Detector Ensemble

Shengyang Dai Ming Yang Ying Wu Aggelos Katsaggelos
EECS Department, Northwestern University, Evanston, IL 60208, U.S.A.
{sda690,mya671,yingwu,aggk}@ece.northwestern.edu

Abstract

Component-based detection methods have demonstrated their promise by integrating a set of part-detectors to deal with large appearance variations of the target. However, an essential and critical issue, i.e.*, how to handle the imperfectness of part-detectors in the integration, is not well addressed in the literature. This paper proposes a* detector ensemble *model that consists of a set of substructure-detectors, each of which is composed of several part-detectors. Two important issues are studied both in theory and in practice, (1) finding an optimal detector ensemble, and (2) detecting targets based on an ensemble. Based on some theoretical analysis, a new model selection strategy is proposed to learn an optimal detector ensemble that has a minimum number of false positives and satisfies the design requirement on the capacity of tolerating missing parts. In addition, this paper also links ensemble-based detection to the inference in Markov random field, and shows that the target detection can be done by a max-product belief propagation algorithm.*

1. Introduction

The success of learning-based methods for face detection motivates the exploration of these methods for general targets. However, it is found that these methods are often limited when it is impractical to include sufficient representative training data to cover the large uncertainty of the target's visual appearances. Such challenging situations are not uncommon in practice, and occur when the target experiences large deformations, or when the target is partially occluded (some examples are shown in Fig. 1). In these situations, since the variability in appearance changes is enormous due to, for example, virtually unlimited possibilities of partial occlusion, it is generally very difficult, if not infeasible, to train such holistic detectors.

This has motivated the investigation of component-based detection [1, 2, 3, 4, 7, 9, 10, 12, 13, 15, 17, 18, 19, 20, 25, 26], where the detection of the entire target is through the integration of the detection of its parts by matching isomorphic graphs. This approach may provide a powerful means to overcome the limitation of the holistic object detectors mentioned above, because the variability of the appearance of the entity is decomposed into the local variability of the appearances of its parts and the variability of the structure among these parts. Thus, the combination of multiple part-detectors may ease the learning requirement based on an impractical amount of training data.

Figure 1. Appearance variations.

However, an essential and critical issue, *i.e.*, how to handle the imperfectness of part-detectors in the integration step, is not well addressed in the literature. Although the training of part-detectors is more likely to converge, part-detectors generally tend to produce larger number of false positives, because the parts tend to be less discriminative. In addition, parts may be missing due to occlusions and/or the imperfectness of part-detectors. These situations have challenged many existing component-based detectors. To handle the false positives, geometrical constraints among the parts usually need to be considered. But the missing parts bring difficulties. If the constraints are defined on all (or a large portion) of the parts, a small number of missing parts (or even a single one) will make these constraints invalid. Thus detectors based on such constraints are not likely to have good properties in tolerating missing parts.

This has motivated our idea of using *substructures* to distribute the geometrical constraints. It tends to be more robust because a small number of missing parts is less likely to invalidate all the substructures. In this paper, we propose a *detector ensemble* model that consists of a set of substructure-detectors, each of them consisting of several

part-detectors. A substructure-detector provides a positive detection only when all the related part-detectors are positive and the constraints among those parts are satisfied. This integration rule tends to reduce the false positives in detection. Moreover, the ensemble provides a positive detection when at least one substructure-detector is positive. This rule tends to increase the probability of detecting the target.

This detector ensemble should not be arbitrarily composed, because different compositions of the ensemble will certainly have different capacities of tolerating missing parts and have different detection performances. Therefore, we ask the following interesting question: given a tolerance capacity, does there exist an optimal detector ensemble that has the minimum false positive rate?

In this paper, we analyze the capacity of tolerating missing parts based on the concept of a covering set in graph theory, and obtain an analytical form that characterizes the tolerance capacity. Combining this tolerance capacity and our analysis on the detection rates and false positive measure of the substructures and the ensemble, we formulate a novel constrained optimization problem, whose solution gives the optimal ensemble. Based on that, a new model selection strategy is proposed. In addition, we show the relation between the ensemble-detector and the inference of Markov random field, thus the detection task can be done by the max-product belief propagation algorithm efficiently.

The proposed detector ensemble approach has several advantages. First of all, its training is relatively simple. It does not require a huge training set to cover virtually unlimited cases of partial occlusion. In addition, as demonstrated in our experiments, even if the part-detectors are not fine-tuned and only trained on very limited cases, the entire ensemble can still achieve a better performance than finely-tuned holistic detectors on many difficult cases including partial occlusions. More importantly, the proposed method has a theoretical guarantee for the detection rate.

2. Related work

Various algorithms have been proposed to integrate component detection results. Global methods try to characterize the relationship among all parts simultaneously. The model can be deterministic by placing the component detectors at specific locations [15, 18], or probabilistic, such as modeling the relative position among parts as a joint Gaussian distribution [5, 12]. In [1], the spatial relations of the parts are encoded as a high-dimensional feature vector to train a sparse network of Winnows. Mutual occlusion is investigated in [25] under a Bayesian framework.

