# AN EXPLORATION OF ON-ROAD VEHICLE DETECTION USING HIERARCHICAL SCALING SCHEMES

*Yi-Min Tsai, Keng-Yen Huang, Chih-Chung Tsai, and Liang-Gee Chen*

DSP/IC Design Lab., Graduate Institute of Electronics Engineering
National Taiwan University, Taipei, Taiwan

## ABSTRACT

This paper targets at detecting preceding vehicles in a wide range of distance. We propose an Adaboost-based approach combined with hierarchical image and sub-window scaling schemes. The relationship is investigated among object characteristics, image structures and image scales. A parameter set is developed to easily adjust overall performance, which benefits researchers to establish a vehicle detection system. It achieves 96.6% detection rate with 2.0% false alarm rate along proposed methodology. The benchmark of several learning-based vehicle detection approaches is also provided. The results show the outperformance of the proposed method.

***Index Terms*—** Vehicle detection, Adaboost, Haar-like, Image scaling, Detection rate

## 1. INTRODUCTION

Nowadays, vehicular technologies are developed to solve not only safety but also energy-saving problems that have drawn intensive attention. Owing to the maturity of vision sensors, vision-based systems play an essential role in many applications. Blind Spot Warning Systems (BSWS) attempt to monitor and detect objects not seen by drivers. Advanced Driver Assistance Systems (ADAS) provide driving guideline for drivers. Moreover, Collision Warning Systems (CWS) prevent vehicles from sudden crashes. These applications all involve localization and identification of on-road vehicles, which suggests the demand of robust vehicle detection and recognition.

## 2. PRIOR ARTS

Generally, a vision-based object analysis flow can be classified into two stages after video capturing (Fig. 1). Object detection is to locate where candidates are in a frame and object recognition is to verify if the candidates are the desired patterns. Sun et al. made an overview for vision-based on-road vehicle analysis [1], which is strictly divided into hypothesis generation and hypothesis verification corresponding to detection and recognition respectively.

For object detection, knowledge-based methods utilize edge [2], corner [3], and symmetry [2, 3] to identify vehicles. Yet, the approaches are sensitive to environmental factors such as changing illumination. Motion-based methods use motion vectors such as optical flow [4] to locate objects with large displacement but such methods suffer from correspondence problems.

As for object recognition, template-based methods [5] utilize predefined vehicle template to verify suspected patterns through correlation. However, their performance may decisively rely on the created templates. In recent years, machine-learning makes progress in object recognition [6–12]. Features such as Haar-like [6, 7] and Gabor [9, 10] are cooperated with classifiers, including SVM and Adaboost, for vehicle recognition and categorization. These learning-based methods yield a decent performance in the recent literatures.

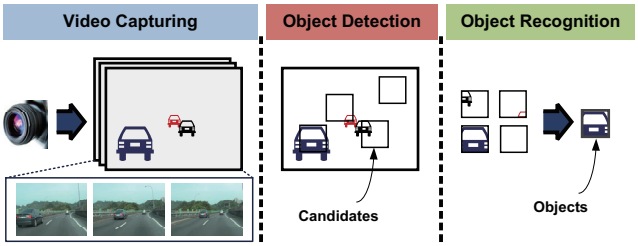



**Fig. 1**. General object detection and recognition flow.

This paper is organized as following. We briefly introduce the Adaboost-based approach and its pitfalls in Sec. 3. Proposed methods are presented in Sec. 4 and Sec. 5. In Sec. 6, experiment results are discussed. Sec. 7 summarizes our exploration and contributions.

## 3. ADABOOST-BASED APPROACH: OVERVIEW

Viola and Jones [6] proposed an object recognition method using Adaboost with Haar-like features. The method is combined with a sub-window scaling detection routine [8] (Fig. 2(a)). To ensure not to miss any suspected patterns, every scaled sub-window is viewed as a candidate and is verified through a boosted classifier.

**Detection**: To detect both small-sized and large-sized objects, sub-windows progressively scan entire image and are up-scaled with a window scaling factor $WSF$ for the next scan phase (Fig. 2(a)).

**Recognition**: It becomes a binary classification problem. To distinguish these scaled sub-windows, a stage cascade composed of trained weak classifiers is constructed statistically. Non-object sub-windows can be rapidly rejected with early stage classifiers. More complex stage classifiers are performed subsequently on object-like sub-windows only if they passed through previous stage classifiers.

