Deep Visual Features Matching Method for Vehicle Model Re-Identification

Nikolay Nemcev[1][0000−0003−4801−3284] nicknemcev@gmail.com and Elena Vasilenko[1][0000−0002−1914−7753]apfelrobbe@gmail.com

ITMO University, Saint Petersburg, 197101, Russian Federation www.ifmo.ru

Abstract. In this paper, a study of the existing methods for identifying and comparing the features of objects used in the task of a vehicle model re-identification by its image was conducted, which is one of the most important tasks facing automated traffic control systems, and solved by comparing the vehicle features being verified with a certain set of features obtained by the monitoring system earlier, and deciding whether the compared samples belong to the same vehicle model or to different ones. The article describes the approach for vehicle model re-identification according to its image, based on the method of feature vector extraction, using classification convolutional neural network, and on the criterion for feature vectors matching via the sub-counting corresponding features. The proposed method shows a lower computational complexity than modern analogous approaches, uses smaller feature vector, demonstrates comparable re-identification accuracy in scenarios when the testing data have characteristics that coincide with training ones (similar camera model and level of lighting and noise, models of re-identifiable vehicles are contained in the dataset used for training) and achieves significantly higher relative accuracy in cases when testing data very different from the training dataset. The proposed approach is practically applicable in vehicle re-identification task for highly loaded traffic control systems.

Keywords: Visual data processing · machine learning· convolutional neural networks · feature extraction · feature comparison · Alexnet

1 Introduction

The computer vision algorithms are used to solve various tasks in the automotive industry: from detecting vehicles, measuring its speed and counting its number for creating environmental analysis functions for autonomous moving devices. The task of re-identifying vehicles is one of the most important ones that are being solved with traffic control systems, and it may be expressed through selection of vehicle's features (both each one and some specific group of vehicles of the same model) for its further comparison with some set of previously extracted features in order to determine the conformity of the samples.

The task of re-identifying a car is related to the task of facial re-identifying, however, it also has its own specifics - so in the task of identifying a car it is necessary to ensure a reliable comparison of the car's attributes regardless of the angle of view (front, side, back, at different angles), while face identification is usually done from one angle (generally in full face). Also, different car models can visually differ only from a certain angle (especially if car models from the same manufacturer are compared), and images of the same car, taken from different angles, may contain only few common details, which greatly complicates the task of re-identification [1].

Conditionally, all approaches for re-identifying objects can be divided into ones using classic feature extraction methods [2, 3], and ones based on the approaches of feature extraction using convolutional neural networks [1, 4], meanwhile neural network (based on the use of artificial neural networks) approaches can be divided into ones working according to the classical scheme in which features are determined for each compared image separately and its matching is placed in distinct module [5], and approaches based on the use of Siamese neural networks [6] in which two input images are processed in parallel within the same network, and the similarity metric is calculated directly on the last layer [1, 7].

Classical approaches of extracting and comparing features are hardly practically applicable in the task of verifying a vehicle model due to the fact that in real use cases, comparisons are often made for vehicles taken at a considerable distance, images of which do not have a resolution high enough to highlight a significant number of features [8]; the procedure of comparing images of vehicles is being reduced to counting the number of matches between the received descriptors [9], what does not allow you to adjust the compared objects taken from different angles (sets of features obtained from different angles for a same vehicle model will differ).

Neural network approaches for comparing of object in an image use classification network architecture for extracting feature vectors of objects (f.e. [10, 11]) and process its further comparison in the separate module [5] or on the last network layer in case of siamese networks [6]. Main advantages of this approach include possibility of multi-angle comparison and absence of strict resolution requirements for the compared images. The main disadvantage is the feature extraction module's necessity of training on corresponding data set, for example, for correct identification of characteristics of a particular car model classifier should be trained to classify the most complete set of car models, ideally containing model wanted to be verified, meanwhile the visual characteristics of used dataset should correspond to the real ones as much as possible (for reliable identification of the car model in the night time, used classifier must be trained on a data set containing night photos)[12].