To decompose the model variety, and achieve higher tolerance for missing part, distributed methods have been extensively studied recently, which model the consistency of one part subset at each time, and describe the entire object by choosing a number of such part subsets. For part subset consistency, pairwise model is widely used due to its simplicity. The relative position of two parts is usually modeled as a Gaussian distribution [6, 10, 17]. To achieve rotation invariant, triple [13, 20] or higher order [2] relation is also explored. To compromise between efficiency and accuracy, various subset selection strategies are proposed, mainly based on methods from graph theory. The entire graph topology can be constructed by a mixture tree model [10], acyclic graph [4], star model [6], k-fan model [2], minimal spanning tree [26], or triangulation method [13, 17]. A compositional model [16] is another distributed method motivated by the study of human vision, where intermediate abstraction of images is learnt in a generative manner for object category recognition. Although the existing methods can achieve the missing part tolerance ability to some extend, as far as we are aware, no quantitative analysis of the relationship between model selection and detection performance has been constructed in the literature.

3. Detector ensemble

3.1. Model overview

The overview of the proposed *detector ensemble* model is shown in Fig. 2. Suppose an object O contains n parts $\{p_1, p_2, \ldots, p_n\}$. Each part p_i is associated with a *part-detector* whose detection rate is d_i^p and false positive rate f_i^p. The part-detectors could be any existing detectors, and they are generally far from perfect. A lot of false positives may be present since a part is less discriminative than the entire object. In addition, parts are likely to be missing due to appearance variations and occlusions. Here we do not differentiate between these two cases, and use d_i^p to characterize both of them.

A *substructure-detector* is composed of a set of part-detectors and the constraints among those parts. A substructure-detector is positive only when all its part-detectors are positive and the constraints are satisfied. One key feature of the substructure-detectors is that they do not allow missing parts. In other words, all related parts must be detected in order to form a valid detection of the substructure. The detection rate and false positive measure of a substructure S_i are denoted by d_i^s and f_i^s, respectively.

The *detector ensemble* consists of a set of substructure-detectors. It gives a positive detection if and only if at least one substructure is detected positively, and none of them has a negative response. We denote by $d^{\mathcal{E}}$ and $f^{\mathcal{E}}$ its detection rate and false positive measure, respectively.

3.2. Substructure detectors

Assume that there are k_i detections for part p_i, which we call *part candidates*. Each object candidate can be described by a labeling state $L = \{l_1, l_2, \ldots, l_n\}$, where $l_i \in \{0, 1, \ldots, k_i\}$ indicates the selected candidate for part p_i, and $l_i = 0$ if p_i is missing.

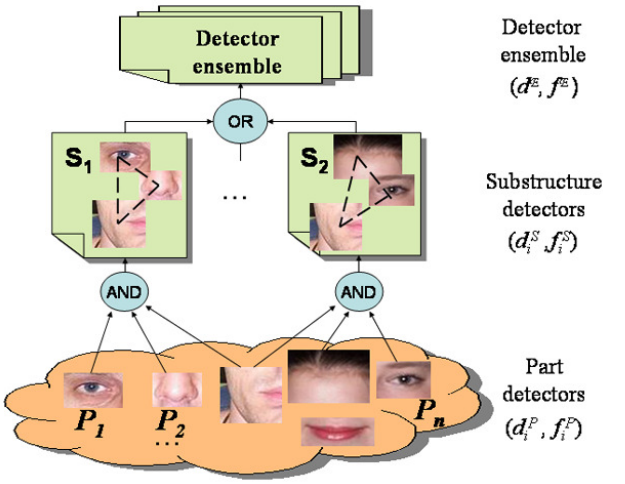


Figure 2. Architecture of a *detector ensemble*.

Each substructure $S_j = \{P_j, h_j(L_j)\}$ has two components, $P_j \subseteq O$ contains all the related parts, L_j is the projection of labeling state L on P_j, and $h_j(L_j)$ is a decision function that determines how well the candidate set L_j fits the training data. The decision function is as follows:

$$h_j(L_j) = \begin{cases} 0 & \exists\, i, p_i \in P_j, l_i = 0 \\ \log \frac{H_j(L|1)}{H_j(L|0)} - \lambda_j & \text{otherwise,} \end{cases} \quad (1)$$

where H_j is the likelihood term derived from the training data, and the log likelihood ratio $\log \frac{H_j(L|1)}{H_j(L|0)}$ is compared with a pre-defined threshold λ_j.

A labeling state L is *valid* for substructure S_j if and only if $h_j(L_j)$ is positive. A valid substructure S_j must satisfy two conditions: (1) no part in S_j is missing, and (2) the log likelihood ratio must be larger than λ_j to enforce the part consistency. For a substructure-detector, we denote by d_j^S the detection rate of S_j (the probability that the set of parts from a real object is detected as valid substructure by this detector, given that no related part is missing) and by f_j^S the false positive measure (the false positive number on negative training set). The benefit of using substructure-detectors is that its training is quite easy since no missing part need to be considered.

It is critical to chose an appropriate substructure pool. If a substructure contains too many parts, the chance to detect such a substructure is low. On the other hand, if it has very few parts, the corresponding decision function can be easily satisfied and thus tends to induce many false positives. We choose part triples as substructure candidates, because they can achieve a trade-off between modeling efficiency and false positive control. In addition, this is also consistent with our intuition: if three parts are detected, then it is likely that there exists a real object. Even more so, the use of part triples can achieve rotation invariance during integration, which is not achievable in pairwise modeling.

