise \end{cases} \end{aligned} \tag{1}$$

where $(x_c, y_c)$ is central pixel in a $3 \times 3$ window, $j_i$ and $j_c$ represent gray values of the central pixel and its N-neighbor pixels. $h(x)$ is sign function. According to Eq.1, the LBP feature map can be obtained, and then we directly append it with

the corresponding RGB image to construct a four-channel image.

### 2.3. MLP for BoW-CN vector

BoW-CN feature is a color based feature, which combine bag-of-word model and Color Name (CN) features. As for Bag-of-Word model, it needs to train a codebook with the image samples. In [14], the codebook is trained by the k-means algorithm. For each image sample, it is divided into $4 \times 4$ patches. The CN features are extracted from these patches. Then the average inverse document frequency, weak geometric constrains, multiple assignment and background suppression are applied to refine the BoW-CN features. After that, we can obtain a 5600-dimension feature vector from a resized $64 \times 128$ test image by using the pre-training Bag-of-Word model. In our approach, we treat the BoW-CN features as an input of MLP and combine its output with deep features in our multi-modal architecture.

### 2.4. The training details of Multi-modal Architecture

In this section, we describe the training details of the proposed multi-modal deep learning architecture. The training details contain four components: loss function, pre-training and fusion process.

**Loss function.** In our deep metric learning architecture, the triplet loss function is employed to optimize the model. Therefore, the triplet samples should be produced for training. Given a vehicle image, its LBP feature map $L_i$ firstly is extracted, then combined with the source RGB image $I_i$ and BoW-CN features $B_i$ to make up a reference sample $\{v_i|v_i = (I_i^v, L_i^v, B_i^v), i \in N^+\}$, where $N^+$ is the set of positive integers. The reference sample combining with a positive sample $\{p_i|p_i = (I_i^p, L_i^p, B_i^p), i \in N^+\}$ that has the same label of the reference sample and a negative sample $\{n_i|n_i = (I_i^n, L_i^n, B_i^n), i \in N^+\}$ that has a different label forms a triplet input data $(v_i, p_i, n_i)$. Through the network, the reference sample $v_i$ (equally for $p_i$ and $n_i$) is projected into a vector $f(v_i) \in R^d$, where $d$ represents the feature dimension. The main idea of the triplet loss function is increasing the distance between the reference samples and negative samples and decreasing the distance between the reference samples and positive samples. Given the triplet projected vector $(f(v_i), f(p_i), f(n_i))$, the triplet loss function can be defined as follows:

$$L(v_i, p_i, n_i) = \sum_{i=1}^{M} [\|f(v_i) - f(p_i)\|_2^2 - \|f(v_i) - f(n_i)\|_2^2 + \alpha]_+ \quad (2)$$

where $\|\cdot\|_2$ is $l_2$-normalization between two vectors. The $[\cdot]_+$ is the max function, $M$ is the number of the triplet samples and $\alpha$ is a constance. In our triplet network, $\alpha$ is set to 0.5.

**Pre-training.** Since the BoW-CN features are hand-crafted features, which have different feature space with learning features of CNN modality, if we forcibly integrate them without initialization, the multi-modal network with high possibility fails in the training stage. Therefore, in order to address this issue, we apply the pre-training strategy into our network training. In the implementation, the CNN modality and the MLP modality are respectively trained by their corresponding triple inputs. Then the initial learning rate is reduced and the fusing multi-modal architecture is trained for the second time.

**Fusion process.** The fusing method of the modalities we adopt in this architecture is a concatenation layer, which is straightforward but effective. Here, the concatenation layer can be represented as:

$$Y = h(\sum_{i=1}^{M} W_i * X_i) \quad (3)$$

where $X_i$ is the projected vector of previous layer from the $i$-th modality, $W_i$ is its corresponding weight parameter, $M$ is the number of modalities, here $M$ is 2. In the process of implementation, $X_1$ corresponds the output of FC-bof in Fig. 1 and $X_2$ corresponds the output of FC6. In our architecture, the units of FC-bof and FC6 are set to 3072 and 1024, respectively. $h(\cdot)$ is the activation function, which is Relu function in our network, $Y$ is the concatenating output in this layer.

