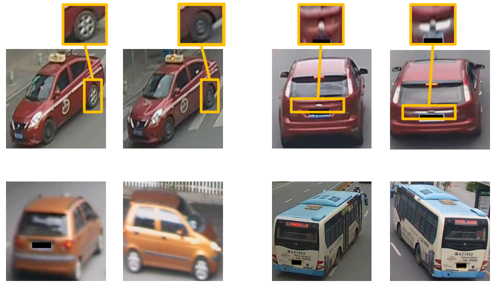
Orientation Invariant Feature Embedding and Spatial Temporal Regularization for Vehicle Re-identification

Zhongdao Wang, Luming Tang, Xihui Liu, Zhuliang Yao, Shuai Yi, Jing Shao,
Junjie Yan, Shengjin Wang, Hongsheng Li, Xiaogang Wang
Tsinghua University SenseTime Group Limited The Chinese University of Hong Kong

International Conference on Computer Vision 2017

Vehicle Re-Identification

- Vehicle re-identification (ReID) is of great importance in urban surveillance, aiming at associating vehicle images across cameras and temporal periods.
- Sometimes vehicles' license plate is occluded or cannot be seen clearly, vehicle ReID methods can be used in these scenarios to effectively locate vehicles of interest from surveillance databases.
- Difficulties and challenges:
  - Vehicle ReID needs fine-grained feature. Some vehicles are quite similar and local patterns are keys to determine whether they are the same ones.
  - Vehicles have different orientation. Comparing global features directly between vehicles with different orientations is not optimal.

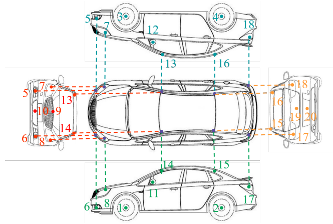


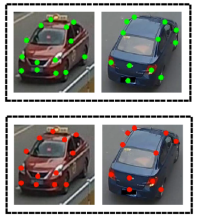
Fine-grained differences

Different orientations

Our Contributions

- We defined 20 vehicle key points and adopted an hourglass-like fully convolution network to generate vehicle key point response maps.
- A deep learning framework is proposed to extract ***fine-grained*** and ***orientation-invariant*** vehicle appearance feature, which contains four main components:
  - Orientation-based region proposal module
  - Feature extraction module
  - Feature aggregation module
  - Spatio-temporal regularization module
- State-of-the-art performance is achieved on existing vehicle retrieval datasets.

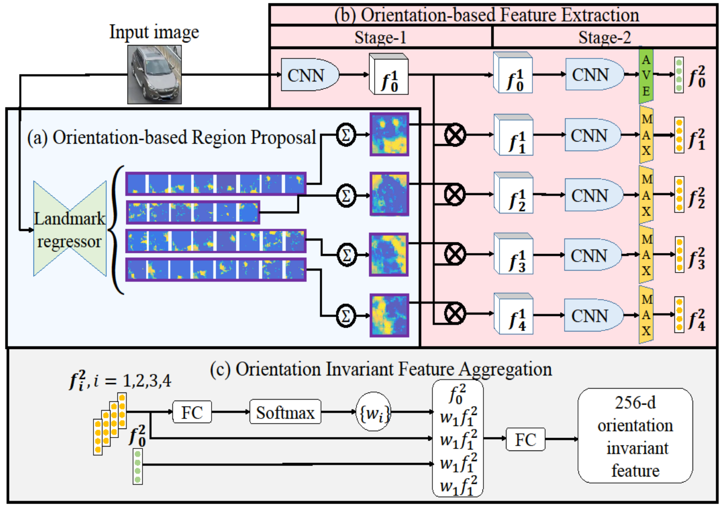




Ground truth Key points

Predicted Key points

Orientation Invariant Feature Embedding



(1) Orientation-based Region Proposal

- Response maps of 20 vehicle key points are generated by an hourglass-like network (landmark regressor). High value in one map means the corresponding key point may lie in this location with high probability.
- Response maps are assigned to four clusters according to their orientation(front, back, left and right), e.g., response maps of front auto logo, front license plate, headlights and fog lamps are assigned to a 'front-face' cluster.
- Response maps in each cluster are element-wisely summed up to give an orientation-based region mask, representing the salience region in one face.
- Four region masks are resize to the same size as the feature map of the vehicle image after convolution module f_0^1 , then region masks are element-wisely multiplied to the feature map to obtain the orientation-related feature maps.

(2) Orientation-based Feature Extraction

- Four orientation feature maps and the original feature map are further convolved by two more convolution modules in different branches. Each branch outputs a 1536-d feature.

(3) Orientation Invariant Feature Aggregation

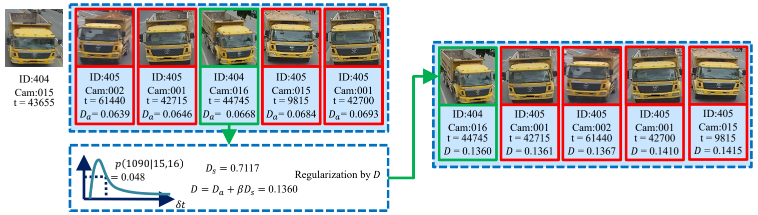
- Five feature vectors are aggregated in an attention fashion, finally a 256-dimension feature vector is outputted, which is a ***fine-grained*** and ***orientation-invariant*** embedding for vehicle images.

Spatial Temporal Regularization

- The location and time stamp of a vehicle in surveillance are easy to obtain, so it is possible to refine vehicle search results with the help of such spatio-temporal information
- We model the vehicles' transition interval between pairs of cameras as a random variable that follows ***logarithmic normal distribution*** with parameter μ and σ**.** Then parameters of every camera pair are estimated by maximizing log-likelihood function.
- During the retrieval process, the appearance distance D_a is first computed via the proposed orientation-invariant feature aggregation framework, while the spatio-temporal distance D_s of two vehicle is computed for regularization.:

$$D_s = 1/(1 + e^{\alpha(p(\tau|l,e;\mu_{l,e},\sigma_{l,e}) - 0.5)}).$$

  where p is the probability destiny function estimated in the training phase.
- The final distance of two vehicle is computed as weighted summation of D_a and D_s.



Experimental Results

- Our proposed orientation invariant feature embedding outperforms previous works by a large margin.
- Spatial temporal regularization is proved to be efficiency to further improve the ReID performance.
- By experiments we show that metric learning methods like KISSME and MLAPG can be utilized to further refine the ReID result.

VeRi-776	mAP	HIT@1	CMC@1	CMC@5
BOW-CN [28]	12.20	33.9	-	-
LOMO [6]	9.64	25.3	-	-
KEPLER [11]	33.53	68.7	48.2	64.3
PROVID [9]	27.77	61.4	-	-
Baseline	45.50	88.66	62.8	86.7
Ours	48.00	89.43	65.9	87.7
Ours + ST	**51.42**	**92.35**	**68.3**	**89.7**

VehicleID	CMC@1	CMC@5
KEPLER [11]	45.4	68.9
VGG + Triplet Loss [8]	31.9	50.3
VGG + CCL [8]	32.9	53.3
Mixed Diff + CCL [8]	38.2	61.6
Baseline	63.2	80.6
Ours	**67.0**	**82.9**