In our experiments, the positive substructure likelihood is defined as a combination of the inner angle distribution (2D distribution), the pair-wise scale ratios (three of them), and the pair-wise scale to distance ratios (three of them). Only the inner angle distribution is considered for negative likelihood. λ_j is chosen to guarantee a high detection rate ($\geq 99\%$) in training.

3.3. Ensemble detector

Although each substructure requires that all related parts must be present, the cooperation of multiple substructures will enable the entire detector to be robust to missing parts. The *ensemble-detector* $\mathcal{E}$ consists of a set of substructure-detectors. The decision for a labeling state L can be made by fusing the decisions of all substructures. It is positive if and only if: (1) there exists at least one valid substructure in the labeling state L, and (2) no substructure gives negative decision on L.

How should substructures be chosen to compose the ensemble? Does there exist an optimal one? To answer these questions, we need to analyze the detection rate and false positive number of the ensemble based on the fusion rule.

Detection rate: Including more substructure-detectors in the ensemble tends to raise the detection rate of the ensemble in general, because more occlusion cases can be covered. The quantitative analysis is given as follows.

It is reasonable to assume that the missing parts are independent of each other, since part-detectors work independently. Then the detection rate of the ensemble $\mathcal{E}$ is

$$d^{\mathcal{E}} = \sum_{D \subseteq O} \big(p(D)d(D)\big), \quad (2)$$

where D is a detection event indicating which parts are present, the sum is over all possible D (2^n cases),

$$p(D) = \prod_{p_i \in D} d_i^p \prod_{p_i \notin D} (1 - d_i^p) \quad (3)$$

is the probability that the event D happens, and $d(D)$ is the probability that one of the substructures is actually valid given the detection event D. If each substructure-detector reaches decisions independently, we have

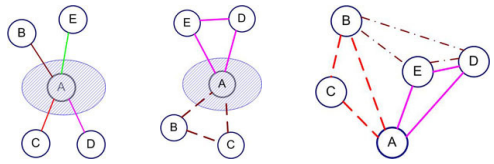
$$d(D) = 1 - \prod_{S_j \in \mathcal{E}, P_j \subseteq D} (1 - d_j^S). \quad (4)$$

Without the independency assumption, a conservative estimation gives:

$$d(D) \geq 1 - \min_{S_j \in \mathcal{E}, P_j \subseteq D} (1 - d_j^S) \quad (5)$$

$$= \begin{cases} \max_{S_j \in \mathcal{E}, P_j \subseteq D} d_j^S & \exists \text{ such } j \\ 0 & \text{otherwise.} \end{cases} \quad (6)$$

Furthermore, if we assume that all part-detectors have roughly the same detection rate, or $d_i^p = d^p$, and all



(a) (b) (c)

Figure 3. Illustration of the idea of covering set.

substructure-detectors have roughly the same detection rate $d_j^S = d^S$ (since they are all close to 1), the detection rate for the ensemble can be further simplified as follows:

$$d^{\mathcal{E}} = \sum_{D \subseteq O} \left((d^p)^{|D|} (1-d^p)^{n-|D|} (1-(1-d^S)^{|\{j|P_j \subseteq D\}|}) \right). \tag{7}$$

False positive: Since each substructure-detector alone can make positive decisions for the entire detector ensemble, the total number of false positives is approximately equal to the summation of the false positive numbers from all substructure-detectors. Although all substructure-detectors have similar detection rates, their false positive numbers can be very different. Substructure-detectors in which the parts present large shape deformation of relative position tends to have a high false positive measure, as the requirement for making positive decision has to be relatively loose to accommodate deformation.

To summarize, an ensemble with more substructure-detectors tends to have a higher detection rate, as well as a larger number of false positives. Therefore, there must exist a trade-off that gives the best ensemble.

4. Learning the optimal ensemble

Selecting a subset from all possible triple substructures has a combinatorial complexity $O(2^{n^3})$. This poses a great difficulty in finding the optimal solution. Before we can learn the optimal ensemble, we need to characterize its capacity of tolerating missing parts. To answer this question, we introduce the concept of the covering set that is related to the graph topology and can be used to characterize the tolerance capacity.

4.1. Covering set

Inappropriate choice of the ensemble $\mathcal{E}$ will result in low detection rate due to vulnerability to missing parts. Fig. 3 shows a simple example to explain this phenomenon. In Fig. 3(a), $\mathcal{E}$, containing four substructures (each of which is a part-pair), is used to detect a five-part object. If part A is missing, none of the substructures will be valid and the entire object cannot be appropriately detected. This is not "fair" since the miss detection of the target is only because of the missing of one particular part. In this case, the detection rate of the ensemble is no more than that of part A, which can be very small if A is a frequently missing part. Similarly, if we use triple-substructures as shown in Fig. 3(b), two triples cannot cover all the "unfair" cases where one missing part (A) will lead to the missing detection of the entire target. But once more substructures are involved, these "unfair" cases are less likely to happen. Fig. 3(c) shows an ensemble that has three triple-substructures, and this ensemble has a better capacity of tolerating missing parts. No matter which part is missing, there always exists at least one valid substructure. As long as such remaining substructures are detected, the entire target can be detected. This implies a fairly high detection rate. Besides, the number of substructures in this covering set is much smaller than the total number of all possible triple-substructures ($\binom{5}{3} = 10$). This example explains that the capacity of tolerating missing parts and detection rate of the ensemble are closely related to some set covering property. Next, we formally introduce the concept of covering set, and show that a high detection rate can be guaranteed by this covering set property.

