2011 11th IEEE International Conference on Computer and Information Technology

Vehicle Detection on Aerial Images by Extracting Corner Features for Rotational Invariant Shape Matching

Sheng Wang{Sheng.Wang-1@student.uts.edu.au}
School of Computing and Communications
University of Technology, Sydney
Sydney, NSW 2007

***Abstract*—Vehicle detection from aerial images has been extensively studied in many research papers and it is an important component of an intelligent transportation system. In the meantime, it is still a difficult problem with many open questions due to challenges caused by various factors such as low resolution of the aerial images, features restricted to a particular type of car, noise from other objects or object shadows, and occulsion in urban environments. By investigating several benchmark methods and frameworks in the literature, this paper proposes a novel feature fusion framework which successfully implements an effective vehicle detection method based on shadow detection followed by a rotational invariant shape matching of corner features. Promising results are obtained from the experiments.**

I. Introduction

A lot of problems has been caused by an increasing number of vehicles on the road, such as an increasing number of traffic accidents, daily traffic jams, difficulty for parking, air polution, increasing energy demands, and so on. Such problems are prevalent in all countries around the world, especially in capital cities and metropolis. Nevertheless, those problems can be alleviated by better planning of road construction and developments of Intelligent Transportation Systems (ITSs) for traffic route optimization. As a result, there is an increaing demand for intelligent transportation systems in urban environments. In an ITS, the most crucial part is vehicle detection, this paper will focus on vehicle detection in aerial images.

Vehicle detection in aerial images has attracted a lot of research attentions, not only becasue it has introduced an non-intrusive vehicle detection method compare to other methods such as an induction loop or a Radio Frequency Identification (RFID) approach, but also because it can provide a global panorama of the traffic situation in a specified district. Such global information, in addition to security surveillance, can also serve other meaningful purposes such as traffic census, urban planning and so on.

Although vehicle detection in aerial images has been widely studied, there are still many challenging problems, such as low resolution of the aerial images, in which vehicles become tiny objects; features that is only obvious on one particular type of vehicle object (e.g. medium sedan); potential noise introduced by other similar objects or shadows of those objects; occlusions from either objects such as trees or shadows of buildings; varying lighting conditions; varying background and varying vehicle density [1] [2] [3] [4] [5].

In order to overcome those challenges, many approaches try to integrate context information from other sources into image processing, such as 3D depth information and road network information obtained from GIS data [1] [4].

In [1], the authors have defined a total of 14 features, those features can be divided into two categories: spatial features and gray level features. Among the 8 spatial features, 6 were selected in combination with all 6 gray level features for classification. The image is first pre-classified with a rule-based classifier, which effectively eliminates many false positives (e.g. heterogeneous patches), then a statistical classifier based on Quadratic Discriminant Analysis (QDA) is used to perform the final classification.

In [2], an image is first segmented with the mean shift clustering algorithm, then a Gabor filter is used to measure the symmetry of each segment (blob). Only symmetric blobs will be considered for further feature extraction based on Shape Context feature [6]. A simple thresholding method based on the Euclidean distance of Shape Context feature descriptors will be used for classification.

In [4], 3D depth information is used together with 2D aerial image for object detection. Ground plane estimation is performed based on 3D depth information, such estimation can effectively eliminate false positives that is not on the ground. Then a variant of the edgelet feature is defined for feature extraction [7]. Finally, the cascade Adaboost algorithm is used to perform training and classification [8]. Differs from those conventional approaches which consider object shadows as noisy information and aims to eliminate the shadows, our approach makes use of object shadows to estimate the location of a potential object. Moreover, we use a new definiton of sample points to enhance the saliency of Shape Context feature.

In this paper, we propose a vehicle detection method based on static aerial images. Those images are manually collected from Google Map. The study site is Sydney, an urban environment where occlusion occurs frequently.

