CONTENT-BASED VEHICLE RETRIEVAL USING 3D MODEL AND PART INFORMATION

Ming-Kuang Tsai, Yen-Liang Lin, Winston Hsu

National Taiwan University,
Taipei, Taiwan

Chih-Wei Chen

Industrial Technology Research Institute,
Hsin-Chu, Taiwan

ABSTRACT

Content-based vehicle retrieval in unconstrained environment plays an important role in surveillance system. However, due to large variations in viewing angle/position, illumination, and background, traditional vehicle retrieval is extremely challenging. We approach this problem in a different way by rectifying vehicles from disparate views into the same reference view and searching the vehicles based on informative parts such as grille, lamp, and wheel. To extract those parts, we fit 3D vehicle models to a 2D image using active shape model (ASM). In the experiments, we compare different 3D model fitting approaches and verify that the impact of part rectification on the content-based vehicle retrieval performance is significant. We propose a model fitting approach with weighted jacobian system which leverages the prior knowledge of part information and shows better results. We compute mean average precision of vehicle retrieval with L1 distance on *NetCarShow300* dataset, a new challenging dataset we construct. We conclude that it benefits more from the fusion of informative rectified parts (e.g., grille, lamp, wheel) than a whole vehicle image described by SIFT feature for content-based vehicle retrieval.

***Index Terms*—** 3D model construction, 3D model fitting, content-based vehicle retrieval, part rectification

1. INTRODUCTION

Vehicles are one of the most important subjects in surveillance environment when surveillance cameras become ubiquitous and more and more surveillance video data is available. However, millions of surveillance videos are so large-scaled that it is impossible for human to deal with. Therefore, effective vehicle retrieval is becoming increasingly significant. As a result, we propose an effective content-based vehicle retrieval approach to satisfy the needs. See Figure 1.

To address diversity of viewpoints and shape variation, it comes to our minds that using rectified parts extracted from fitted 3D vehicle models must be more powerful than a whole image for content-based vehicle retrieval. In other domains, such as face [1] and people [2] search, salient attributes or parts have been utilized to identify targets. However, the utilization of parts for vehicles has not achieved similar successes. In addition, unlike people or face recognition, using 2D models to extract parts of a vehicle within the bounding box generally fails due to dramatic variations in viewing angles. Employing 3D models is more suitable for our work. In fact, using 3D vehicle model is one of the major line of research in the fields of vehicle detection [3], pose estimation [4][5], classification [6][7], etc. To deal with the lack of details in simple polyhedral models for vehicle model fitting, M. J. Leotta et al. [8] and Y. Tsin et al. [9] use more delicate 3D vehicle models which provide rich constraints to match vehicles reliably and refine 3D-to-2D alignment until convergence.

Searching for vehicles in surveillance videos, Feris et al. [10] build a surveillance system capable of vehicle retrieval based on semantic attributes. To deal with different viewpoints, they train 12 motionlet detectors from a set of city surveillance cameras. They define several attributes as possible descriptions, such as dominant color, direction, and vehicle dimensions. To estimate vehicle dimensions in world coordinates, they manually do camera calibration and use a simple 3D model fitting approach on the basis of several assumptions (i.e., a vehicle's location on the ground plane, orientation of heading direction, and the scale of the model).

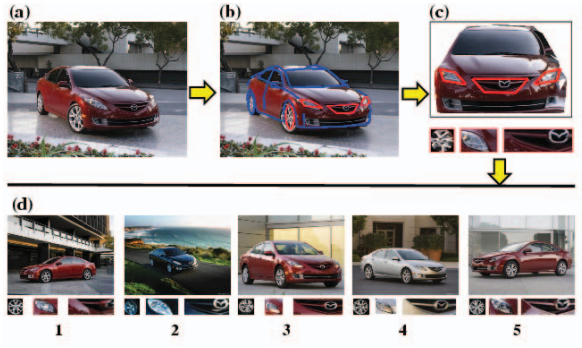

Fig. 1. An overview of the proposed system. (a) Input image. (b) Aligning 3D model to 2D image. (c) Rectifying vehicles to the same reference view points and extracting vehicle parts (e.g., grille, wheel, and lamp). (d) Top 5 searching results by fusing three parts. Best viewed in color.