Definition 1 *Given a set O containing n elements, denoted by $C_m = \{c \,|\, c \subseteq O, |\, c\,| = m\}$ the set containing all subsets of N with cardinality m. We call a set $S \subseteq C_m$ a (t, m)* cover *of set O if and only if for any $c \in C_t$, there is at least one $s \in S$, satisfying $s \subset c$, where C_t is the set containing all subset of O with cardinality t ($t \geq m$). A* minimum (t, m) covering set *is the covering set which has no other (t, m) covering set as its subset.*

In other words, a (t, m) covering set is a set of m element subsets, such that for any t elements, there must exists at least one subset in this covering set, whose elements are all included by those t elements. If we select the ensemble as a (t, m) covering set of the part set O, then for a real object, and if t parts are correctly detected, there must exist at least one substructure, whose corresponding parts are all detected. Thus the entire object will have a good chance to be detected, and a high detection rate can be guaranteed.

This definition is closely related to the concept of *Turán* number in graph theory [8], which is defined as

$$T(n, t, m) = \min\{\, |\, c\,| \,|\, c \text{ is a } (t, m) \text{ cover of } O\}. \tag{8}$$

It is the minimum size of a (t, m) covering an n element set ($n \geq t \geq m$). Some *Turán* numbers for small n are shown in Table.1 [8]. When $t = 3$, $T(n, 3, 3) = \binom{n}{3}$, all triples should appear in the covering set. Fig. 3(c) is an example of a $(4, 3)$ covering of a 5 element set. Although the *Turán* number is not achievable for all the minimum covering sets, those with the same covering property usually have roughly the same cardinality as confirmed experimentally.

For object detection, we define the missing tolerance number $\mathcal{K}(\mathcal{E})$ for a given ensemble $\mathcal{E}$ as follows:

$$\mathcal{K}(\mathcal{E}) = \max\{\, p \,|\, \mathcal{E} \text{ is a } (n-p, 3) \text{ cover of } O\}. \tag{9}$$

It means that even if $\mathcal{K}(\mathcal{E})$ parts are missing, there is still at least one substructure in the ensemble $\mathcal{E}$ such that all its parts are detected.

Table 1. $T(n,t,3)$ and upper bound of missing rate ($\bar{m}\%$) estimated by Eqn. 10 when $d^p = 0.8$, $d^S = 0.95$. The values are shown in the format $T_{\bar{m}}$, and * marks the value higher than $1-d^p$.

$t\backslash n$	5	6	7	8	9	10
3	$10_{10.5}$	$20_{6.6}$	$35_{5.4}$	$56_{5.1}$	$84_{5.0}$	$120_{5.0}$
4	$3_{30.0^*}$	$6_{14.4}$	$12_{8.2}$	$20_{6.0}$	$30_{5.3}$	$45_{5.1}$
5	$1_{68.9^*}$	$2_{37.7^*}$	$5_{19.1}$	$8_{10.4}$	$12_{6.9}$	$20_{5.6}$
6	-	$1_{75.1^*}$	$2_{45.2^*}$	$4_{24.3^*}$	$7_{13.1}$	$10_{8.1}$

For any given ensemble $\mathcal{E}$, based on Eqn. 7 and Eqn. 6, the lower bound of the detection rate of the entire object can be further estimated as follows:

$$\begin{aligned} d^{\mathcal{E}} &\geq \sum_{|D|\geq q} \left((d^p)^{|D|}(1-d^p)^{n-|D|}(1-(1-d^S)) \right) \\ &= d^S \sum_{k=q}^{n} \binom{n}{k} (d^p)^k (1-d^p)^{n-k}, \end{aligned} \tag{10}$$

where $q = n - \mathcal{K}(\mathcal{E})$. Thus a larger $\mathcal{K}(\mathcal{E})$ will result in a larger tail of a binomial distribution. As a result, a higher detection rate of the entire target can be guaranteed. The above analysis is only a conservative estimation, since in the first step, all the terms correspond to $|D| < t$ are dropped. Some missing rates estimation with given parameters are listed in Table. 1. It shows that by integrating the part-detectors, the detection rate of the object can be much higher than that of each part.

4.2. Model selection strategy

There is a trade-off between detection rate and false positive number. In general, more substructures in the ensemble $\mathcal{E}$ will result in a higher detection rate and a larger number of false positives. Given the missing tolerance ability requirement t, *i.e.*, to cover or tolerate all cases at most t missing parts, the optimal ensemble $\mathcal{E}_{opt}(t)$ is defined as follows:

$$\begin{aligned} \mathcal{E}_{opt}(t) &= \arg\min_{\mathcal{E}} \{ \sum_{S_j \in \mathcal{E}} f_j^S \}, \\ &s.t. \quad \mathcal{K}(\mathcal{E}) \geq n-t, \mathcal{E} \subset T, \end{aligned} \tag{11}$$

where $T = \{S_1, S_2, ..., S_m\}$ is the set of all substructure candidates ($m = \binom{n}{3}$ for triple substructures), and f_j^S is the false positive measure of S_j defined in Sec. 3.2. The optimal ensemble has the minimum number of false positives while satisfies a required tolerance ability of missing parts.

