Sequential Architecture for Efficient Car Detection

Zhenfeng Zhu[1,2]
[1] Institute of Information Science,
Beijing Jiaotong University,
P. R. China
[2] National Key Laboratory on
Machine Perception, Peking
University, P.R. China
zhfzhu@.bjtu.edu.cn

Yao Zhao
Institute of Information Science,
Beijing Jiaotong University,
P. R. China
yzhao@bjtu.edu.cn

Hanqing Lu
National Laboratory of Pattern
Recognition, Institute of Automation,
CAS, P. R. China
luhq@nlpr.ia.ac.cn

Abstract

Based on multi-cue integration and hierarchical SVM，we present a sequential architecture for efficient car detection under complex outdoor scene in this paper. On the low level, two novel area templates based on edge and interest-point cues respectively are first constructed, which can be applied to forming the identities of visual perception to some extent and thus utilized to reject rapidly most of the negative non-car objects at the cost of missing few of the true ones. Moreover on the high level, both global structure and local texture cues are exploited to characterize the car objects precisely. To improve the computational efficiency of general SVM, a solution approximating based two-level hierarchical SVM is proposed. The experimental results show that the integration of global structure and local texture properties provides more powerful ability in discrimination of car objects from non-car ones. The final high detection performance also contributes to the utilizing of two novel low level visual cues and the hierarchical SVM.

1. Introduction

Object detection in 2-D images has been a deeply investigated field in vision community. Performance of object detection depends mainly on how to characterize the objects to be detected and what classifiers to be applied. Currently much attention has been mainly paid to three classes of object detection such as face detection, pedestrian detection and car (vehicle) detection. Here we focus our attentions on side-view car detection under complex outdoor scene.

A statistical method for vehicle detection based on local features [1] has been investigated and the EM algorithm was applied to learning the model parameters of the constellation of local features. In [2,3], a novel sparse parts-based car detection system was proposed, in which the SNoW architecture was learned based on the parts automatically extracted from some Förstner interest points and the position relations among parts. To the extension of this work, A.Garg et al. [4] proposed a method by fusing global (ICA-based) and local (parts-based) information to obtain a high object detection accurate. But both them pay no more attentions to the computational cost. Some other part-based or interest-point based object detection techniques can be found in [5, 6].

In [7], Schneiderman et al. successfully developed a wavelet histogram based statistic method for multi views face and car detection. But high computational cost is its main problem. Based on Harr wavelet and SVM, Papageorgiou et al. [8] proposed a general object detection framework and applied it to face, pedestrian and car detection. In addition, Viola[9] has advanced a more popularly recognized technique for face detection, in which the Adaboost classifier was trained on a series of Harr wavelet-like coefficients and the requirement of real time was achieved by the technique of integral image and cascade strategy. Furthermore, Sun et al. [10] provided an on-road vehicle (rear views) detection using Gabor filter and SVM. But it relies on some other vehicle locating methods and only can be taken as a verification step.

As an excellent classifier, the nature of SVM is to find a discriminate hyper-plane that optimally separates two classes of objects by using structural risk minimization [11] ,and it has been widely used in the task of object detection [8,10,12,25]. To improve the computational efficiency of SVM, a fast solution approximating method for SVM was developed [13, 14] and successfully applied to face detection by S. Romdhani [15]. In fact a mean shift based optimization technique [16] is adopted to find the approximating solution in their work. As discussed in [16], the mean shift based optimization can't be guaranteed to be convergent with negative weights. Hence, as in [17] a pair-wise method was proposed to solve the problem of convergence. But the object function to be minimized also can't be guaranteed to be monotonously decreased. Considering the relativity among training data, a greedy subspace seeking method based on kernel mapping was proposed by G. Baudat [18] and V. Franc [19] to find the approximating function.

In this paper we present a novel sequential architecture for efficient car detection based on multi cues integration

1-4244-1180-7/07/$25.00 ©2007 IEEE

and hierarchical SVM under complex outdoor scene. Both global structure cue and local texture cue are introduced in our work. To approximately characterize the global structure property we propose an odd Gabor moments via the Angular Radial Transform (***ART***) on the odd Gabor filter responses. Meanwhile, the multi-channel even Gabor filters based local texture is modeled as a Gaussian distribution to have an enhance object representation. To rapidly reject large parts of background without containing any candidate objects at the cost of missing few of the true car objects, two novel low level visual cues based templates (Edge Area Template ,***EAT*, and Corner Area Template, ***CAT***) are designed, which could be applied to forming the identities of visual perception to some extent. Considering the efficiency of general SVM, a sequential greedy search method is proposed to seek a subspace approximating solution to the original hyper plane of SVM, and thus a two-level hierarchical SVM can be constructed. The final experimental results show that the proposed architecture can obtain a high detection performance and low computational cost compared to [2, 4].

2. Discrimination based on identities of visual perception

For most cases of object detection, there exist a huge number of non-objects (corresponding to sub-windows in a general image) that can be rapidly eliminated only based on some low level features without need to introduce higher-level mechanism. In the first phrase of our car detection architecture, both edge area and corner area template are automatically constructed and then applied to forming the identities of visual perception to some extent to perform such task.