Beside that, the effect of "overfitting" is observed in all neural network approaches, in which some model of the classifier, or of the verifier in this case, demonstrates good results on the training data set, but significantly loses accuracy on the data far different from the training [13]. Nevertheless, both feature set extraction and feature set comparison modules based on machine learning methods are prone to this effect [14], which considerably reduces the universality of such approaches.

The approaches based on usage of siamese networks additionally to problems with “overfitting” have problems with network convergence at training stage caused by the heterogeneity of the input data [15], which complicates its usage in some cases.

The proposed approach combines an approach for extracting of a short feature vector based on the usage of a modified classification network Alexnet [10], trained on a specially prepared data set, and a simple similarity metric that operates on small feature vectors and based on the principle of estimating the number of matching features. The use of this metric is due both to the need to reduce the computational complexity of the task and to optimize the computational process of re-identification of the vehicle model for use in highly loaded traffic control systems, and to reduce the ”overfitting” effect of the machine learning algorithms on stability the operation of a system that processes data significantly different in their characteristics from those used in the process of extracting features of the object training module [13].

The obtained results show that, despite its simplicity, the proposed approach demonstrates the accuracy of solving the problem of car model verification comparable with other modern approaches with higher universality and less high computational complexity.

2 Feature extraction model

The modified network Alexnet [10] was used as feature set extraction module in this paper.

Instead of ReLU (Rectified Linear Unit) which can be calculated as:

$$f(x) = max(0, x), \tag{1}$$

where x is the activation function input value, it was used RReLU (Randomized Rectified Linear Unit) [16] as the activation function:

$$f(x) = \begin{cases} x, & \text{if } x \geq 0, \\ \alpha * x, & \text{otherwise} \end{cases}, \tag{2}$$

where $\alpha \subset U(l, u)$,l ¡ u,$l, u \in [0, 1), U(l, u)$ are some uniform distribution (it was used $U(3, 8)$ distribution, and $\alpha = \frac{2}{l+u}$ at the testing stage).

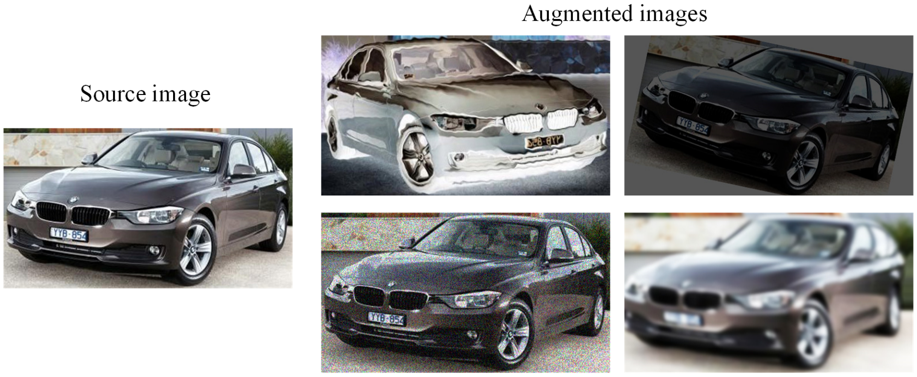
Classical ReLU can break down during training process [16], for example, large gradient passing through ReLU may lead to such an update of weights that the neuron will never be activated again [10]. If it happens, from this moment gradient passing through this neuron will always be evaluated as 0 what negatively affects the effectiveness of classifier training. The researches presented in [16] showed that the usage of some ”loss” for $x < 0$ lets us both decrease the possibility of neuron failing on training stage and some kind decrease ”overfitting” effect of the network due to the random nature of the parameter α.

In order to accelerate the convergence of the network (reduce training time) and increase the stability of its operation it was used a standard approach at the training stage based on the addition of special normalizing layers (BacthNorm, batch normalization [17]).