Finding $\mathcal{E}_{opt}(t)$ is a combinatorial optimization problem. A randomized strategy is used for searching. The ensemble $\mathcal{E}$ is initialized to T. In each step, one substructure is removed if the covering property can be maintained. It is randomly selected according to the false positive measure. Those with a larger false positive measure have a larger probability to be removed. This is done until a minimum covering set is obtained. The entire process runs for N ($= 1000$) times, and the ensemble that gives the least value of the objective function in Eq. 11 is selected.

Input Part number n, substructure candidate set $T = \{S_1, S_2, ..., S_m\}$, false positive measure f_j^S for S_j, missing tolerance number t ($n \geq t \geq 3$).
Output Optimal ensemble $\mathcal{E}_{opt}(t) \subseteq T$.

1. For $i = 1$ to N
   (a) Initialize $\mathcal{E}_i = T$.
   (b) While $\mathcal{E}_i$ is not a minimum $(t,3)$ covering set
      i. Find all removable substructures $\{S_{i_1}, S_{i_2}, \cdots, S_{i_k}\}$.
      ii. Choose one of them with probability $\alpha^{f_{i_t}^S}/z$, with z a normalization factor and α a constant. Assume S_{i_0} is chosen.
      iii. $\mathcal{E}_i = \mathcal{E}_i \backslash S_{i_0}$.
2. $\mathcal{E}_{opt}(t) = \arg\min_{\mathcal{E}_i} \{ \sum_{S_j \in \mathcal{E}_i} f_j^S | 1 \leq i \leq N \}$.

Figure 4. Learning the optimal model $\mathcal{E}$.

Given the model complexity (*i.e.*, the number of substructures), there are some other model selection strategies, such as the triangulation-based model, star model [6], k-fan structure [2], or random selection. It is easy to see that if some parts are shared by many substructures in the ensemble, the risk of detection failure can be high if these common parts are frequently missing parts. In view of this, the triangulation-based model, star model, and k-fan model are less preferable. The random selection strategy might be better than these methods, due to the decentralization of the substructures, although this is not guaranteed. The concept of covering set constructs an explicitly connection between the topology of substructures and the detection rate of the ensemble. As shown in our experiments, the model selected by the proposed algorithm outperforms others.

5. Ensemble-based object detection

Detecting the target from the part candidates is a combinatorial optimization problem. In this section, we connect this task with the inference of a Markov random field which can be solved efficiently with belief-propagation.

Define the energy of a labeling state L w.r.t. the given ensemble $\mathcal{E}$ as follows

$$E(L) = \sum_{p_i} \phi_i(l_i) - \sum_{S_j \in \mathcal{E}} h_j(L_j), \tag{12}$$

where $\phi_i(l_i)$ is the likelihood computed by the i-th part-detector, which can be ignored given the binary decision. The second term measures the consistency with each substructure. The MAP labeling state is the one with minimal energy, which corresponds to the most likely object in the input image.

Positive detection decision is made if and only if the MAP labeling state L satisfies $E(L) < \lambda$, where λ is a pre-determined energy threshold. We have the following remark which can relate the above MAP decision strategy with our ensemble-detector.

Remark 1 *The MAP labeling state L satisfies $E(L) > 0$, if and only if there exists a valid substructure in L.*

Proof: if no valid substructure exists, then all $h_j(L_j)$ will be non-positive ($\forall S_j \in \mathcal{E}$), thus $E(L) \leq 0$, for all possible L, including the MAP one. On the other hand, if there is a valid substructure, then assume that L_0 is the labeling state that contains parts that are only present in this valid substructure and all the other parts are missing. Then the energy of L_0 will be equal to the h value returned by this valid substructure, which is positive. So the MAP labeling state must also have positive energy. □

The searching, therefore, of valid substructures can be transformed to the searching of the MAP labeling state which is the inference problem of a Markov random field defined by Eq. 12. We employ the max-product belief propagation algorithm [11, 24] to find the MAP labeling state on a factor graph, where each variable node corresponds to one part label, and each function node corresponds to a substructure. If and only a part is included in a substructure, the corresponding variable and function nodes are connected.

Assume that each part candidate can only belong to at most one object, then multiple objects can be detected sequentially. In each iteration, the most likely object is detected. Those part candidates overlapped with previous detections are removed before detecting the next one.

6. Experiments

The proposed approach is tested on frontal face detection for comparison with the state-of-the-art detection methods [14, 23], and further on car rear detection to demonstrate its general applicability. Part-detectors are all trained by AdaBoost [14].