**Grouping**: Each detected object may contain multiple overlapping verified neighboring sub-windows, which indicates multiple hits. Neighboring sub-windows are grouped into a single window representing the detected object. The more neighboring sub-windows are combined, the higher confidence the object has. Eventually, if the object's confidence is greater than a designed threshold, it is accepted as a verified object.

Although the detection scheme achieves low-miss outcome on account of full-scanning procedure, it also introduces many false alarms at the same time. Besides, the method is mainly for detecting objects in similar depth, which is not appreciated in situations like detecting vehicles in a variety of distances.

In this paper, we proposed hierarchical image and window scaling schemes. Observations and prior knowledge are involved to significantly improve overall accuracy and reduce computational time. In addition, multiple cascades are considered to further enhance recognition capability.

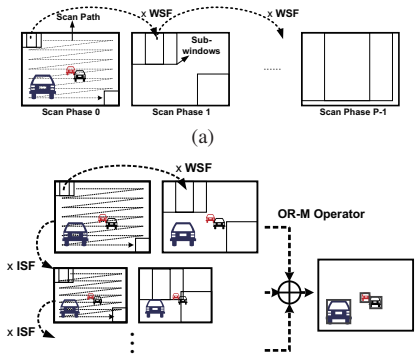



**Fig. 2**. Detection routines. (a)Conventional sub-window scaling routine. (b)Proposed hierarchical approach.

## 4. PROPOSED METHODS: DETECTION

The proposed hierarchical detection routine considers both sub-window scaling and image scaling. We conduct observations on object size corresponding to support detecting distance in different image scales. By analyzing object trajectories and image structures, region of interests are marked off. These correlations are expressed with a principal parameter set $\{ISF, WSF, \alpha, s, \beta, \gamma\}$.

### 4.1. Hierarchical Detection Routine

There are two steps in the routine (Fig. 2(b)). Firstly, an image pyramid is constructed through down-sampling with an image scaling factor (ISF). For each level in the pyramid, sub-window scaling is adopted except that the size of scaled sub-windows is constrained. Secondly, the detection results of each level are combined through an OR-Maximum (OR-M) operator.

In the first step, we determine the maximum size of scaled sub-windows in each level along a relationship between detecting distance and object width in pixel. Fig. 3(a) shows the object width in pixel is approximately inversely proportional to the distance for each level, called O-D curves which can be formulated by,

$$\begin{aligned} wo_l \times D &= K \times W_l \\ &\cong ws_l \times D \end{aligned} \tag{1}$$

where $D$ is the detecting distance. The subscript $l$ stands for level count ranging from 0 to $L-1$. $wo_l$ and $W_l$ are the object width and image width in pixel for each level respectively. $K$ is a constant ranging from 2.0 to 2.5 depending on pixel aspect ratio and sensor parameters. The formula can be approximated with sub-window width $ws_l$ if WSF is close to 1. According to Eq. 1, we shift the regression line of the bottom O-D curve, that is level $L-1$, to derive the maximum scaled sub-window width $ws_{max_l}$. The shifted regression curve is written as,

$$(ws_{L-1} + ws_{min}\alpha) \times D = K \times W_{L-1} \tag{2}$$

where $ws_{min}$ is the minimum sub-window width (the same as training patch width) and is same for each level. $\alpha$ is a shifting factor. Form Eq. 1 and Eq. 2, $ws_{max_l}$ is then given by,

$$\begin{aligned} ws_{max_l} &= \frac{ws_{min}\alpha \times W_l}{(W_l - W_{L-1})} \\ &= \frac{ws_{min}\alpha}{1 - ISF^{L-1-l}} \end{aligned} \tag{3}$$

A distance overlap (DOV) is defined as the overlapped detecting distance between two layers under constraints $ws_{max_l}$ and $ws_{min}$ (Fig. 3(a)). Therefore, by adjusting $\alpha$ to a higher value, DOVs are enlarged. Larger DOVs imply same objects may be detected in different levels, which benefits the OR-M operator in the next step. This also suggests that if an object is not detected in level $L-n$, it may be detected in level $L-(n+1)$ or $L-(n+2)$. Consequently, the larger the DOVs are, the better the performance is.