Network architecture Alexnet [10] was used because it's investigated and has rather shallow depth (low count of hidden layers), alleviating process of classifier training and providing comparatively to other networks architectures low computing complication of feature extracting process.

It was used StanfordCars [18] dataset train part, which contains 8,144 images of static cars of 196 different models (totally dataset contains 16,185 images), for classifier training. In purposes of increasing of classification accuracy, ”overfitting” effect decreasing and raising of classifier universality the source dataset was augmented by 10 times with additional data balancing was done [19] (was equalized amount of images of each car models up to 450 samples). The augmentation of the original images was carried out by using affine transformations, perspective transformations, contrast changes, Gaussian noise, hue/saturation changes, cropping/padding, blurring. For each image set of used transformations was chosen randomly and parameters of each transformation was selected like values of uniform distribution with transformation-specific predefined range 1.

Fig. 1. Example of a augmentation results



Classification network receives on its input a relevant image to get an object's features vector and extracts a set of coefficients after the last MaxPool layer represented as a value matrix sized by $256 \times 6 \times 6$:

$$F_i = \frac{\sum_{k=1}^{6} \sum_{m=1}^{6} (C_{i,k,m} - min(C))}{max(C) - min(C)}, \tag{3}$$

where $F_i \in \mathbf{F}$ is en element of a result feature vector containing 256 elements, $C_{i,k,m} \in \mathbf{C}$ is an element of coefficient matrix extracted from a network.

3 Feature vector similarity criterion

Used feature vector similarity criterion receives on its input two object's feature vectors $\mathbf{F}$ and $\mathbf{F}'$ computes some similarity criterion $S \in [0,1]$:

$$S = \begin{cases} 1 - \frac{\sum_{i=1}^{256} J(F_i,F_i')}{\sum_{i=1}^{256} I(F_i,F_i')}, \text{ if } \sum_{i=1}^{256} I(F_i,F_i') > 0, \\ 0.5, \qquad\qquad\qquad \text{otherwise} \end{cases}, \tag{4}$$

where $F_i \in \mathbf{F}, F_i' \in \mathbf{F}', I(F_i,F_i')$ and $J(F_i,F_i')$ are indicators computed as:

$$I(F_i,F_i') = \begin{cases} 1, \text{ if } F_i > thr \, or \, F_i' > thr, \\ 0, \text{ otherwise} \end{cases}, \tag{5}$$

$$J(F_i,F_i') = \begin{cases} I(F_i,F_i'), \text{ if } |F_i - F_i'| > \frac{1}{2} max(F_i,F_i'). \\ 0, \qquad\qquad \text{otherwise} \end{cases}. \tag{6}$$

4 Assessment of the effectiveness of the proposed approach

A comparative assessment of the effectiveness of the proposed criteria for the similarity of features is based on the evaluation of the Euclidean distance between the compared vectors, an approach using the support vector method (SVM [14]), as well as an approach based on the use of siamese neural networks [6] and other modern analogues, both on a test subset of the data set StanfordCars used for training [18] and on examples from the data set CompCars, which contains images of moving vehicles was captured by surveillance cameras with large appearance variations due to the varying conditions of light, weather, traffic, etc [5].

As the test data of the set StanfordCars [18], we used 5000 "positive" examples consisting of images of cars of one model, and 5000 "negative" pairs of images consisting of images of cars of different models, randomly selected from a subset of images that were not used in the training process. As test data of the set CompCars we used the original structure of test data from three difficulty levels, containing at each difficulty level 20,000 compared pairs of images (10,000 ”positive” and ”negative” examples) [5]. Each image pair in the "easy set" is selected from the same viewpoint, while each pair in the "medium set" is selected from a pair of random viewpoints. Each negative pair in the "hard set" is chosen from the same car make.