6.1. Experiments on frontal face detection

For face, the training set contains 995 frontal upright faces collected from the Caltech-101 Object Categories [5], the GTAV Face Database [21], and the Internet. Eight parts (forehead, left/right eye, nose bridge, nose tip, left/right check, lips) are manually labeled. The first image in Fig. 7(a) shows an example of these parts. The baseline detector for the entire face and the part-detectors for the ensemble are all AdaBoost detectors with extended Haar-like feature [14]. Each part-detector has 16 stages, and the training time is about 10 hours. Some results for part detection are shown in Fig. 5. The testing set is composed of 561 images (225 images from GTAV with appearance variations, 39 images taken by ourselves in office environments, and other images collected from the Internet). Our ground truth data contain 686 faces that exhibit appearance variations to different extents. This set only contains faces with height larger than 100 pixels, since we do not expect part-based detectors to work on low-resolution objects.

The learned optimal ensembles for $(4, 3)$ and $(5, 3)$ covering contain 22 and 10 substructures respectively. Their corresponding ROC curves are shown in Fig. 6. They are compared with $(3, 3)$ covering ensemble (all 56 triples included), AdaBoost.1 (AdaBoost face detector trained by [14]) and AdaBoost.2 (AdaBoost detector trained on the same set of training data). The ROC curves are obtained by changing the operating point of part-detectors.

The proposed algorithm shows a better performance than a finely-tuned AdaBoost detector (AdaBoost.1). The improvement is basically from faces with large appearance variations or severe occlusion. Since such extreme cases only take a small portion of the entire testing set, the improvement over the entire testing set is not significant. For those relatively easy cases, similar results can be achieved. The improvement over the holistic face AdaBoost detector trained on the same data set (AdaBoost.2) is substantial, since our method can work very well on cases with appearance change and occlusion, which are not included in the limited number of training data. Some example results are shown in Fig. 7. Various appearance changes due to occlusion, illumination, rotation are included. Since the part-detectors are only trained with up-right samples, some faces with large in-plane rotation (such as in the third image of Fig. 7(c)) are not detected due to failure of part-detectors, however, the proposed ensemble framework itself is rotation invariant given rotation variant part-detectors.

Compared to the $(3, 3)$ covering ensemble, the performance of the learned optimal $(4, 3)$ covering ensemble and optimal $(5, 3)$ covering ensemble degrades slightly, with a much simpler ensemble structure. Given the same part detection result, $(4, 3)$ and $(5, 3)$ covering ensemble can greatly suppress the false alarm number without much drop of the detection rate.

Figure 6(b) and (c) show the comparison of our model selection strategy with others. The performance of 7 randomly generated ensembles with the same number of substructures as in the optimal $(4, 3)$ and $(5, 3)$ covering ensemble are shown in (b) and (c) respectively. Besides random selection strategy, (c) also shows the performance of a triangulation-based substructure selection strategy (this ensemble contains 8 substructures from Delaunay triangulation, and two more other randomly selected substructures, to make the total number of substructures the same as the optimal $(5, 3)$ covering ensemble), which is very low. The optimal ensemble found by our algorithm outperforms others with a considerable gain in detection rate and a relatively small false alarm number.

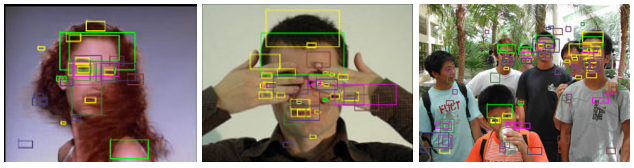


Figure 5. Sample detection results of the parts of the face.