## 3. EXPERIMENTS

### 3.1. Experiments setup

To validate the proposed approach, we conduct the experiments on VeRi [7], a large-scale urban surveillance vehicle dataset for re-identification. This dataset contains 49,375 images of 776 vehicles, each of which is collected by 20 cameras in real-world traffic surveillance environment. The 37,778 images of 576 vehicles are in the training set, whereas the remaining images are in the test set. The cross-camera search [7] is introduced to evaluate our approach. The rule is to search the same vehicle images from different cameras in the test set.

For performance evaluation, we use the cumulative matching characteristic (CMC) curve and the mean average precision (mAP). The CMC curve is a traditional evaluation criterion in re-identification field. The mAP reflects average precision of overall queries. Meanwhile, Rank-1 and Rank-5 identification rates are also employed to evaluate our experiments, which represent the percentages of the right results that appear within the first and fifth ranks, respectively.

### 3.2. Evaluation with vehicle Re-Id approaches

In order to show the effectiveness of our proposed approach, we compare our approach with several state-of-the-art methods as following on VeRi dataset:

1) **BoW-CN** [14]. It is a Bag-of-Word-based method, which uses hand-crafted features for vehicle Re-Id.

**Table 1**. Comparison results of different approaches with evaluation measures of mAP, Rank-1 and Rank-5 on VeRi dataset.

| Method | mAP | Rank-1 | Rank-5 |
|---|---|---|---|
| BoW-CN [14] | 9.03 | 33.31 | 48.69 |
| AlexNet [18] | 11.59 | 31.23 | 45.47 |
| FGCV [4] | 17.03 | 58.22 | 71.57 |
| FACT [7] | 17.38 | 57.75 | 71.93 |
| BoW-CN+MLP | 12.75 | 36.17 | 51.78 |
| RGB image+CNN | 27.97 | 51.66 | 71.09 |
| RGB image+LBP map+CNN | 29.01 | 54.11 | 70.38 |
| RGB image+CNN+BoW-CN+MLP | 31.64 | 58.87 | 76.69 |
| Combining Network | **33.78** | **60.19** | **77.40** |

2) **AlexNet** [18]. This approach is a deep learning model for ImageNet classification. The features extracted from its penultimate fully-connected layer are used for computing the similarity among the vehicle images.

3) **FGCV** [4]. This method employs the GoogLeNet model [19] to fine-tune with their CompCars dataset. In our experiment, we utilize their fine-tuning model to extract deep features (Pooling5 layer) as the representation of the vehicle images.

4) **FACT** [7]. This is late fusing feature method, which integrates GoogLeNet deep features, color and texture hand-crafted features.[1]

5) **BoW-CN + MLP**. This is one modality in our network, which combines the BoW-CN features and MLP model.

6) **RGB image + CNN**. The scheme only uses the RGB images as the input of convolutional network. In order to compare one of modality in our network, the network structure is the same with our CNN modality.

7) **RGB image + LBP map + CNN**. This is the other modality in our network, whose input is a multiple channel image. The image contains RGB channels and a LBP feature channel.

8) **RGB image + CNN + BoW-CN + MLP**. This scheme is comparative experiment, which contains all of input data except LBP feature map.

9) **The combining network**. The proposed approach in this paper contains the two modalities and the final model is with the pre-training strategy. In this experiment, the feature we extracted to compute the similarity is from the penultimate fully-connected layer, which is a 4096-dimension vector.