2.1. Edge Area and corner area templates

The original idea of constructing area templates comes from Hausdorff distance based object matching [20]. To extract the edge and interest point maps, Canny edge detector and Harris corner detector [21] are used. For consideration of positioning deviation, the Edge and corner maps are dilated with 3×3 and 5×5 structures respectively. Thus we define the edge area template (***EAT***) T_e and corner area template (***CAT***) T_c as

$$T_e(x,y)=\begin{cases}1 & \frac{1}{N}\sum_{n=1}^{N}E_n^d(x,y)\geq th_e\\ 0 & \frac{1}{N}\sum_{n=1}^{N}E_n^d(x,y)<th_e\end{cases},T_c(x,y)=\begin{cases}1 & \frac{1}{N}\sum_{n=1}^{N}C_n^d(x,y)\geq th_c\\ 0 & \frac{1}{N}\sum_{n=1}^{N}C_n^d(x,y)<th_c\end{cases} \quad (1)$$

where N is the number of positive training examples, $C_n^d(x,y)$ denotes the dilated corner map, and $E_n^d(x,y)$ the dilated edge map， th_{e} and th_{c} are two thresholds which are set 0.25 and 0.7 respectively in our work . Fig.1 gives an illustration of the constructing procedure of corner area template (***CAT***) and edge area template ***(EAT***). To eliminate the inverse influence from background, we filter the corner map by an ellipse mask in advance.

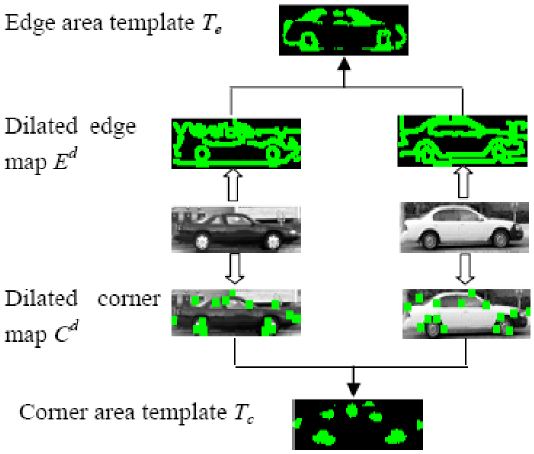


Figure 1 Illustration of the constructing procedure of corner area template (***CAT***) and edge area template ***(EAT***)

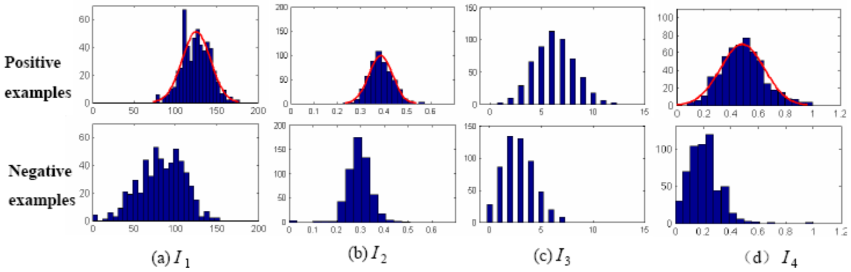
2.2. Identities of visual perception

Given the inputted sub-window $\mathbf{I}(x,y)_{m\times n}$ to be discriminated, let $C(x,y)$ and $E(x,y)$ denote the corresponding corner and edge map respectively, $n_1^c=\sum_{x,y}C(x,y)\cdot T_c(x,y)$ denotes the number of corners contained in corner area, $n_2^c=\sum_{x,y}C(x,y)\cdot 1_{m\times n}$ is the number of all detected corners in the given image $\boldsymbol{I}(x,y)$, $n_1^e=\sum_{x,y}E(x,y)\cdot T_e(x,y)$ is the number of edge pixels contained in edge area, and $n_2^e=\sum_{x,y}E(x,y)\cdot 1_{m\times n}$ is the number of all edge pixels contained in image $\boldsymbol{I}(x,y)$. Thus for the given image $I(x,y)_{m\times n}$, four qualifications based on the edge area and corner area templates, which reflect the 'identities' of visual perception to some extent, could be given as:

$$\mathbf{Q}_1(I)=\begin{cases}1 & I_1=n_1^e\prec C_1\\ 0 & else\end{cases}, \quad \mathbf{Q}_2(I)=\begin{cases}1 & I_2=\frac{n_1^e}{n_2^e}\prec C_2\\ 0 & else\end{cases}$$

$$\mathbf{Q}_3(I)=\begin{cases}1 & I_3=n_1^c\prec C_3\\ 0 & else\end{cases}, \quad \mathbf{Q}_4(I)=\begin{cases}1 & I_4=\frac{n_1^c}{n_2^c}\prec C_4\\ 0 & else\end{cases} \quad (2)$$

where C_i denotes the corresponding decision domain and I_i s denote the primary (edge and interest point) map based identities of visual perception. Thus the final



(a) I_1 (b) I_2 (c) I_3 (d) I_4

Figure 2 distributions of different identities I_i s of visual perception for positive car objects and negative non-car ones

qualification $\mathbf{Q}$ for $\boldsymbol{I}(x,y)$ can be obtained as

$$\mathrm{Q}(I)=\mathrm{Q}_1(I)\wedge\mathrm{Q}_2(I)\wedge\mathrm{Q}_3(I)\wedge\mathrm{Q}_4(I) \qquad (3)$$

When $\mathbf{Q}(I)$ is true, the current sub-window $\mathbf{I}(x,y)_{m\times n}$ will be passed to the next phase for further verification; otherwise, it will be rejected without need of further consideration. From Eq. (2) and (3), we can see that the key factor for forming the final qualification $\mathbf{Q}$ is how to determine the aforementioned decision domains C_i s. As shown in Fig.2， since all the identities I_i s can be approximately characterized by Gaussian normal distributions, the corresponding decision domain can be naturally determined by following the $3\sigma_i$ strategy, where σ_i denotes the variance of corresponding Gaussian distribution.