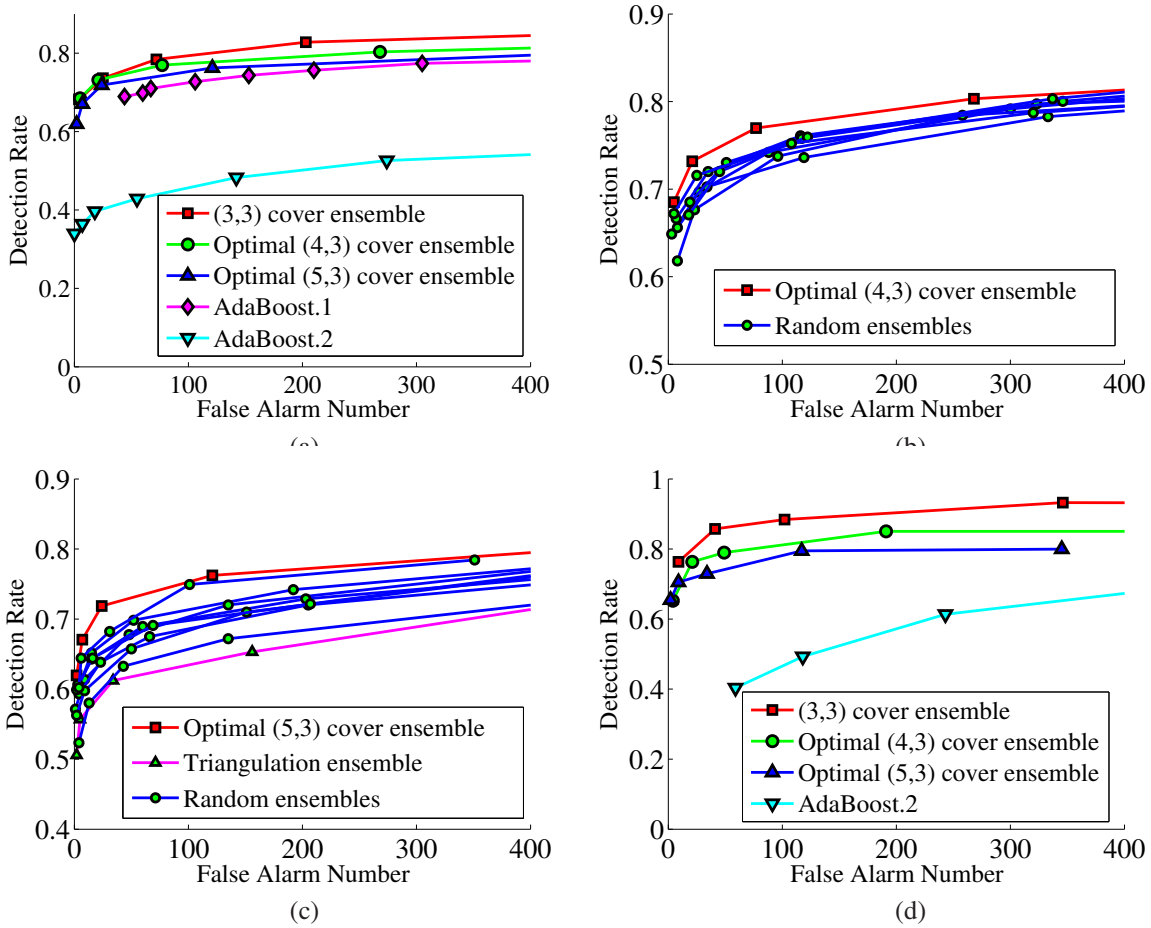
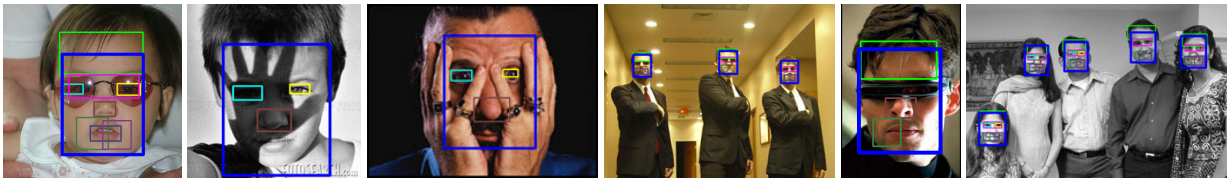
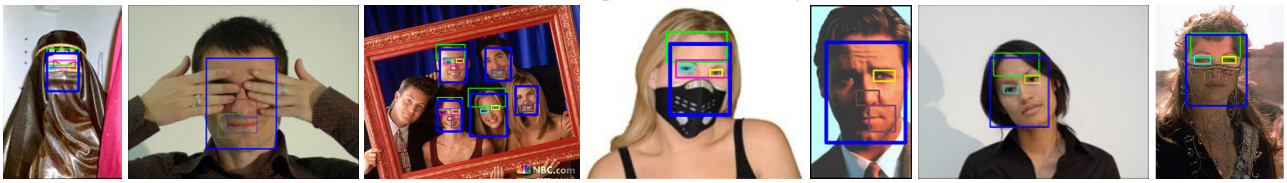


Figure 6. Performance comparison for (a) face detection, (d) car rear detection, and (b) comparison between optimal $(4, 3)$ covering ensemble with random ensemble, (c) comparison between optimal $(5, 3)$ covering and random generated or triangulation-based model.



(a) Some results with (3,3) covering ensemble. The first image also illustrate the part selection.

(b) Some results with optimal (4,3) covering ensemble.



(c) Some results with optimal (5,3) covering ensemble.

Figure 7. Face detection results and corresponding parts.

6.2. Experiments on car rear detection

For car rear detection, seven parts are selected (plate, left/right light, left/right base, left/right window). There are 652 images from the Caltech image database. We randomly select 225 and label them as training data, and use 336 of them as testing data (some very similar images looking like successive frames from a video are removed). We label those whose widths are greater than 50 pixels as the ground truth, and there are 414 ground truth cars in total. The above ground truth selection makes the task more challenging than the traditional object recognition task, since those cars not in the center of each image exhibit large appearance variations due to occlusion, illumination changes, and car direction changes. The ROC curves and some sample results are shown in Fig. 6(d) and Fig. 8 respectively. The proposed algorithm has an excellent generalization ability.

6.3. Complexity

On a PentiumIV 3.4 PC, without code optimization, the speed is evaluated on 266 car rear images (360×240), since they have the same image size. On average, there are 7.57 candidates for each part. Part detection takes 1.59 seconds. The average time for detector ensemble is 1.76 seconds for $(3, 3)$ covering ensemble (35 substructures), 0.58 seconds for optimal $(4, 3)$ covering ensemble (13 substructures), 0.17 seconds for optimal $(5, 3)$ covering ensemble (5 substructures). The complexity is approximately linear with the number of substructures in the ensemble. In fact, with the idea of sharing feature [22], the complexity for detecting multiple parts could potentially be largely reduced.