As a quality criterion, a re-identification accuracy metric was used, calculated as:

$$Accuracy = 100 * \frac{TP + TN}{N}, \%, \tag{7}$$

where TP is an amount of correctly recognized ”positive” image pairs (the pair contains images of vehicles of the same model, and the verifier evaluates them as elements of one subset), TN is an amount of correctly recognized "negative" pairs, N is an amount of compared images.

Evaluation of the effectiveness of the proposed approach and the approach on the test subset of the training data set is given in Table 1, while for evaluation of the effectiveness of the proposed criterion and metric based on the evaluation of the Euclidean distance, it was used a threshold selected to ensure maximum efficiency (accuracy) of verification on the training data.

Table 1. Estimation of the effectiveness of the proposed criteria for the similarity of features (StanfordCars).

Approach	Accuracy, (%)
Proposed feature extraction method + Euclidean distance	65,5
Proposed feature extraction method + SVM [4]	71,2
Proposed feature extraction method + proposed feature criteria for the similarity of features	69,8
Siamese network (Triplet Loss)	-
Random selection	50

Comparison presented in the Table 1 allows us to conclude that during processing data as close as possible to training, the proposed criterion for comparing feature vectors is somewhat inferior to the comparison method updated on machine learning techniques (in this case, the method of support methods was used) and surpasses the approach based on estimating the Euclidean distance between vectors. It should also be noted that the lack of results for the siamese network [6] is due to the fact that, despite the enumeration of training hyperparameters, it was not possible to ensure the convergence of the network on the StanfordCars data set, which is a known problem in the process of training networks with Siamese architecture on heterogeneous data [1].

A comparative analysis of the effectiveness of the proposed approach on the CompCars [5] data set is given in Table. 2.

Basing on the analysis of the table. 2, we can draw the following conclusions:
- the proposed metric of similarity of feature vectors allows you to maintain the relative accuracy of verification when switching to a data set significantly different from the training, while the approach for comparing features using the support vector method significantly loses accuracy due to the "overfitting" effect;

- the proposed vehicle model verification system demonstrates on a set of data significantly different from the training accuracy comparable to the accuracy of similar basic algorithms (GoogleNet + SVM [5]), trained on a subset of the data close to the test ones, however, it loses much to modern verification approaches (Mixed Diff + CCL [5]), the training of which was also carried out on a subset of data similar in their characteristics to the test ones.

Table 2. Evaluation of the effectiveness of the proposed criteria for the similarity of features (CompCars).

Approach	Training Dataset	Accuracy, %		
		Easy	Medium	Hard
Proposed feature extraction method + Euclidean distance	StanfordCars[18]	62,1	60,3	57,6
Proposed feature extraction method + SVM [4]	StanfordCars[18]	67,1	63,7	61,5
Proposed feature extraction method + the proposed criteria for the similarity of features	StanfordCars[18]	72,4	70,1	66,8
GoogleNet + SVM [5]	CompCars[5]	70	69	65,9
Mixed Diff+CCL [1]	CompCars[5]	83,3	78,8	70,3
Random select	-	50	50	50

In addition, it should be noted that the proposed approach for verifying a car model operates with shorter feature vectors compared to GoogleNet + SVM [5] (256 for the proposed approach, 4096 for GoogleNet + SVM [5]) and does not require a long training process for a module of similarity defining. And also, unlike Mixed Diff + CCL [1], which belongs to the class of Siamese networks, it allows to save the feature vector separately for further comparison without computationally complex features extraction operations.

It is important for the class of cyberphysical systems under consideration to ensure the probability of completing tasks in a given time if the proposed approach is used in real-time, which, as shown in [19–21], is achievable with redundant calculations.

5 Conclusion

The proposed approach for selecting and comparing features of objects from their image is used in the task of verifying a vehicle model. Selecting is being occured by modifying of a known artificial neural network trained on a specially prepared (augmented) data set. The proposed criterion for the similarity of feature vectors is based on comparison techniques and has extremely low computational complexity. The proposed vehicle verification method demonstrates accuracy comparable to similar modern methods in those use cases when the processed data have the same characteristics as the training ones (a similar camera model, similar level of lighting and noise, etc.), and demonstrates higher relative accuracy in processing data that are significantly distinguished from training.