3. Object representation based on odd Gabor moments Mathematics

How to find an appropriate or compact representation for the object to be detected is a key step for the task of object detection. Although there exist currently many ways to take, the types of features effective for discriminating some object class may be less useful for other object classes. To obtain an effective representation in our case, the global structure property for an object is extracted by using odd Gabor moments.

Gabor filters, which have been shown to fit well the receptive fields of the majority of simple cell in the primary visual cortex [22], are modulation products of Gaussian and complex sinusoidal signals. A 2D Gabor filter oriented at angle θ is given by [23]:

$$\mathbf{G}_j(x,y)=k_j^2\exp(-\frac{k_j^2x^2}{2\sigma^2})\exp(i\vec{k}_j\bar{x})$$
$$\vec{k}_j=\begin{pmatrix}k_{jx}\\ k_{jy}\end{pmatrix}=\begin{pmatrix}k_v\cos(\theta_u)\\ k_v\sin(\theta_u)\end{pmatrix},\qquad k_v=2^{-\frac{v+2}{2}}\pi \qquad (4)$$

where σ denotes the standard deviation of the circle Gaussian, and $\vec{k}_j$ denotes the spatial frequency. The even real part and odd imaginary part of 2D Gabor filter is denoted by $\mathbf{G}_j^{\mathrm{e}}(x,y)$ and $\mathbf{G}_j^{o}(x,y)$ respectively.

For a given input image $I(x,y)$, the Gabor filter response is denoted by

$$\mathbf{R}_j(x,y)=\mathbf{G}_j(x,y)*\mathbf{I}(x,y) \qquad (5)$$

where * denotes two-dimensional convolution operator. Here we denote R^e as even Gabor filter response, which will be served for local texture based object representation in next section, and R^o as odd Gabor filter response. Thus the final energy map for odd Gabor response, which can effectively reflect the structure property [24], is defined as

$$\mathbf{E}(x,y)=\left\|\mathbf{R}_j^o(x,y)\right\|_2 \qquad (6)$$

where $\|\cdot\|_2$ denotes Euclidean norm. Note that in odd Gabor case only two orientations and one scale are adopted, i.e. $\theta_u=\frac{u\pi}{2}, j=u+v, u=0,1, v=0$. To reduce the inverse influence of illumination, the final energy map $E(x,y)$ is mapped into $[0,255]$.

As we can see that the dimension of $E(x,y)$ keeps the same as the input image $I(x,y)$ with the size of $m\times n$. To obtain a low dimension representation or to have a compact object representation, the Angular Radial Transform (ATR)[25] is considered and we named F_{lp}, the Angular Radial Transform on the above Energy map $\boldsymbol{E}(x,y)$, as odd Gabor moment.

$$F_{lp}=\sum_{x=-n/2}^{x=n/2}\sum_{y=-m/2}^{m/2}\mathbf{V}_{lp}(\rho,\theta)\mathbf{E}(x+n/2,y+m/2),$$
$$\mathbf{V}_{lp}(\rho,\theta)=\frac{1}{2\pi}e^{jp\theta}R_l(\rho),\ R_l(\rho)=\begin{cases}1 & l=0\\ 2\cos(\pi\rho) & l\neq 0\end{cases} \qquad (7)$$

where $\rho=\sqrt{x^2+y^2}$ and $\theta=\arctan(y/x)$.

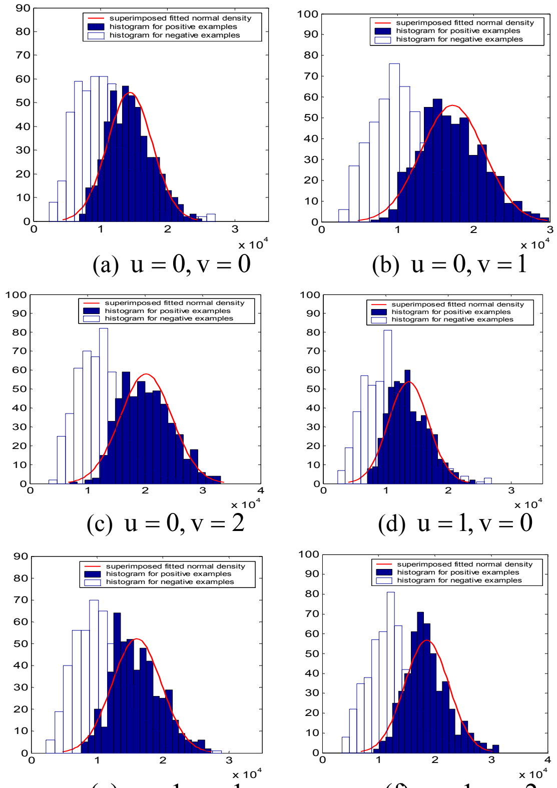
To perform the computing of odd Gabor moments more efficiently, the Angular Radial Transform function is regularized and we have:

$$\mathbf{V}_{lp}^1(\rho,\theta)=\begin{cases}1 & V_{lp}>0\\ 0 & V_{lp}\leq 0\end{cases},\ \mathbf{V}_{lp}^2(\rho,\theta)=\begin{cases}1 & V_{lp}\leq 0\\ 0 & V_{lp}>0\end{cases} \qquad (8)$$