Taking advantage of previous works, we propose to augment content-based vehicle retrieval by aligned 3D vehicle model and fusing informative parts (cf. Figure 1). First, we establish consistent shape representation between several 3D vehicle models (Section 2) and align 3D model to natural images (Section 3) (cf. Figure 1(b)). Second, informative parts (e.g., grille, lamp, and wheel) are extracted and rectified into one reference view (Section 4), and the parts are represented by several features for retrieval (Section 5) (cf. Figure 1(c)). Third, we evaluate our approach under different situations. The result shows that we improve the retrieval performance significantly even in diversity of viewpoints.

The main contributions of this work include:

- We implement and compare current state-of-the-art 3D model fitting algorithms and evaluate on a challenging dataset.
- We argue to improve 3D model fitting precision by leveraging the prior knowledge of those informative parts.
- We investigate the impacts of rectified parts on the content-based retrieval performance.
- To our best knowledge, this is the first content-based vehicle retrieval approach that uses informative parts and analyzes the detailed parameterizing components for the framework.

2. 3D VEHICLE MODEL CONSTRUCTION

Considering shape variation of vehicles, we build an active shape model (ASM) for vehicles. We manually select 128 3D points for a half vehicle model to make sure the correspondence of the same physical shape. The other half can be obtained by mirroring. Procrustes analysis is done to align shapes of each instance. Then, we apply principal component analysis (PCA). The M eigenvectors corresponding to the M largest eigenvalues of the covariance matrix define the vehicle space. By projecting a 3D vehicle model to the vehicle space, we can get a vehicle shape with a distribution in M-dimensional weight space. The projected weights control the variability of the shape of a vehicle model.

In our experiment, we use 11 3D vehicle models as training instances, including 3 sedans, 2 wagons, 1 pickup truck, 1 crossover, 2 hatchbacks, and 2 SUVs, with totally 256 salient points for two sides and 342 triangular faces to describe wheels, radiator grille, lamps, and other semantic parts. According to the mean reconstruction error estimated by the ratio between average distance error and the vehicle length, we find that less than 0.4% reconstruction error results from 8 eigenvectors; that is, the error is only about 4 pixels if the length of a projected vehicle is 1000 pixels, which is relatively low.

3. 3D VEHICLE MODEL FITTING APPROACH

In order to extract parts of vehicles, 3D model fitting is essential. In the model fitting step, we assume that initial position and pose of a vehicle in an image can be estimated by multi-view object detection approaches (e.g., [3][4]) and the direction and detected objects for the moving objects. Content-based vehicle retrieval is based on information about the target vehicle.

We investigate and compare two different state-of-the-art approaches in [8] and [9]. One depends on point registration and the other solves a Jacobian system. This is the first work comparing these two approaches. Moreover, we propose to leverage the prior knowledge of semantic parts (e.g., grille, lamp, and wheel) and further improve the challenging 3D alignment problem.

In general, the approaches start with some initial parameters. Then, a set of hypothetic edges is generated, and a collection of correspondences between the observed and projected edges is determined by local search. After iterative updates for the correspondences, the shape and pose will converge.

3.1. Model Fitting Methods

3.1.1. Fitting by Point Registration

Here we apply point registration (PR) algorithms to find corresponding points, solve equations, and obtain projected weights and translations as [9]. First of all, given an initial pose, each landmark point is reconstructed from the mean shape and projected according to the general camera equation. Second, for each projected salient edge point, we find all nearby points in the normal direction as a candidate point set. Third, we apply a point registration approach, Kernel Correlation (KC) [11] or Coherent Point Drift (CPD) [12]. The step finds a rigid or non-rigid transformation which maximizes the correlated distribution between two point sets. Finally, when we assume only shape and translation parameters are unknown, the model fitting problem is formulated as a least square problem based on the correspondence and other known factors, and it can be solved by repeating the steps until it converges.