Our contributions can be summarised as follows

Firstly, we use shadow detection algorithm to quickly locate potential objects, this will greatly reduce the computational

978-0-7695-4388-8/11 $26.00 © 2011 IEEE
DOI 10.1109/CIT.2011.56

cost compared to the scanning window approach. Moreover, comparing with other image segmentation approaches which aims to reduce the computational cost, such as clustering based segmentation [2], thresholding based segmentation [5], and segmentation with prior knowledge [1] [4], the candidate segments given by a shadow detection algorithm is more reliable and no prior knowledge is needed.

Secondly, different from the original definition of sample points for Shape Context descriptor, which was extracted along the edge of the object subject to minimum distance requirement, we extract the sample points from the Harris Corner Response Map (HCRM). This is because the HCRM is more robust than the Edgemap (See Fig. 1). From Fig. 1b, we can see that the Harris corner response from vehicle edges is reasonably stronger than that of the tree edges (at the top-left corner of the image); while such an effect is not obvious in the edgemap because the edge responses from both the vehicle edges and tree edges are strong. The shape context descriptor is made rotational invariant by measuring relative angles of other sample points to the gradient direction of current sample point.

The rest of this paper will be organized as follows. In Sec. II, we introduce the framework of our approach. In Sec. III, we demonstrate our experimental results. Sec. IV gives conclusion on this paper.

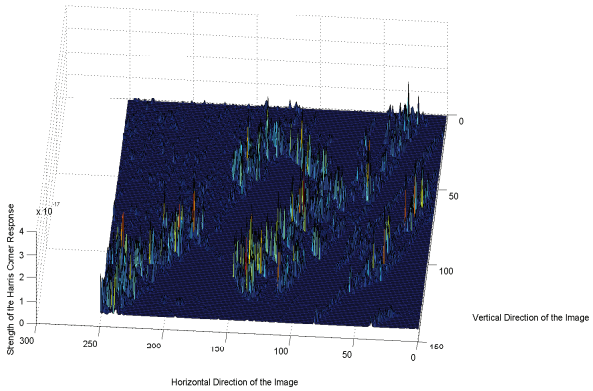
II. Framework

Fig. 2 gives an overview of our proposed framework. Input color aerial images are first put into shadow detection algorithm, as proposed in [3], then, from the candidate regions given by the shadow detection algorithm, Harris Corner Response Map (HCRM) is calculated, based on the HCRM, the top N points which gives the highest corner response are selected. Rotational invariant local shape features are extracted from those N points using Shape Context feature extraction algorithm. A classifier based on Euclidean distance is used to classify the candidiate regions into objects or non-objects.

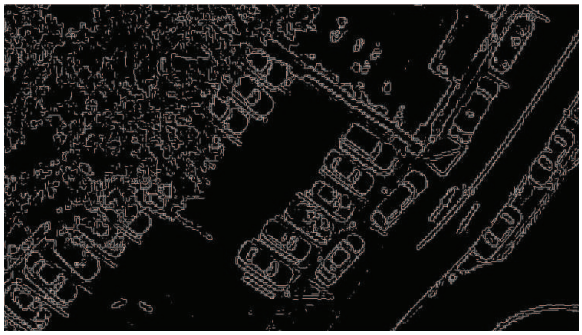
A. Shadow Detection

In many object detection algorithms, object shadows are considered as noise and the ultimate purpose is to remove the shadows. Shadow detection algorithms have been proposed in [3] and [9]. However, in this paper, we are going to use the object shadows as a way to detect potential object locations. In recent works, image segmentation is frequently used prior to feature extraction in order to reduce the computational cost. As mentioned in Sec. I, those works can be categorized into clustering based segmentation, thresholding based segmentation, and segmentation with prior knowledge. For clustering based segmentation (e.g. mean shift), challenges come from finding a good initial location, a reasonable search radius, and an appropriate bandwidth. For thresholding based segmentation, an optimized threshold is crucial. For segmentation with prior knowledge, it is a prerequisite to have accurate prior knowledge. Differing from those approaches, we propose to use shadow-based image segmentation. Without

(a) RGB color image



(b) 3D surface of HCRM



(c) Edgemap

Figure 1: HCRM vs. Edgemap

the need of any prior knowledge, using shadow-based image segmentation is more efficient than the scanning windows approach and clustering based segmentaion, more reliable than a thresholding based segmentation.