Note that only real part of $\boldsymbol{V}_{lp}$ with $l=(0\sim 2), p=(0\sim 5)$ is adopted in our case. Thus, with the zeros component eliminated, a 35-D odd Gabor moment vector $\boldsymbol{F}=\{F_{lp}\}$ can be obtained for each training example.

4. Modeling multi-channel Gabor responses in corner area

Although the odd Gabor moments introduced in section 3 can effectively portray the structure property of an object to be detected, to have an enhanced object representation some other properties, e.g. texture property, should also be exploited. Since Gabor filter has shown powerful abilities in characterizing texture property, the multi channels even Gabor filters are adopted in our case.



(a) $u=0, v=0$ (b) $u=0, v=1$

(c) $u=0, v=2$ (d) $u=1, v=0$

(e) $u=1, v=1$ (f) $u=1, v=2$

Figure 3 Local texture distribution based on multi-channels even Gabor filters

For consideration of computational cost, only three scales and two orientations are applied for building multi-channel even Gabor filters. Let R_j^e be a jet of multi-channel even Gabor filters based responses with $\theta_u=\frac{u\pi}{2}+\frac{\pi}{4}, j=u+2v, u=0,1, v=0,1,2$. Here we only consider the local texture property extracted from the corner area, and then a local texture feature vector $\bar{f}=\{f_j\}$ can be obtained, where $f_j=\sum_{x,y}[R_j^e(x,y)\cdot T_c(x,y)]$.

To model the distribution of local texture, an assumption is made that $\{f_j\}$ follows a Gaussian distribution with mean $\bar{\mu}_f$ and covariance Σ and we have

$$P(\bar{f})=(2\pi)^{-\frac{d}{2}}|\Sigma|^{-\frac{1}{2}}\exp[-\frac{(\bar{f}-\bar{\mu}_f)\Sigma^{-1}(\bar{f}-\bar{\mu}_f)^T}{2}] \qquad (9)$$

Then a threshold τ_1 can be set on $P(\bar{f})$ for further discrimination. For positive examples as we can see from Fig.3， the real distribution of local texture based on multi-channel even Gabor filters can be ideally fitted by a normal density, which verifies the above assumption quite well. In addition, it is also can be found that the local texture cue provides good discrimination ability between positive car objects and negative non-car ones.

5. Hierarchical SVM classifier based on solution approximating

5.1 Support vector machines

Let $X=\{x_i\in R^d, \mathrm{i}=1,2,\ldots \mathrm{n}\}$ denotes the training set in d dimension space, $y_i=\{-1,1\}$ is the corresponding class label of the example contained in X. As an excellent binary classifier, the nature of SVM is to find a hyper-plane for discrimination that optimally separates two classes of objects by using structural risk minimization. The final discrimination indicator L(x) can be given as [11]:

$$L(x)=\theta[S(x)]=\begin{cases}+1 & if \quad S(x)>0\\ -1 & if \quad S(x)<0\end{cases} \qquad (10)$$

$$S(x)=\sum_{i=1}^{n} y_i\alpha_i\phi(x_i)\cdot\phi(x)+b \qquad (11)$$

where $\theta(\cdot)$ is a indicative function , S(x) denotes the final discrimination output, b is a bias, and $\phi(\cdot)$ is a nonlinear mapping function. In our case the Gaussian kernel: $k(x_i,x)=\phi(x_i)\cdot\phi(x)=\exp(-\|x_i-x\|^2/(2\sigma^2))$ is adopted to perform inner dot between two high dimension feature vectors. Hence, all training examples x_i s corresponding to nonzero α_i s are named by support vectors. Furthermore, as we can see from Eq.(11), $S(x)$ can be taken as a kernel expansion and every nonlinear mapping $\phi(x_i)$ with non-zero α_i can be thought as a basis function in nonlinear feature mapping space $F=\{\phi(x_i)\}_{i=1\ldots n_s}$ with n_s being the number of support vectors. As a mater of fact, the computational cost of discrimination output of $S(x)$ for a given example x depends mainly on the number of nonzero α_i i.e. $\mathrm{n_s}$. For the case of a large number of support vectors, i.e. the case of non-sparse kernel expansion, the approximating solution $S^a(x)$ to $S(x)$ should be considered to reduce the computational cost.