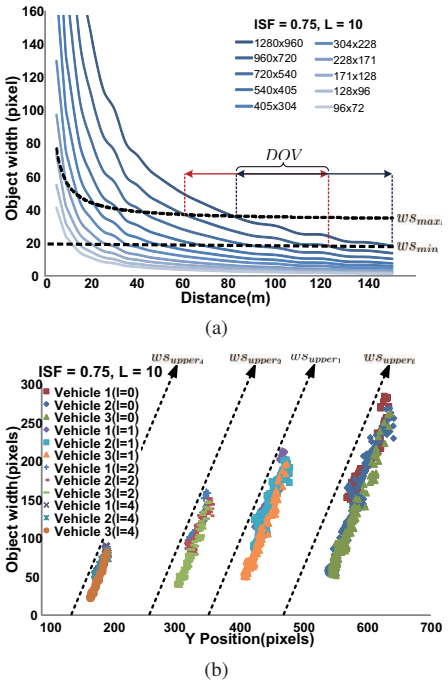



**Fig. 3**. Observations on object characteristics. (a)Relationship between detecting distance and width of the targeted objects in pixel. A DOV is illustrated (the overlapping segment). (b)Trajectories of objects in different resolution layers.

In the second step, an OR-M binary operator ($\bigoplus$) is defined as selecting the detected objects with highest conference among levels. A lower grouping threshold is applied to each layer before OR-M operator but a higher one is set afterwards. Accordingly, it has higher probability to accept true positives through the OR-M operator.

### 4.2. Intra-level Sub-window Reduction

In a front-faced view image, objects do not appear in some locations such as sky. Hence, there are many redundant scaled sub-windows. An intra-level sub-windows reduction constraint is proposed.

Fig. 3(b) records trajectories of three types of vehicles for four levels, and each point represents a 2D position corresponding to that vehicle in temporal domain. An observation is made that an object's width in pixel is approximately a linear function of the projected 2D trajectory (Y-coordinate) of that object. Therefore, an upper-bound of scaled sub-window size is applied in each scan phase. The upper-bound $ws_{upper_l}$ is a regression of experiential data and statistical data, which can be formulated by,

$$\begin{aligned} ws_{upper_l} &= (s \times PosY_0 - \beta_0) \times ISF^l \\ &= s \times PosY_l - \beta_l \end{aligned} \tag{4}$$

where $s$ is a position correlated factor and $\beta$ is a position offset depending on the image horizon. $PosY_l$ is denoted as image Y-coordinate. In a certain y position, a sub-window with width $ws_l$ larger than $ws_{upper_l}$ is not processed for further recognition. Accordingly, there are scanning and non-scanning regions in an image (Fig. 4(a)). Therefore, redundant sub-windows are eliminated, which implies the higher chance to reduce false alarms.

### 4.3. Inter-level Region of Interest Selection

Suppose the optical axis is parallel to the ground, an image horizon is defined as the horizontal line passing through the optical center. Intuitively, with the front-faced camera setup, the closer to an image horizon an object is, the smaller it is. It is obvious that far-objects only appear in a narrow strip near the image horizon. Thus, we restrict the region of interest (ROI) of each level from full-frame region to a narrow strip. Besides, we detect far-objects in higher resolution level and leave near-object detection in lower resolution level (Fig. 4(b)). On the other hand, near-objects locate at the narrow strip ROI in lower resolution because of image down-sampling.

From Eq. 3, there exists a level $l^{'}$ such that Eq. 5 come into existence. A distance gap $h_{l^{'}}$ is decided and described by Eq. 6,

$$wo_0 \times ISF^{l^{'}} \leq ws_{min}\alpha \leq \frac{ws_{min}\alpha}{1-ISF^{L-1-l^{'}}} \tag{5}$$

$$(pos_0 - IH_0) \times ISF^{l^{'}} \leq h_{l^{'}} \tag{6}$$

where $pos_0$ and $IH_0$ are denoted as Y-coordinate value of a sub-window center and that of the image horizon in the original resolution respectively. From Sec. 4.2, object width can be approximated as a linear function of its 2D position and be expressed by,

$$\begin{aligned} wo_0 &\simeq \kappa \cdot pos_0 - \eta \\ &= \kappa^{'}(pos_0 - IH_0) \end{aligned} \tag{7}$$