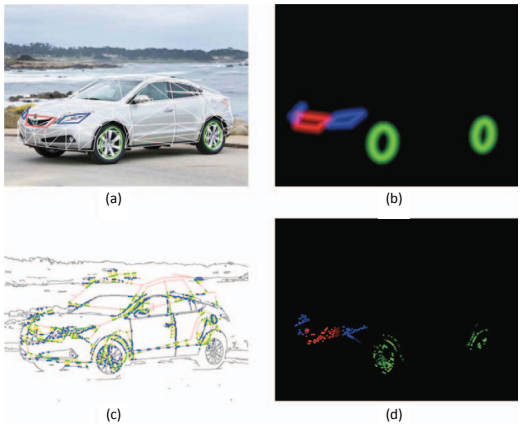

Fig. 2. Illustration of 3D model fitting process. (a) The input image with superjacent 3D ground truth data. (b) The synthetic weight map of grille, lamp and wheel drawn in different colors. For each part, the color strength represents the weights. (c) Intermediate result in 3D model fitting process. Red line segments are hypothetic edges of current vehicle pose. The points on the hypothetic edges are green and the corresponding points on the observed edges are blue. Each correspondence is linked by a yellow line which represents the error measurement. (d) Intermediate weight value of each observed points. Higher weight values imply higher probability of the observed point belonging to the correct part. Best viewed in color.

3.1.2. Model Fitting by Jacobian System

As mentioned earlier, we assume there are initial parameters. Given a collection of correspondences between observed and projected edges, each corresponding edge point produces one error measurement e_i. The fitting problem can be formulated as a Jacobian system (JS) in [8]: $J\Delta p = e$, where e is the vector of signed errors, Δp is the vector of parameter displacement updated at each iteration, and J is the Jacobian matrix with current parameters. The solution is found by a least square method and iteratively optimizing the parameters until convergence.

3.2. Model Fitting with Part Information

Most of the model-fitting algorithms find corresponding points by local search of the projected edges depending only on some low-level features, such as edge intensity and edge orientation, which are still likely to fail and converge to local maxima in common cases due to cluttered background or complexities of edges on the surface of vehicles. We are interested in whether it is possible to improve the fitting algorithm with some prior knowledge of parts. That is, we can give different weights to different correspondences and lead to better fitting results. To validate our assumption for sure, we generate synthetic weight maps of parts by using annotated ground truth data (Figure 2(b)), and formulate this problem into a weighted Jacobian system (WJS): $WJ\Delta p = We$, where W is a diagonal weight matrix with each diagonal element w_{ii} representing the weight of each correspondence. We take two important weights into consideration, distance weight w_{dist} and part weight w_{part}. w_{ii} is computed by a linear combination with λ: $w_{ii} = \lambda \cdot w_{dist} + (1 - \lambda) \cdot w_{part}$, where w_{dist} is based on the Beaton-Tukey weight [13] and w_{part} is determined by the value of the location of observed edge point in the weight map (Figure 2(d)). Our experiments show that the 3D model fitting precision is improved with the aid of the prior weight map.

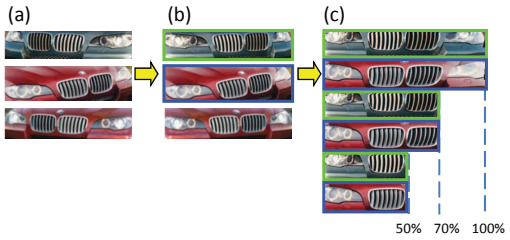

Fig. 3. Illustration of part rectification. (a) The original extracted frontal regions after 3D fitting. (b) The frontal parts flipped to the same side. (c) The flipped parts rectified to specific pose and retaining 70% and 50% width respectively.