References

1. Liu H. et al. Deep relative distance learning: Tell the difference between similar vehicles // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. P. 2167–2175.
2. Rublee E. et al. ORB: An efficient alternative to SIFT or SURF // ICCV. 2011. V. 11. N 1. P. 2.
3. Pan X., Lyu S. Region duplication detection using image feature matching // IEEE Transactions on Information Forensics and Security. 2010. V. 5. N 4. P. 857–867.
4. Zapletal D., Herout A. Vehicle re-identification for automatic video traffic surveillance // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops. 2016. P. 25–31.
5. Yang L. et al. A large-scale car dataset for fine-grained categorization and verification // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2015. P. 3973–3981.
6. Koch G., Zemel R., Salakhutdinov R. Siamese neural networks for one-shot image recognition // ICML deep learning workshop. 2015. V. 2.
7. Cheng D. et al. Person re-identification by multi-channel parts-based cnn with improved triplet loss function // Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition. 2016. P. 1335–1344.
8. Ke Y. et al. PCA-SIFT: A more distinctive representation for local image descriptors // CVPR (2). 2004. V. 4. P. 506–513.
9. Ng P.C., Henikoff S. SIFT: Predicting amino acid changes that affect protein function // Nucleic acids research. 2003. V. 31. N 13. P. 3812–3814.
10. Krizhevsky A., Sutskever I., Hinton G.E. Imagenet classification with deep convolutional neural networks // Advances in neural information processing systems. 2012. P. 1097–1105.
11. Szegedy C. et al. Going deeper with convolutions // Proceedings of the IEEE conference on computer vision and pattern recognition. 2015. P. 1–9.
12. Tanner M.A. Tools for statistical inference: observed data and data augmentation methods. Springer Science and Business Media, 2012. V. 67.
13. John G.H., Kohavi R., Pfleger K. Irrelevant features and the subset selection problem // Machine Learning Proceedings 1994. Morgan Kaufmann, 1994. P. 121–129.
14. Joachims T. Making large-scale SVM learning practical. Technical report, SFB 475: Komplexitätsreduktion in Multivariaten Datenstrukturen, Universität Dortmund, 1998. N 1998, 28.
15. Hoffer E., Ailon N. Deep metric learning using triplet network //International Workshop on Similarity-Based Pattern Recognition. Springer, Cham, 2015. P. 84–92.
16. Xu B. et al. Empirical evaluation of rectified activations in convolutional network // arXiv preprint arXiv:1505.00853. 2015.
17. Howard A.G. et al. Mobilenets: Efficient convolutional neural networks for mobile vision applications // arXiv preprint arXiv:1704.04861. 2017.
18. Krause J. et al. 3d object representations for fine-grained categorization // Proceedings of the IEEE International Conference on Computer Vision Workshops. 2013. P. 554–561.
19. A. V. Bogatyrev, V. A. Bogatyrev and S. V. Bogatyrev, "Multipath Redundant Transmission with Packet Segmentation," 2019 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF), Saint-Petersburg, Russia, 2019, pp. 1-4. doi: 10.1109/WECONF.2019.8840643

20. V. A. Bogatyrev, S. V. Bogatyrev and A. V. Bogatyrev, "Model and Interaction Efficiency of Comput-er Nodes Based on Transfer Reservation at Multipath Routing," 2019 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF), Saint-Petersburg, Russia, 2019, pp. 1-4. doi: 10.1109/WECONF.2019.8840647

21. Bogatyrev A.V., Bogatyrev S.V., Bogatyrev V.A. Analysis of the Timeliness of Redundant Service in the System of the Parallel-Series Connection of Nodes with Unlimited Queues//2018 Wave Electronics and its Application in Information and Telecommunication Systems (WECONF), 2018, pp. 8604379