From Eq. 5 and Eq. 7, we rewrite Eq. 6 into Eq. 8 to describe $h_{l^{'}}$. $i$ is substituted for $i^{'}$ without losing generalities. At last, Eq. 9 depicts the desired ROI strip $H_{ROI_l}$ is two times the height of $h_l$ plus an error term $\delta$ caused by video oscillation. Eq. 9 suggests a constant height ROI strip in every layer. In addition, the greater the $ISF$ is, the narrower a ROI strip is.

$$\begin{aligned} (pos_0 - IH_0) \times ISF^{l^{'}} &\leq ws_{min}\frac{\alpha}{\kappa^{'}} \\ &= \gamma \cdot ws_{min} \\ &= h_{l^{'}} \end{aligned} \tag{8}$$

$$\frac{1}{2}H_{ROI_l} = \gamma \cdot ws_{min} + \delta \tag{9}$$

## 5. PROPOSED METHODS: RECOGNITION

The physical size of vehicles is similar in width but diverse in height, which results in different object aspect ratio. Consequently, multiple cascades are created instead of using a single cascade for recognition. The resultant object confidence is the average of those independently produced by each cascade. The resultant confidence $C_R$ is given by Eq. 10 and S is the total number of cascades.

$$C_R = \frac{\sum_{i=1}^{S} C_i}{S} \tag{10}$$

## 6. EXPERIMENTAL RESULTS

Real sequences were captured from a 24fps CMOS front-mounted camera with 640×480 and 1280×960 intermediate resolutions. 2034 vehicles' rear parts and 2540 non-vehicles are trained to build four 20-stage cascades with Adaboost. The trained window sizes are 20×20, 20×16, 20×12, and 20×10 ($ws_{min}$=20). For comparison, we re-implemented VJ's method using OpenCV library. Programs are run on a PC with Intel Duo Core 2.67GHz CPU equipped.

Videos with 1280×960 resolutions are selected for the following experiments. We examine detection rate (DR) and false alarm rate (FAR) under different $\{ISF, WSF, \alpha, s, \beta, \gamma\}$ sets. Receiver operating characteristic (ROC) curves are constructed afterwards. The term *conventional* indicates VJ's method with sub-window scaling.

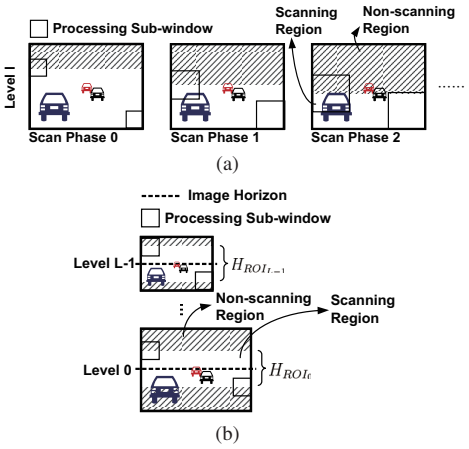



**Fig. 4**. Sub-window scanning orders. (a)With intra-level sub-window reduction. (b)With inter-level region of interest selection.

In Fig. 5(a), the hierarchical approach is applied without intra-level sub-window reduction (intra- for shorthand) and inter-level region of interest selection (inter- for shorthand). Image sub-sampling is executed through bilinear interpolation. Besides, $ISF$ is equal to the inversion of $WSF$ in conventional method. The minimum down-sampled resolution is one-sixteenth of the original resolution. In conclusion, 4-cascade structure is superior to 1-cascade and 2-cascade structures. The proposed method outperforms the conventional one. Under 3% FAR, the proposed method with three parameter sets all outperforms *conventional* with 4 to 10% margin. Among them, the hierarchical method with $ISF$ 0.88 achieves best performance and yields 95.6% DR with 4.0% FAR.

Fig. 5(b) and 5(c) show the improvement on performance with intra- and inter- involved. $s$ and $\beta$ are equal to 1.8 and 800 respectively. The greater $ISF$ implies a larger DOV, a narrower ROI, and more resolution levels. As $\alpha$ increases, the performance raises but saturates at the boundary where $\alpha$ is about 4.0. This infers further raise $\alpha$ is less beneficial. Our method yield 96.6% DR with 2% FAR and 95% DR with only 1% FAR.