The results of the of the above methods with the evaluations of mAP, Rank-1 and Rank-5 are presented in Table 1. Similarly, the CMC curves of them are shown in the Fig. 2. According to the results, we can observe that the BoW-CN, that are hand-crafted features, can surpass the deep features extracted from the pre-training model of AlexNet in Rank-1 and Rank-5. AlexNet, FGCV and FACT explore deep learning to obtain the robust deep features. Especially, the FACT

[1]In our experiments, we carefully re-implement their algorithm because the code is not public available

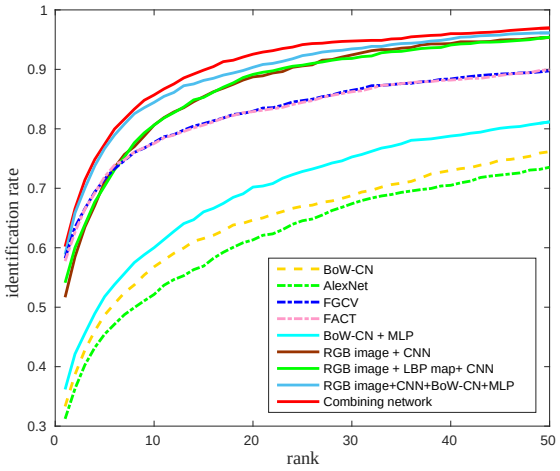


**Fig. 2**. Comparison of the vehicle Re-Id methods in terms of CMC curves.

implements the late fusion strategy with the integration of advantages of the deep features and hand-crafted features, which conduct the best performance on VeRi dataset.

In our architecture, we firstly integrate the BoW-CN features and the MLP model. Compared with only using the BoW-CN features, this modality can refine the BoW-CN features and has a slight improvement. Then, the modality of integrating the RGB image and CNN is tested and this modality achieve good performance. The result has surpassed the FACT in mAP, but it is not good enough in Rank-1 and Rank-5. Following this scheme, we combine the RGB image and its LBP feature map as the input data of CNN. The performance in mAP and Rank-1 are also improved. As a comparative experiment, all of input data except LBP map is set on our multi-modal framework. At last, the combining network achieves 33.78% on mAP, 60.19% on Rank-1 and 77.40% on Rank-5. The results outperform the state-of-the-art by a large margin. The results demonstrate that our method not only retains characteristics of the different features, but also obtains their complementary correlation through the multi-modal architecture, and then produces a more robust as well as discriminative fusing features for vehicle Re-Id.

## 4. CONCLUSION

In this paper, we propose a multi-modal learning architecture to fuse the hand-crafted and learning-based features in an end-to-end optimization network for vehicle re-identification. Furthermore, we employed the triplet network to obtain robust and discriminative feature representations. The experiments on the vehicle Re-Id dataset from real traffic surveillance system demonstrated that our approach achieves higher performance compared to other state-of-the-art approaches.

**Acknowledgment** This work is supported by the National Natural Science Foundation of China under Grand No.61401287 and the Natural Science Foundation of Shenzhen under Grant No.JCYJ20160307154003475, No.JCYJ-20160506172651 25.

## 5. REFERENCES


[1] Junping Zhang, Fei-Yue Wang, Kunfeng Wang, Wei-Hua Lin, Xin Xu, and Cheng Chen, "Data-driven intelligent transportation systems: A survey," *IEEE Transactions on Intelligent Transportation Systems*, vol. 12, no. 4, pp. 1624–1639, 2011.

[2] Yu Zheng, Licia Capra, Ouri Wolfson, and Hai Yang, "Urban computing: concepts, methodologies, and applications," *ACM Transactions on Intelligent Systems and Technology*, vol. 5, no. 3, pp. 38, 2014.

[3] Karric Kwong, Robert Kavaler, Ram Rajagopal, and Pravin Varaiya, "Arterial travel time estimation based on vehicle re-identification using wireless magnetic sensors," *Transportation Research Part C: Emerging Technologies*, vol. 17, no. 6, pp. 586–606, 2009.