B. Extracting Corner Responses from Edgemaps

Early corner detectors (e.g. Moravec Corner Detector) defines a corner to be a point with low self-similariy (i.e. the patch centered on the pixel is compared with its surroundings, largely overlapping patches using Sum of Squared Distance (SSD) Measurement) [10]. Lower SSD value indicates higher self-similarity. Rather than using shifted patches, as used by

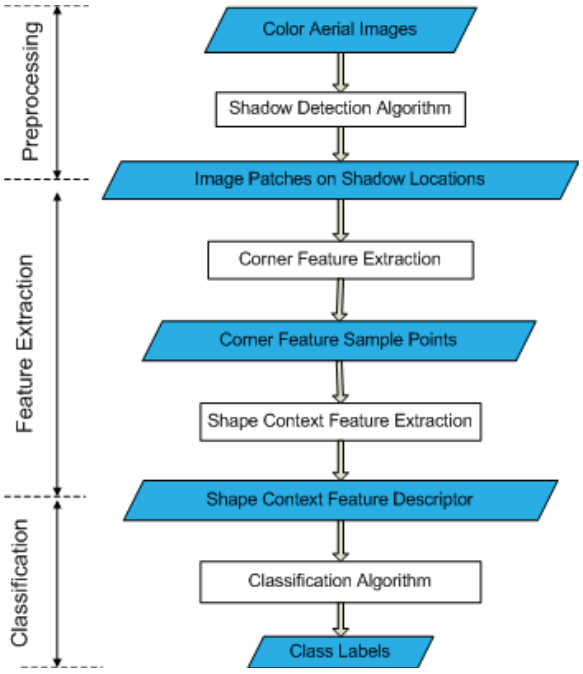
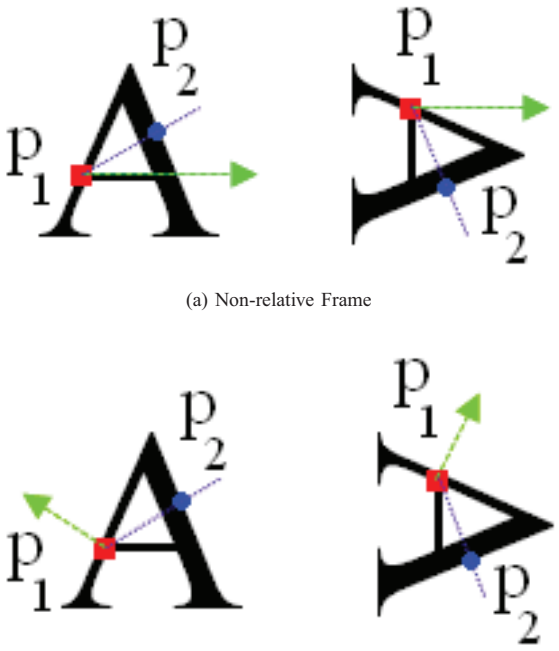


Figure 2: Framework of the Approach



(a) Non-relative Frame

(b) Relative Frame

Figure 3: Non-relative Frame vs. Relative Frame

the Moravec Corner Detector, the Harris Corner Detector advances by directly considering the differential of the corner score with respect to direction [11].

In our approach, the Harris Corner Response Map (HCRM) is used together with Edgemap of the RGB image to determine the candidate set of sample points for Shape Context feature extraction [12]. Denote E as the edgemap of the original color image. Denote R as the HCRM of original image.