4. PART RECTIFICATION

Depending on the estimated pose of each vehicle (Figure 1(b)), we extract the parts after the state-of-the-art 3D model fitting approach. Before feature extraction and feature comparison, we then rectify the parts by projecting them into specific angles, such as the front view or side view.

Figure 3 depicts the process of part rectification. First, parts may be contrary in different vehicle images due to different viewpoints. Therefore, utilizing symmetry of vehicle shape, we flip the visible parts into the same side before applying feature extraction. Second, we adopt barycentric coordinates for image warping after comparing other methods (e.g., global affine transformation). By bilinear interpolation and inverted mapping, each point in the projected view can find the corresponding point in the original image and get the mapped pixel value. Furthermore, we can remove background pixels which are out of projected triangles. Third, the visible parts are under different poses, so there are some distorted regions after rectification. In other words, warping may enlarge originally small regions and cause distortion. As a result, removing the regions may seem redundant but actually improve the performance. Empirically, we remain 70% and 50% ratio of width to investigate the influence.

5. EXPERIMENTS

In the following, we conduct several experiments on a challenging dataset to investigate the performance of our content-based retrieval approach and the comparison of 3D model fitting. To prove our idea and remove uncertain factors in content-based vehicle retrieval for leveraging informative parts, our retrieval experiments are based on extracted parts from the ground truth. But we also show the retrieval results based on those parts extracted by model fitting in Section 5.3. Similarly, in order to evaluate the influence of knowledge of part information for 3D vehicle model fitting, weight maps are generated from the ground truth.

5.1. NetCarShow300 Dataset

We collect 300 images from *NetCarShow.com*[1] , the *NetCarShow300* dataset, where the size is comparable to commonly used vehicle recognition image datasets. There are 30 vehicle instances. Each instance has 10 images respectively. Each image contains one main vehicle of which the frontal part is visible. The vehicles are presented in different environments, including noisy background, different illumination, and shadows. Moreover, a vehicle may be extremely projective, and the surface has reflection. No doubt the diversity challenges the model fitting and retrieval. Also, vehicle instances made by the same manufacturer, e.g., *Honda Odyssey*, *Honda Pilot*, may influence the performance if we focus on retrieving those which belong to the same instance as the input image. The ground truth of *NetCarShow300* is obtained by aligning the projected models manually. That is, we set several hard-constrained corresponding points between a projected model and a vehicle image. Given the hard-constrained correspondence, we can update the shape distribution iteratively by 3D model fitting. Finally, we get approximate 3D vehicle models and a 2D projected vehicles as the ground truth.

[1] http://www.netcarshow.com

5.2. Vehicle Retrieval Performance

In this experiment, we apply several descriptors on extracted parts which are resized to the same number of pixels while keeping the ratio between height and width. Those retrieved vehicle instances which have the same label as the query instance are correct. We compare the mean average precision (MAP) performance on different sources including a whole vehicle image and three parts, grilles, lamps, and the most visible wheel. Then, we do sensitivity tests to select the late fusion weights and obtain the best parameters.

The following are the descriptors we tested. Firstly, Difference of Gaussian (DoG) detector and SIFT descriptor are used. Each feature point are transformed into a visual word according to a codebook containing 512 entries. Secondly, we use Pyramid Histogram of Oriented Gradients (PHOG) which computes the histogram of gradient in a region with several levels. We concatenate the vectors into a 168-dimension descriptor. Thirdly, we adopt the rotation-invariant feature, Local Binary Pattern Histogram Fourier (LBPHF) [14]. It applies to a whole region and describes the appearance locally based on the signs of differences of neighboring pixels. Three circular neighborhoods are used and result in 478-dimension descriptors.