5.2 Solution approximating of Support vector machine

Given a fixed order of approximation $n_a < n_s$, we follow the idea of [13, 14] to form the problem of solution approximating of SVM as a minimization:

$$(\beta, F^a) = \arg\min \left\| \psi - \psi^a \right\|^2 \tag{12}$$

where $\Psi = \sum_{i=1}^{n_S} \gamma_i \phi(x_i)$, $\psi^a = \sum_{j=1}^{n_a} \beta_j \phi(z_j)$, $\beta = \{\beta_j \in R\}_{j=1\ldots n_a}$, $F^a = \{\phi(z_j)\}_{j=1\ldots n_a}$, $\gamma_i = \alpha_i y_i$.

Table 1 the greedy subspace seeking procedure

```
1. Set $F^0 = \Phi$
2. $z_1 = \arg\min_{x_k} \left\| \sum_{i=1}^{n_s} \gamma_i \phi(x_i) - \beta_0 \phi(x_k) \right\|^2$
     where $\beta_0 = [\phi(x_i)^T \cdot \phi(x_i)]^{-1} [\phi(x_i)^T \cdot \Psi]$
       $F^1 = F^0 \oplus \{\phi(z_1)\}$
       $\beta^1 = [\phi(z_1)^T \cdot \phi(z_1)]^{-1} [\phi(z_1)^T \cdot \Psi]$
3. For t=2: $n_a$
   $z_t = argmax_{x_k} \left| \sum_{i=1}^{n_s} \gamma_i k(x_i - x_j) - \sum_{j=1}^{t-1} \beta_j^{t-1} k(z_j, x_k) \right|$
   $F^t = F^{t-1} \oplus \phi(z_t)$
   $\beta^t = [(F^t)^T \cdot F^t]^{-1} [(F^t)^T \cdot \Psi]$
4. $S^a(x) = \sum_{i=1}^{n_a} \beta_i^a k(z_i, x) + b^a$
   where $b^a$ is a new bias that can be computed as [15].
```

Instead of using 'mean shift' technique to obtain the optimal z_i as in [13, 14], a sequential greedy search method that can ensure the convergence of the approximating procedure is applied to seeking a subspace F^a of F for minimizing Eq. (12). The detail implementation of the proposed subspace seeking algorithm is illustrated in Table 1. For obtaining the inverse matrix $[(F^t)^T \cdot (F^t)]^{-1}$ in the step 3, its higher computational cost wouldn't be avoided with the gradually expanded subspace. In our case, the partition method for obtaining inversion matrix is adopted, which can reduce the computational complexity from $O(M^3)$ to $O(M^2)$. Let $K(t+1) = [(F^{t+1})^T (F^{t+1})] = \begin{pmatrix} K(t) & [(F^t)^T \cdot \phi(z_{t+1})] \\ [(F^t)^T \cdot \phi(z_{t+1})]^T & \phi(z_{t+1}) \cdot \phi(z_{t+1}) \end{pmatrix}$, thus we have:

$$K(t+1)^{-1} = \begin{pmatrix} X_{11} & X_{12} \\ X_{22} & X_{22} \end{pmatrix} \tag{13}$$

where $\begin{cases} X_{12} = A_{11}^{-1} A_{12} (A_{12}^T A_{11}^{-1} A_{12} - A_{22}^T)^{-1} \\ X_{11} = A_{11}^{-1} - A_{11}^{-1} A_{12} X_{12} \\ X_{22} = A_{22}^{-1} - A_{22}^{-1} A_{12} X_{12} \end{cases}$ and

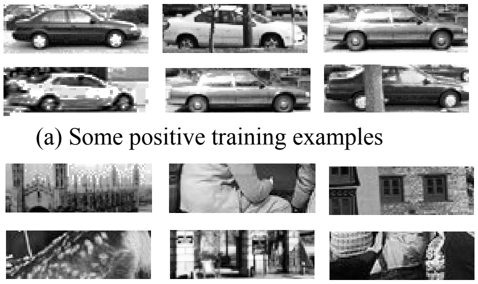
$$\begin{cases} A_{11} = K(t) \\ A_{12} = [(F^t)^T \phi(z_{t+1})] \\ A_{22} = \phi(z_{t+1}) \phi(z_{t+1}) \end{cases}.$$

Given an example that is far away from the discriminate hyper-plane $S(x)$, it is sufficient enough to give its class label only by its approximating solution $S^a(x)$, which can lead to reducing the computational cost quite efficiently. Thus a hierarchical SVM classifier can be constructed with different n_a s.

6. Experimental results and analysis

6.1 Experimental dataset specification

The proposed scheme is evaluated on the car image database [26]. An additional 500 negative examples collected from website are added to the database. Thus the whole training database contains 550 positive and 1000 negative images, and 170 test images containing 200 cars in all. Note that all images in the database are natural images. They are taken from different sources and are roughly subject to possible occlusion， different illumination and cluttered scene. In addition, the size of training images is clipped to 30×90 manually from 40×100 to eliminate the influence from background and all positive examples are adjusted to the same direction. Fig.4 gives some positive and negative training examples.