Fig. 6 demonstrates the effects with different intra- and inter-parameters. We consider the situation under low FAR (less than 10%). In Fig. 6(a) , as the intercept $\beta$ increases, the performance drops because the upper-bound $ws_{upper}$ becomes stricter. In Fig. 6(b) , the saturation occurs when $\gamma$ is equal to 0.75 which leads to 50-pixel $H_{ROI}$ in height if oscillation term is equal to 5 pixels. As expected, DR increases as $\gamma$ increases.

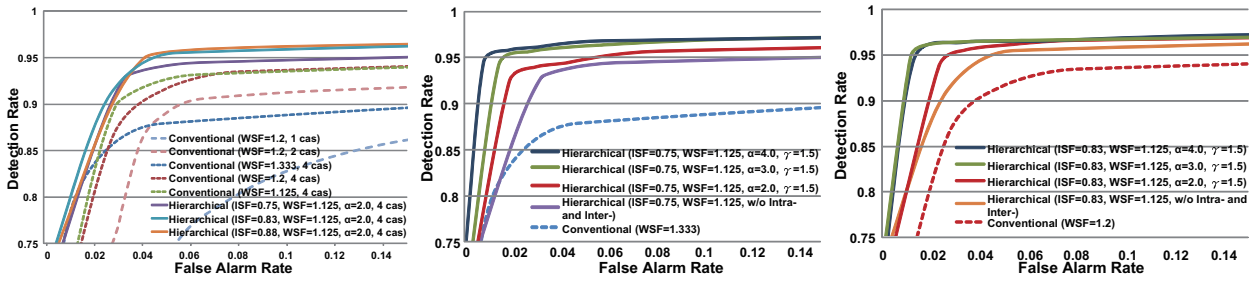
The state-of-the-arts are analyzed in five aspects (Table 1). The results show the proposed method outperforms other methods. The average processing time is listed for two resolution specifications. 0.18 and 2.13 seconds per frame are needed for 640×480 and 1280×960 videos respectively. However, 10 to 20 fps at least is required for a real-time vision-based system. This suggests the necessity of hardware participation, such as DSPs or ASICs. At last, Fig. 7 illustrates detection results in different driving environments, including tunnel, highway, and typical road.

## 7. CONCLUSIONS

Our contribution is twofold. (1)A recognition-oriented vehicle detection flow is proposed, which integrates image and sub-window scaling into a complete routine. (2)A parameterized set is organized to explore the trade-off between accuracy and efficiency, which provides researchers to establish related systems. Totally, the proposed framework declares convincing detection for on-road vehicles.

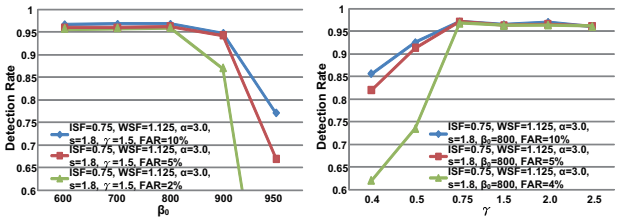
| Methods | Detection | Recognition (feature/classifier) | Accuracy(%) (DR / FAR) $\{ISF, WSF, \alpha, s, \beta, \gamma\}$ | Average Processing Time (s/frame) (640×480/1280×960) |
|---|---|---|---|---|
| Fu [11] | Static ROIs | Edge/SVM | 87.6 / N/A | N/A |
| EGFO [10] | Edge-based | Gabor/SVM | 91.0 / 6.4 | N/A |
| BGF [9] | Static ROIs | Gabor/Boosting+SVM | 95.8 / 8.8 | N/A |
| VJ [6] | Scaled sub-window | Haar-like/Adaboost | 96.4 / 15.5 | 3.98 / 12.81 |
| Proposed | Scaled sub-window + scaled image + intra- + inter- + | Haar-like/Adaboost | 95.8 / 2.0 $\{0.75, 1.125, 3.0, 1.8, 800, 1.5\}$ 96.6 / 2.0 $\{0.83, 1.125, 3.0, 1.8, 800, 1.5\}$ | 0.18 / 2.13 0.57 / 3.62 |