We collect the leave-one-out results with these descriptors combined with L1 or cosine distance. Table 1 shows that PHOG descriptor with L1 measure (PHOG+L1) outperforms other descriptors and achieves an MAP of 54.84% with 50% frontal parts composed of half grille and a lamp. The reason may be that PHOG maintains the structural consistency, which is benefited a lot by the rectified parts. Using 70% or 50% part regions which are more undistorted increases MAP by around 2–6%. Table 1 also indicates that flipping alignment from "Original Side" to "Same Side" improves the performance by about 5–10%. However, it shows that SIFT descriptor becomes worse after part rectification. The reason may be that DoG detector fails to find good feature points and matched visual words become fewer after the rectification step.

Considering each parts, the grille and lamp are more discriminative than the wheel. The grille part has an MAP of 47% when retaining 70% rectified region, and the lamp part also achieves 47% with PHOG. It is obvious that the composition of grilles and lamps is distinct between vehicle instances, but wheels are not very helpful when distinguishing the vehicle instances. One explanation is that the wheel structure may be not consistent in one vehicle instance. The other reason is that the internal structure of wheel parts may be blurred and unidentifiable when the vehicle is in motion.

Furthermore, the last row in Table 1 shows the result when we combine the three parts, grille, lamp, and wheel, to do late fusion:

$$S_{fusion} = w_{grille} \cdot S_{grille} + w_{lamp} \cdot S_{lamp} + max(1 - w_{grille} - w_{lamp}, 0) \cdot S_{wheel}, \quad (1)$$

where S means the similarity score, and w_{grille} and w_{lamp} are the weights of the grille and lamp respectively. After a sensitivity test, the achieved MAP is 63.08% on $w_{grille} = 0.4$ and $w_{lamp} = 0.5$, and it significantly outperforms the previous unfused results.

Table 1. The performance (in MAP) for vehicle. "Fusion of 70% Grille, Lamp, and Wheel" obtains the best MAP of 63.08% and is much better than our baseline (32.01% with SIFT+L1). "Rectified Body Same Side" refers to the top image in Figure 1(c). "Original Front Original Side" and "Original Front Same Side" refer to the Figure 3(a) and (b) respectively. "Rectified Same Side" with "Front", "70% Front", and "50% Front" refer to Figure 3(c). The meanings of the rest rows are similar.

Descriptor+Distance Measure	SIFT+L1	SIFT+COS	LBPHF+L1	LBPHF+COS	PHOG+L1	PHOG+COS	WJS+PHOG+L1
Rectified Body Same Side	32.01%	29.30%	21.10%	17.47%	31.61%	23.76%	25.26%
Original Front Original Side	36.96%	32.98%	20.17%	17.54%	22.87%	20.02%	18.14%
Original Front Same Side	39.42%	34.63%	20.08%	17.40%	35.63%	29.87%	29.22%
Rectified Front Same Side	37.95%	34.00%	25.27%	21.23%	48.99%	38.87%	37.80%
Rectified 70% Front Same Side	38.95%	32.99%	26.89%	22.77%	51.76%	41.16%	41.88%
Rectified 50% Front Same Side	29.27%	26.48%	25.58%	22.66%	**54.84%**	45.05%	44.58%
Original 50% Front Same Side	30.98%	26.88%	20.83%	18.39%	44.96%	35.19%	36.93%
Rectified Grille Same Side	31.85%	27.17%	35.57%	31.64%	45.13%	34.53%	34.40%
Rectified 70% Grille Same Side	30.72%	27.55%	34.30%	30.99%	47.38%	36.87%	31.66%
Rectified Lamp Same Side	13.17%	12.24%	25.77%	23.78%	47.31%	42.21%	28.93%
Rectified Wheel Same Side	13.78%	11.87%	10.80%	9.62%	14.00%	12.13%	9.86%
Fusion of 70% Grille, Lamp, Wheel	34.26%	30.23%	43.24%	38.54	**63.08%**	53.79%	42.89%

Table 2. 3D Model fitting precision. Weighted Jacobian System (WJS) outperforms other approaches.