[4] Linjie Yang, Ping Luo, Chen Change Loy, and Xiaoou Tang, "A large-scale car dataset for fine-grained categorization and verification," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 3973–3981.

[5] Daeho Ha, Jongcheol Lee, and Yong Deak Kim, "Neural-edge-based vehicle detection and traffic parameter extraction," *Image and Vision Computing*, vol. 22, no. 11, pp. 899–907, 2004.

[6] Rogerio Schmidt Feris, Behjat Siddiquie, James Petterson, Yun Zhai, Ankur Datta, Lisa M. Brown, and Sharath Pankanti, "Large-scale vehicle detection, indexing, and search in urban surveillance videos," *IEEE Transactions on Multimedia*, vol. 14, no. 1, pp. 28–42, 2012.

[7] Xinchen Liu, Wu Liu, Huadong Ma, and Huiyuan Fu, "Large-scale vehicle re-identification in urban surveillance videos," in *Proceedings of the IEEE International Conference on Multimedia and Expo*, 2016, pp. 1–6.

[8] Xinchen Liu, Wu Liu, Tao Mei, and Huadong Ma, "A deep learning-based approach to progressive vehicle re-identification for urban surveillance," in *Proceedings of the European Conference on Computer Vision*, 2016, pp. 869–884.

[9] Jiquan Ngiam, Aditya Khosla, Mingyu Kim, Juhan Nam, Honglak Lee, and Andrew Y Ng, "Multimodal deep learning," in *Proceedings of the 28th International Conference on Machine Learning*, 2011, pp. 689–696.

[10] Yanhua Cheng, Xin Zhao, Rui Cai, Zhiwei Li, Kaiqi Huang, and Yong Rui, "Semi-supervised multimodal deep learning for rgb-d object recognition," in *Proceedings of International Joint Conference on Artificial Intelligence*, 2016.

[11] Natalia Neverova, Christian Wolf, Graham Taylor, and Florian Nebout, "Moddrop: adaptive multi-modal gesture recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, vol. 38, no. 8, pp. 1, 2014.

[12] Di Wu, Lionel Pigou, Pieter Jan Kindermans, and L. E. Nam, "Deep dynamic neural networks for multimodal gesture segmentation and recognition," *IEEE Transactions on Pattern Analysis and Machine Intelligence*, pp. 1–1, 2016.

[13] Eunbyung Park, Xufeng Han, Tamara L. Berg, and Alexander C. Berg, "Combining multiple sources of knowledge in deep cnns for action recognition," in *Proceedings of the IEEE Winter Conference on Application of Computer Vision*, 2016, pp. 1–8.

[14] Liang Zheng, Liyue Shen, Lu Tian, Shengjin Wang, Jingdong Wang, and Qi Tian, "Scalable person re-identification: A benchmark," in *Proceedings of the IEEE International Conference on Computer Vision*, 2015, pp. 1116–1124.

[15] Cheng Zhang, Wu Liu, Huadong Ma, and Huiyuan Fu, "Siamese neural network based gait recognition for human identification," in *Proceedings of the IEEE International Conference on Acoustics, Speech and Signal Processing*, 2016, pp. 2832–2836.

[16] Florian Schroff, Dmitry Kalenichenko, and James Philbin, "Facenet: A unified embedding for face recognition and clustering," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 815–823.

[17] Elad Hoffer and Nir Ailon, "Deep metric learning using triplet network," in *International Workshop on Similarity-Based Pattern Recognition*, 2015, pp. 84–92.

[18] Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton, "Imagenet classification with deep convolutional neural networks," *Advances in Neural Information Processing Systems*, vol. 25, no. 2, pp. 2012, 2012.

[19] Christian Szegedy, Wei Liu, Yangqing Jia, Pierre Sermanet, Scott Reed, Dragomir Anguelov, Dumitru Erhan, Vincent Vanhoucke, and Andrew Rabinovich, "Going deeper with convolutions," in *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition*, 2015, pp. 1–9.