7. Conclusion

In this paper, to detect objects under severe occlusion, the detector ensemble model is introduced as a set of cooperative substructure-detectors. Our theoretical analysis provides the condition that guarantees the tolerance ability of missing parts, based on which, the optimality of the detector ensemble is studied, and a randomized search algorithm is designed to find the optimal ensemble. Encouraging results demonstrate the merits and advantages of the ensemble-detector that only uses very limited training data.

Acknowledgments

The authors would like to thank Ting Yu and Gang Hua for very helpful discussion, and Xiaoming Sun for pointing out the literature in graph theory. This work was supported in part by National Science Foundation Grants IIS-0347877 and IIS-0308222, and the Motorola center for seamless communications.

References

[1] S. Agarwal, A. Awan, and D. Roth. Learning to detect objects in images via a sparse, part-based representation. *IEEE Trans. on PAMI*, 26(11):1475 – 1490, 2004. 1, 2

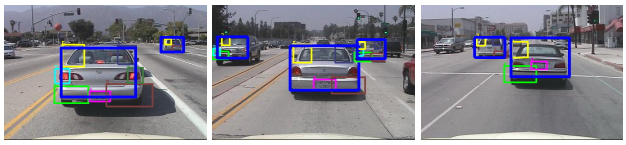


Figure 8. Car rear detection results with optimal $(4, 3)$ covering ensemble and corresponding parts.

[2] D. Crandall, P. Felzenszwalb, and D. Huttenlocher. Spatial priors for part-based recognition using statistical models. In *CVPR*, 2005. 1, 2, 5

[3] B. Epshtein and S. Ullman. Identifying semantically equivalent object fragments. In *CVPR*, 2005. 1

[4] P. F. Felzenszwalb and D. P. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 61(1):55–79, 2005. 1, 2

[5] R. Fergus, P. Perona, and A. Zisserman. Object class recognition by unsupervised scale-invariant learning. In *CVPR*, 2003. 2, 6

[6] R. Fergus, P. Perona, and A. Zisserman. A sparse object category model for efficient learning and exhaustive recognition. In *CVPR*, 2005. 2, 5

[7] S. Gold and A. Rangarajan. Graduated assignment graph matching. In *IEEE Intl. Conf. on Neural Networks*, 1996. 1

[8] D. M. Gordon, O. Patashnik, and G. Kuperberg. New constructions for covering designs. *Journal of Combinatorial Designs*, 3(4):269–284, 1995. 4

[9] B. Heisele, T. Serre, M. Pontil, and T. Poggio. Component-based face detection. In *CVPR*, 2001. 1

[10] S. Ioffe and D. A. Forsyth. Mixtures of trees for object recognition. In *CVPR*, 2001. 1, 2

[11] F. R. Kschischang, B. J. Frey, and H.-A. Loeliger. Factor graphs and the sum-product algorithm. *IEEE Trans. on Infomation Theory*, 47(2):498–519, 2001. 6

[12] T. Leung, M. Burl, and P. Perona. Finding faces in cluttered scenes using random labled graph matching. In *ICCV*, 1995. 1, 2

[13] Y. Li, Y. Tsin, Y. Genc, and T. Kanade. Object detection using 2d spatial ordering constraints. In *CVPR*, 2005. 1, 2

[14] R. Lienhart and J. Maydt. An extended set of haarlike features for rapid object detection. In *ICIP*, 2002. 6

[15] A. Mohan, C. Papageorgiou, and T. Poggio. Example-based object detection in image by components. *IEEE Trans. on PAMI*, 23(4):349 – 361, 2001. 1, 2

[16] B. Ommer and J. M. Buhmann. Learning compositional categorization models. In *ECCV*, 2006. 2

[17] A. Rangarajan, J. Coughlan, and A. L. Yuille. A bayesian network for relational shape matching. In *ICCV*, 2003. 1, 2

[18] H. Schneiderman and T. Kanade. Object detection using the statistic of parts. *IJCV*, 56(3):151–177, Feb. 2004. 1, 2

[19] L. G. Shapiro and R. M. Haralick. Structural descriptions and inexact matching. *IEEE Trans. on PAMI*, PAMI-3:504–519, 1981. 1

[20] Y. Song, X. Feng, and P. Perona. Towards detection of human motion. In *CVPR*, 2000. 1, 2

[21] F. Tarres and A. Rama. Gtav face database. available at http://gps.tsc.upc.es/GTAV/ResearchAreas/UPCFaceDatabase. 6

[22] A. Torralba, K. P. Murphy, and W. T. Freeman. Sharing visual features for multiclass and multiview object detection. *IEEE Trans. on PAMI*, 29(5):854–849, 2007. 8

[23] P. Viola and M. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 6

[24] Y. Weiss and W. T. Freeman. On the optimality of solutions of the max-product belief propagation algorithm in arbitrary graphs. *IEEE Trans. on Information Theory*, 47(2):723–735, 2001. 6

[25] B. Wu and R. Nevatia. Detection and tracking of multiple, partially occluded humans by bayesian combination of edgelet based part detectors. *IJCV*, 2007. 1, 2

[26] D. Zhang and S.-F. Chang. A generative-discriminative hybrid method for multi-view object detection. In *CVPR*, 2006. 1, 2