The combined map R' of the HCRM with the Edgemap is represented by

$$R' = R \oplus E. \tag{1}$$

In particular, if E is a binary image with all edge pixels equals 1 and non-edge pixels equals 0, Eq. 1 can be rewritten as

$$R' = R * E. \tag{2}$$

From the convoluted map R', P points with strongest Harris Corner responses will be selected, those P points are the sample points for Shape Context feature descriptor.

C. Rotational Invariant Shape Context for Classification

The Shape Context feature descriptor was first introduced in [6]. In Shape Context, an object is represented by a matrix of $P \times B$, as mentioned in Sec. II-B, P is the number of sample points extracted from the object edges and B is the number of features for each sample point (i.e. the number of bins to quantize the distance and relative angles of all other sample points to the current sample point). Please refer to [6] and [13] for more details on the Shape Context feature descriptor.

In our test images (aerial images), the objects can be of various orientations, hence a rotational invariant feature descriptor is needed. As mentioned in [6], the Shape Context feature descritor can be made totally rotational invariant by measuring all angels relative to the tangent angle at each sample point. In [6], this frame is defined as the *relative frame*. Fig. 3 gives an illustration of the *relative frame* compared with the rotational variant frame. p_1 is the current sample point, and p_2 is one of other sample points.

From Fig. 3b we can see that the relative angle between $p_1\vec{p}_2$ (the blue dotted line) and the gradient orientation of p_1 (the green dashed line) remains the same when the object rotates clockwise by 90 degree. While in Fig. 3a, the relative angle between $p_1\vec{p}_2$ and the horizontal axis changes when the object orientation changes.

Based on the Shape Context descriptor for each sample point (i.e. a vector), a diagonal symmetric cost matrix C of size $P \times P$ can be obtained. Each element $c(i,j)$ in C represents the cost for matching sample point i on the reference object to sample point j on the matched object. $c(i,j)$ can be decomposed into two parts, the shape cost $c_s(i,j)$ and the appearance cost $c_a(i,j)$. Denote D_i and D_j as the Shape Context feature descriptors for sample point i and sample point j, respectively. $D_i(k)$ is the value for the k-th bin in D_i. The shape cost c_s is represented by

$$c_s(i,j) = \frac{1}{2}\sum_{k=1}^{B}\frac{[D_i(k) - D_j(k)]^2}{D_i(k) + D_j(k)} \tag{3}$$

Denote θ_i and θ_j as the orientations corresponding to sample

point i and sample point j, respectively, c_a is represented by

$$c_a(i,j) = \frac{1}{2}\left\| \begin{pmatrix} cos\theta_i \\ sin\theta_i \end{pmatrix} - \begin{pmatrix} cos\theta_j \\ sin\theta_j \end{pmatrix} \right\| \quad (4)$$

The final weighted cost $c(i,j)$ is

$$c(i,j) = (1-\beta)\cdot c_s(i,j) + \beta \cdot c_a(i,j). \quad (5)$$

In Eq. 5, β is a predefined ratio that controls the weight of shape costs versus the weight of appearance costs. In [6], $\beta = 0.3$.

With the given cost matrix C, Hungarian algorithm is used to obtain the optimal matching between i and j, such that the total matching cost T is minimized.

$$T = argmin \sum_{1\le i\le P,\ 1\le j\le P} c(i,j). \quad (6)$$

In our approach, each potential object image patch is matched against a pre-defined set of N_{ref} reference object image patches, the reference object image patches are selected from those images which contains vehicle objects. The cost matrix that gives the cost for pairwised matching between each sample point on one of the reference object and every sample point on the matched potential object will be calculated. After using the Hungarian algorithm over the cost matrices, a set of minimum matching costs $\{T_i, (1 \le i \le N_{ref})\}$ can be calculated, thresholding $\{T_i\}$ will assign a label L to the matched potential object. $L = 1$ denote that the potential object is a vehicle object, $L = 0$ otherwise. Denote n_T the number of T_i's that is smaller than a predefined threshold, then the classification function is represented by

$$L = \begin{cases} 1 & n_T > \frac{N_{ref}}{2} \\ 0 & otherwise. \end{cases} \quad (7)$$

III. Experiments

All of our experimental data are collected from Google Map satellite images [14]. We collect all sample images from Sydney, starting from the location of University of Technology, Sydney. This is to ensure that the illumination condition will not have large variation across different samples and every object higher above the ground will have shadow projected onto the ground.