Method	APD	STD
Initial Location	47.15	6.06
PR(KC)	39.26	9.90
PR (Rigid CPD)	29.59	6.91
PR (Non-rigid CPD)	26.53	6.84
JS	34.19	6.31
WJS($\lambda = 0.3$)	**18.73**	**4.66**

5.3. Model Fitting Comparison

To compare difference between model fitting approaches, we generate a testing data with noisy initial position. Then, we measure average pixel distance (APD) and standard deviation (STD) of visible vertices between fitted models and ground truth.

In Table 2, PR(Rigid-CPD) (29.59) is better than other approaches (39.2 and 34.19). Besides, non-rigid transformation improves the performance (26.53) because deformation possibility is considered even the model does not actually change the shape. In fact, we find that translating to good location is an important key for good fitting performance because worse translation may increase overall distance. Rotation and shape deformation then adjust the position of each vertex locally and lead to the minor improvement. Furthermore, with the knowledge of salient parts, we can utilize these weighing maps to facilitate the 3D model fitting precision. In the sensitivity test, WJS with λ = 0.3 has the lowest error 18.73 which surpasses other 3D model fitting approaches.

The retrieval performance corresponding to the fitting result (WJS+PHOG+L1) is shown in Table 1. It has lower MAP than the ideal case, but it still achieves relatively better performance than ideal rectified whole image and validates the impact of part rectification.

6. CONCLUSIONS AND FUTURE WORK

In this paper, we effectively utilize 3D vehicle models for novel content-based vehicle retrieval. When robust 3D model fitting approaches are applied, it is possible to extract some discriminative parts. After part rectification, we demonstrate remarkable performance on a challenging dataset. The precision is surely notable and supports our idea on vehicle part information fusion. Besides, our investigation shows that the prior knowledge regarding certain parts has noteworthy impacts on 3D model fitting. While our current application is based on given initial pose and location, we are undergoing an approach to automatically generate the information. In the future, we expect to include vehicle detection and pose estimation steps, and we can build a structural augmenting content-based vehicle retrieval system on more difficult natural images and leverage more informative parts to increase the performance.

7. REFERENCES

[1] N. Kumar et al., "Facetracer: A search engine for large collections of images with faces," in *ECCV*, 2008.

[2] D. A. Vaquero, R. S. Feris, D. Tran, L. Brown, A. Hampapur, and M. Turk, "Attribute-based people search in surveillance environments," in *WACV*, 2009.

[3] M. Stark, M. Goesele, and B. Schiele, "Back to the future: Learning shape models from 3d cad data," in *BMVC*, 2010.

[4] M. Arie-Nachmison and R. Basri, "Constructing implicit 3d shape models for pose estimation," in *ICCV*, 2009.

[5] J. Liebelt et al., "Viewpoint-independent object class detection using 3d feature maps," in *CVPR*, 2008.

[6] Y. Guo et al., "Peet: Prototype embedding and embedding transition for matching vehicles over disparate viewpoints," in *CVPR*, 2007.

[7] S. M. Khan et al., "3d model based vehicle classification in aerial imagery," in *CVPR*, 2010.

[8] Matthew J. Leotta and Joseph L. Mundy, "Predicting high resolution image edges with a generic, adaptive, 3-d vehicle model.," in *CVPR*, 2009.

[9] Y. Tsin, Y. Genc, and V. Ramesh, "Explicit 3d modeling for vehicle monitoring in non-overlapping cameras," in *AVSS*, 2009.

[10] Rogerio Feris et al., "Attribute-based vehicle search in crowded surveillance videos," in *ICMR*, 2011.

[11] Y. Tsin and T. Kanade, "A correlation-based approach to robust point set registration," in *ECCV*, 2004.

[12] Andriy Myronenko and Xubo B. Song, "Point set registration coherent point drift," 2009.

[13] A. E. Beaton and J. W. Tukey, "The fitting of power series, meaning polynomials, illustrated on band-spectroscopic data," in *Technometrics*, 1974.

[14] T. Ahon et al., "Rotation invariant image description with local binary pattern histogram fourier features," in *SCIA*, 2009.