**Table 1**. Benchmark of state-of-the-arts and proposed method.



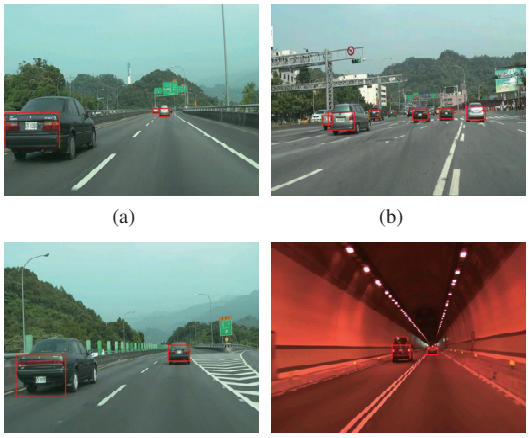

(a) (b) (c)

**Fig. 5**. ROC curves comparison between conventional and hierarchical detection schemes. (a)No intra- and inter- in hierarchical method. Multiple cascades are considered in conventional methods. (b)Hierarchical method with intra- and inter-. $ISF$ is equal to 0.75 with $\alpha$ ranging from 2.0 to 4.0. $s$ is 1.8 and $\beta$ is 800. (c)Similar to (b) but $ISF$ is equal to 0.83.




(a) (b)

**Fig. 6**. Effects on detection rate. (a)Intra-level sub-window reduction with different $\beta$. (b)Inter-level region of interest selection with different $\gamma$.



(a) (b) (c) (d)

**Fig. 7**. Detection results (red bounding boxes) in different situations.

## 8. REFERENCES


[1] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection: A review," *IEEE Trans. on PAMI*, pp. 694–711, May 2006.

[2] D. Alonso, L. Salgado, and M. Nieto, "Robust vehicle detection through multidimensional classification for on board video based systems," in *Proc. IEEE ICIP*, Sept. 2007, vol. 4, pp. IV–321–IV–324.

[3] J. Arrospide, L. Salgado, M. Nieto, and F. Jaureguizar, "On-board robust vehicle detection and tracking using adaptive quality evaluation," in *Proc. IEEE ICIP*, Oct. 2008, pp. 2008–2011.

[4] J. Wang, G. Bebis, and R. Miller, "Overtaking vehicle detection using dynamic and quasi-static background modeling," in *Proc. IEEE CVPR*, June 2005, pp. 64–64.

[5] A. Bensrhair, M. Bertozzi, A. Broggi, P. Miche, S. Mousset, and G. Toulminet, "A cooperative approach to vision-based vehicle detection," in *Proc. IEEE Intelligent Transportation Systems*, Aug. 2001, pp. 207–212.

[6] P. Viola and M. Jones, "Rapid object detection using a boosted cascade of simple features," in *Proc. IEEE CVPR*, 2001, vol. 1, pp. I–511–I–518.

[7] P. Negri, X. Clady, S. Hanif, and L. Prevost, "A cascade of boosted generative and discriminative classifiers for vehicle detection," *EURASIP Journal on Advances in Signal Processing*, 2008.

[8] M. Hiromoto, H. Sugano, and R. Miyamoto, "Partially parallel architecture for adaboost-based detection with haar-like feature," *IEEE Trans. on Circuits and Systems for Video Technology*, vol. 19, pp. 41–52, Jan. 2009.

[9] H. Cheng, N. Zheng, and C. Sun, "Boosted gabor features applied to vehicle detection," in *Proc. IEEE ICPR*, 2006, vol. 1, pp. 662–666.

[10] Z. Sun, G. Bebis, and R. Miller, "On-road vehicle detection using evolutionary gabor filter optimization," *IEEE Trans. on Intelligent Transportation Systems*, vol. 6, pp. 125–137, June 2005.

[11] Chih-Ming Fu, Chung-Lin Huang, and Yi-Sheng Chen, "Vision-based preceding vehicle detection and tracking," in *Proc. IEEE ICPR*, 2006, vol. 2, pp. 1070–1073.

[12] T. Liu, N. Zheng, L. Zhao, and H. Cheng, "Learning based symmetric features selection for vehicle detection," in *Proc. IEEE Intelligent Vehicles Symposium*, June 2005, pp. 124–129.