(a) Some positive training examples

(b) Some negative training examples

Figure 4 Parts of positive and negative training examples

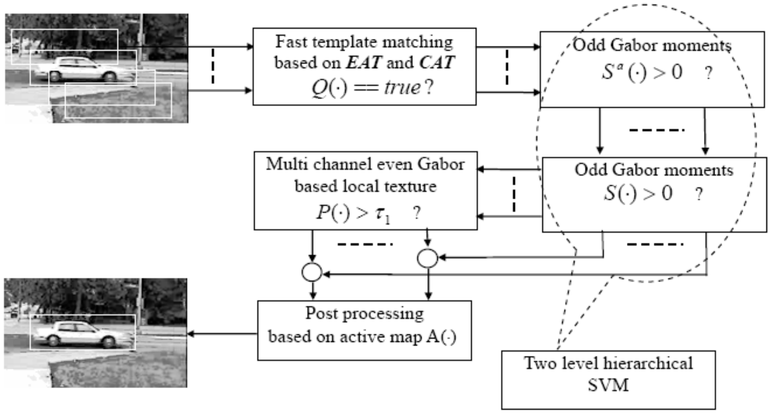


Figure 5 Sequential architecture for car detection based on multi cues integration and hierarchical SVM

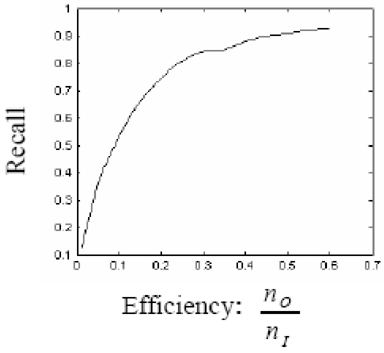


Figure 6 Efficiency of two level hierarchical SVM

6.2 Sequential multi cues based car detection architecture

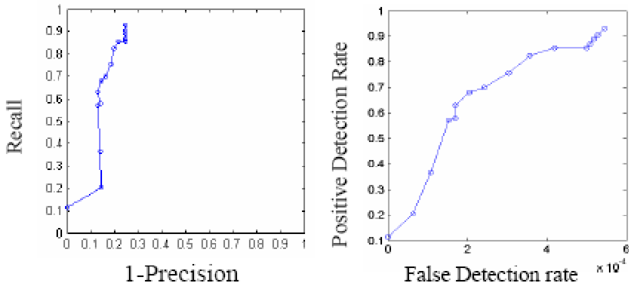
The proposed sequential multi-cue based architecture and hierarchical SVM for efficient car detection is shown in Fig.5. Here the active map $A(\cdot)$ is defined as

$$A(\cdot) = \omega_s S(\cdot) + \omega_p P(\cdot) \tag{14}$$

where ω_s and ω_p are two corresponding weights for $S(\cdot)$ and $P(\cdot)$ respectively, and they are assigned by 0.8 and 0.2 experimentally in our case. For space consideration, the detailed post processing (evaluation scheme on active map) procedure can be found in [2]. Note that to reduce the computational cost the sub-window is moved in steps of 4 pixels and 2 pixels in the horizontal and vertical directions respectively. As we can see only two-level hierarchical SVM with $n_a = 20$, which only takes 6.7% of original 360 support vectors in all among 1550 training examples, is applied in the odd Gabor moment based feature space. To characterize the efficiency of the proposed hierarchical SVM for the task at hand, the ratio of n_O to n_I is considered, where n_I denotes the number of sub-windows passed to the first level SVM and n_O denotes the number of output sub-windows. The relation between positive recall and efficiency of the proposed two-level hierarchical SVM is shown in Fig.6. It is obvious to find that the architecture keeps slightly varied with high recall as ${}^{n_O}\!/_{n_I} > 0.4$, which shows the performing efficiency of the architecture can be improved greatly with slight loss of detection performance.

6.3. Performance evaluation

Fig. 7 illustrates the performance of our car detection system on the test set containing 200 true car objects. The definitions of the related evaluation criterions and the discussions about Recall -Precision and ROC can be found in [2]. The performance comparison of our car detection system with part-based method [2] is shown in Table 2, from which we can see that our detection scheme is greatly superior to S. Agarwal's. In order to make a fair comparison, the No. of correct detection is fixed for both cases. In addition, we also compare our scheme with some other ones in Table 3. With the condition of fixing the No. of false correct detection, although the combination of global (ICA-based) with local (parts-based) information has achieved a good performance with 1% higher than our scheme, it is at the cost of computational efficiency. Note that all the test images are performed on *Matlab* 6.5 platform with P4 2.4 G Hz PC and the average computational cost for processing each test image is no more than 2 seconds, whereas 8 seconds for S.Agarwal's [2] method and more for A. Garg's [4] method. The reason for the low computational cost of our algorithm mainly contributes to the utilization of low level visual cues i.e. Corner area template and Edge area template. Fig. 8 shows an area template matching result for a given input test image. The white pixels shown in right figure denote the candidates of being the centers of car objects for further mining, which only occupy 1.7% of the original input image, i.e. it means most of non-car objects have been efficiently filtered out.