In our experiments, we use a total of 72 images, those images contain a total of 736 visible vehicle objects. The size of each collected image varies from one to another, but the vehicle objects on those images are roughly bounded in a patch of size 100×100 pixels. As the sizes of vehicles are different, some larger vehicles (e.g. a truck or a bus) may not be fully bounded and some smaller vehicles may not occupy the whole patch. Neverthless, our experimental results demonstrate that those differences can be tolerated by the feature. This is because the Shape context feature mostly describe the correlation between sample points instead of describing the property of each individual sample point.

After shadow detection, the centroid for the object shadows will be used to roughly determine the vehicle location. For the feature extraction, we select 200 edge points from the image patch as sample points of Shape Context feature descriptor. Those 200 edge points have the strongest Corner responses compared to other edge points on the image patch.

We extracted 24 image patches as reference objects from 7 images. There are 44 visible objects in those 7 images. As a result the testing set have 65 images with 692 visible vehicle objects.

Of all 692 vehicle objects, our algorithm can successfully identify 641 vehicle objects, the detection rate is 92.6% and at the same time there are 8 false positve regions among all 65 test images.

Fig. 4 gives an illustration of our detection result. For some vehicles, the red dots are not exactly located in the centroid because it is used to roughly determine the location of the vehicles. In fact, it is the centroid of the bounding ellipse for the shadow of that vehicle.

From Fig. 4, we can see that there are repeated detections (i.e. two red dots identifying one vehicle) due to imperfect shadow-based image segmentation. Because for both segments, the matching cost is within the threshold. Also from Fig. 4b, we can see that all vehicles under the shadow are not detected. This is again because the shadow-based segmentation algorithm, apart from detecting shadow centroids, cannot identify objects being cloaked by the shadows. However, even for human, some of the shadowed vehicle objects in Fig. 4b are difficult to identify (e.g. two vehicles on the top left corner).

IV. Conclusion

In conclusion, we have proposed an efficient vehicle detection algorithm for aerial images based on an improved shape matching algorithm. The shadow detection algorithm has reduced the computational cost and corner detection gives more reliable sample points compared to edge detection. The shape context feature descriptor is rotational invariant using the *relative framework*. Our experimental results demonstate that the proposed approach gives satisfactory performance on one of the most popular search engines for aerial images.

References

[1] Line Eikvil, Lars Aurdal, and Hans Koren, "Classification-based vehicle detection in high-resolution satellite images," *ISPRS Journal of Photogrammetry and Remote Sensing*, vol. 64, no. 1, pp. 65 – 72, 2009.

[2] Jae-Young Choi and Young-Kyu Yang, "Vehicle detection from aerial images using local shape information," *Advances in Image and Video Technology*, vol. 5414, pp. 227–236, 2009.

[3] Kuo-Liang Chung, Yi-Ru Lin, and Yong-Huai Huang, "Efficient shadow detection of color aerial images based on successive thresholding scheme," *IEEE Transactions on Geoscience and Remote Sensing*, vol. 47, no. 2, pp. 671 –682, feb. 2009.