(a) Results: Recall-Precision (b) Results: ROC

Figure 7 Performance evaluation of our car detection scheme on test set with 200 side-view car objects contained

Table 2 performance comparisons between our scheme and parts-based method [2] on the test set with 200 car objects contained

No. of Correct detection N	Recall $\frac{N}{200}$	No. of false	Precision $\frac{N}{N+M}$	False detection rate $\frac{M}{112000}$
181	**90.5%**	**98** **59**	**64.9%** **75.4%**	**0.09%** **0.05268%**
178	**89%**	**92** **55**	**65.9%** **76.4%**	**0.08%** **0.04911%**
171	**85.5%**	**76** **47**	**69.2%** **78.4%**	**0.07%** **0.04196%**
162	**81%**	**48** **39**	**77.1%** **80.6%**	**0.04%** **0.03482%**
154	**77%**	**36** **35**	**81.1%** **81.5%**	**0.03%** **0.03125%**
140	**70%**	**29** **27**	**82.8%** **83.8%**	**0.03%** **0.02411%**

*The table is adopted from [2], and the boxed parts are results based on our scheme. The denominator 112000 showed in last column is the No. of negative sub-windows among 147802 sub-windows evaluated from 170 test images. The data for S.Agarwal's parts based method come from [2]

Table 3 Accuracy comparisons of our car detection scheme with some other ones

Detection schemes	No. of correct detection N	Recall rate $\frac{N}{200}$	No. of false detections M	Precision $\frac{N}{N+M}$
ICA[4]	**184**	**92%**	**139**	**56.97%**
Parts[2]	**181**	**90.5%**	**98**	**64.87%**
ICA+Parts[4]	**188**	**94%**	**61**	**75.50%**
Our scheme	**186**	**93%**	**61**	**75.30%**

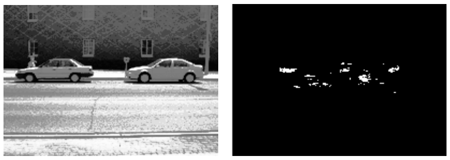


Figure 8 Area template matching result for an input image

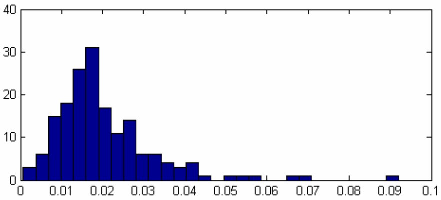


Figure 9 Efficiency of low level visual cues based template matching

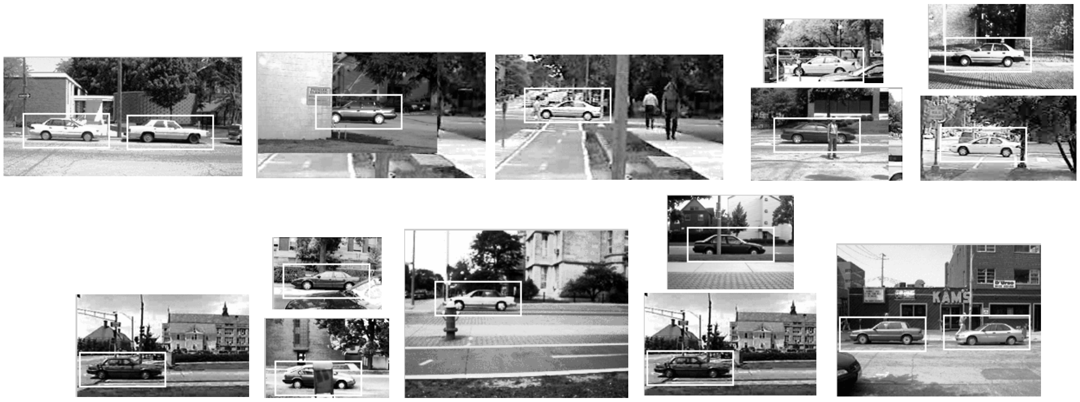


Figure 10 Some examples of correct side-view car detection

Following the architecture shown in Fig.5, let R_{i} denotes the ratio of the number of pixels to pass through the aforementioned low level cues based template matching to the total number of pixels contained in the i th test image. As given in Fig. 9 , most of R_{i} 's are less than 5%; in other word , more than 95% sub windows have been filtered out only based on low level visual cues with low computational complexity. Finally some examples of correct side view car detection are shown in Fig.10, which involves all kinds of variations of occlusion, illumination appearance and so on.

7. Conclusions

In this paper we present a novel sequential architecture for efficient side-view car detection based on multi-cue integration and hierarchical SVM, in which a high detection rate is obtained. The low computational cost contributes mainly to the utilizing of the two proposed low level visual cues based area templates, i.e. Edge Area Template and Corner Area Template and the two level hierarchical SVM. Both global structure cue based on odd Gabor moments and local texture cue based on multi-channel even Gabor filter responses in corner area are considered, which makes a powerful discrimination between car and non-car objects. For our future work we will focus our researches on multi views and multi scales car detection by extending the approach presented in this paper. In addition, the proposed scheme also can be extended to some other highly structured object detection task.

8. Acknowledgements

This work was funded in part by National Natural Science Foundation of China (Nos. 60602030, 90604032), Program for New Century Excellent Talents in University, Specialized Research Foundation of BJTU (No. 2005SZ005), Research Foundation of Beijing Jiaotong University (No. 2005SM013). Open Foundation of National Key Laboratory on Machine Perception.

References

[1] M.Weber, M. Welling, and P. Perona, "Unsupervised Learning of Models for Recognition", In *Proc. Sixth European Conf. Computer Vision*, pp. 18–32, 2000.

[2] S.Agarwal and D. Roth, "Learning a Sparse Representation for Object Detection", In *Proc. Seventh European Conf. Computer Vision*", Vol. 4, pp. 113-130, 2002.

[3] Shivani Agarwal, Aatif Awan, and Dan Roth, "Learning to Detect Objects in Images via a Sparse, Part-Based Representation", *IEEE Trans. on Pattern Analysis and Machine Intelligence*, Vol. 26, No. 11, pp.1475-1490, 2004.

[4] A. Garg, S. Agarwal and Thomas S. Huang, "Fusion of Local and Global Information for Object Detection", In *Proc. IEEE* 16th International Conference on Pattern Recognition, 2002.

[5] D. Liu and T. Chen, "Semantic-Shift for Unsupervised Object Detection", *Workshop on Beyond Patches, in conjunction with CVPR*, 2006.

[6] Bo Wu, Ram Nevatia, "Detection of Multiple, Partially Occluded Humans in a Single Image by Bayesian Combination of Edgelet Part Detectors", In *Proc. 8th International Conf. on Computer Vision*, pp. 90-97, 2005.

[7] H. Schneiderman, T. Kanade, "A Statistical Method for 3d Object Detection Applied to Faces and Cars", In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2000.

[8] C. Papageorgiou and T. Poggio. "A Trainable System for Object Detection", *International Journal of Computer Vision*, Vol.38, No. 1, pp. 15-33, 2000.

[9] P. Viola and M. Jones, "Rapid Object Detection Using a Boosted Cascade of Simple Features", In *Proc. IEEE Conf. Computer Vision and Pattern Recognition*, 2001.

[10] Z. Sun, G. Bebis, and R. Miller, "On-Road Vehicle Detection Using Gabor Filters and Support Vector Machines", *IEEE 14th International Conference on Digital Signal Processing*, 2002.

[11] V.Vapnik, "The Nature of Statistical Learning Theory", Springer, New York, 1995.

[12] E. Osuna, R. Freund, and F. Girosi. "Training Support Vector Machines: an Application to Face Detection", In *Proc. IEEE Conf. of CVPR*, pp.130–136, 1997.

[13] B. Schölkopf,Phil Knirsch,Alex Smola, and Chria Burges, "Fast Approximating of Support Vector Kernel Expansions,and Interpretation of Clustering as Approximation in Feature Spaces" *DAGM Symposium Mustererkennung, Springer Lecture Notes in Computer Science*, 1998.

[14] B. Schölkopf, S. Mika, C. Burges, P. Knirsch, K.-R. Müller,G. Rätsch, and A. Smola," Input Space vs. Feature Space in Kernel-Based Methods", *IEEE Transactions on Neural Networks*, Vol.10,No.5, pp. 1000-1017, 1999.

[15] S. Romdhani, P. Torr, B. Schölkopf, A. Blake, "Computationally Efficient Face Detection", In *Proc. 8th International Conf. on Computer Vision.*, 2001.

[16] D. Comaniciu, V. Ramesh, P. Meer, "Real-Time Tracking of Non-Rigid Objects Using Mean Shift", In *Proc. IEEE Conf. of CVPR*, pp. 142-149,2000.

[17] X. P. Xiao, H. Z. Ai, G.Y. Xu, "Pair-Wise Sequential Reduced Set for Optimization of Support Vector Machines", In *Proc. IEEE International Conference on Pattern Recognition*, pp. 860-863, 2002.

[18] G. Baudat, F. Anouar, "Feature Vector Selection and Projection Using Kernels", *Neurocomputing,* Vol. 55, No.9, pp. 21-38, 2003.

[19] Vojtêch Franc,Václav Hlaváĉ, "Greedy Algorithm for a Training Set Reduction in the Kernel Methods", In *Proc. of Computer Analysis of Images and Patterns*, pp. 426-433 ,2003.

[20] D.P. Huttenlocher, G.A. Klanderman, and W.J.Rucklidge, "Comparing Images Using the Hausdorff Distance", *IEEE Transactions on Pattern Analysis and Machine Intelligence,* Vol. 15, No.9, pp. 850-863, 1993.

[21] C. J. Harris and M. Stephens, "A Combined Corner and Edge Detector", *In Proceedings of the 4th Alvey Vision Conference*, pp. 147-151, 1988.

[22] J.P. Jones and L.A. Palmer,"An Evaluation of the Two-Dimensional Gabor Filter Model of Simple Receptive Fields in Cat Striate Cortex", *Journal of Neurophysiology,* Vol. 58, pp.1233-1258, 1987.

[23] M. Lades, J. Vorbruggen, J. Buhmann, J. Lange, C. Malsburg,"Distortion Invariant Object Rrecognition in the dynamic Link Architecture", *IEEE Transactions on Computers*, Vol.42,No.3, pp.300–311, 1993.

[24] R.Mehrotra, K.R.Namuduri, and N.Ranganathan, "Odd Gabor Filter-Based Edge Detection", *Pattern Recognition,* Vol. 25, No.12, pp.1479-1494, 1992.

[25] J. Fang, G. Qiu, "Face Detection Based on Multiple Regression and Recognition Support Vector Machines", *British Machine Vision Conference*, Norwich, Sept. 2003.

[26] http://l2r.cs.uiuc.edu/~cogcomp/Data/Car/