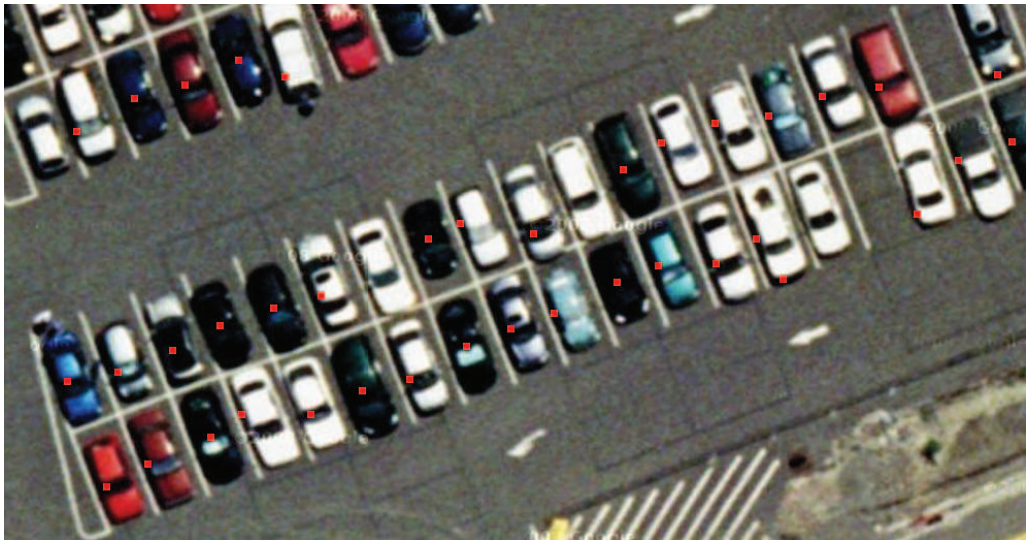
[4] Bo Yang, P. Sharma, and R. Nevatia, "Vehicle detection from low quality aerial lidar data," in *Applications of Computer Vision (WACV), 2011 IEEE Workshop on*, 2011, pp. 541 –548.

[5] G. Sharma, C.J. Merry, P. Goel, and M. McCord, "Vehicle detection in 1-m resolution satellite and airborne imagery," *International Journal of Remote Sensing*, vol. 27, no. 4, pp. 779–797, 2006.

[6] S. Belongie and J. Malik, "Matching with shape contexts," in *Proc. IEEE Workshop on Content-based Access of Image and Video Libraries*, 2000, pp. 20 –26.

[7] Bo Wu and Ram Nevatia, "Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors," *International Journal of Computer Vision*, vol. 75, no. 2, pp. 247–266, 2007.

[8] Paul Viola and Michael Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE conference on Computer Vision and Pattern Recognition*. 2001, pp. 511–518, IEEE.



(a) Example 1

(b) Example 2

Figure 4: Detection Result

[9] HaiYan Yu, JunGe Sun, LiNing Liu, YunHong Wang, and YiDing Wang, "Mser based shadow detection in high resolution remote sensing image," in *Proc. The Ninth International Conference on Machine Learning and Cybernetics*, 2010, pp. 780 –783.

[10] H. Moravec, "Obstacle avoidance and navigation in the real world by a seeing robot rover," *Technical Report CMU-RI-TR-3 Carnegie-Mellon University, Robotics Institute*, 1980.

[11] C. Harris and M. Stephens, "A combined corner and edge detector," in *Proc. 4th Alvey Vision Conference*, 1988, pp. 147–151.

[12] A. Vedaldi and B. Fulkerson, "{VLFeat}: An open and portable library of computer vision algorithms," http://www.vlfeat.org/, 2008.

[13] S. Belongie, J. Malik, and J. Puzicha, "Matching with shape contexts," http://www.eecs.berkeley.edu/Research/Projects/CS/vision/shape/sc_digits.html, 2001.

[14] Google Maps, "Sydney - google maps," http://maps.google.com.au/maps?hl=en&ie=UTF8&ll=-33.883812, 151.200655&spn=0.001178,0.002575 &t=k&z=19, 2011.