Semi-Supervised Learning Based Object Detection in Aerial Imagery

Jian Yao Zhongfei (Mark) Zhang
Department of Computer Science, State University of New York at Binghamton, NY 13905, USA
jyao@binghamton.edu Zhongfei@cs.binghamton.edu

Abstract

Object detection in aerial imagery has been well studied in computer vision for years. However, given the complexity of large variations of the appearance of the object and the background in a typical aerial image, a robust and efficient detection is still considered as an open and challenging problem. In this paper, we have developed a theoretic foundation for aerial imagery object detection using semi-supervised learning. Based on this theory, we have proposed a context-based object detection methodology. Both theoretic analyses and experimental evaluations have successfully demonstrated the great promise of the developed theory and the related detection methodology.

1. Introduction

Object detection in aerial imagery has been well studied in computer vision for years [8,11,14,28,33]. However, given the complexity of large variations of the appearance of the object and the background in a typical aerial image, a robust and efficient detection is still considered as an open and challenging problem.

The object detection problem is typically solved in two stages: *candidate generation* and *candidate classification*. Candidate generation generates regions of an image that may contain the object, and candidate classification further classifies and/or verifies the generated regions. Two types of methods for candidate generation are proposed in the literature. "Exhaustive search" methods [8,11,21,29] consider all the windows in an image as candidates while "segmentation based" methods [12,27,33] consider only the segmented features including regions as the candidates. The limitation of the "exhaustive search" methods is the demanding complexity, especially when the image resolution is very high, which is typically true for aerial imagery. The limitation of the "segmentation based" methods is the imperfect segmentation by nature.

The majority of the classification models used in detection proposed in the literature are based on supervised learning, including boosting model [25], cascade models [21,24,29], neural networks [10,11,20], Bayesian networks [33], generative models [27], and statistical models [26]. The problem with the supervised learning classification methods is that in order to achieve a reasonably good performance, typically a large training data set is required; the larger the training set, the more expensive to ground-truth the training data. Due to these considerations, we have developed a semi-supervised learning based classification theory that simultaneously resolves both problems. For the reference purpose, we call this theory as well as the classification method SLC.

The semi-supervised learning [2,4,9,30,32] has recently received intensive attention in machine learning community. The techniques developed in this research have been applied to solving problems in many areas including computer vision [4,30]. SLC is motivated to specifically take into account the reality for many computer vision problems including the problem this paper is addressing where only a very limited amount of labeled training data is available but at the same time there is always a large amount of unlabelled data available; on the other hand, the accuracy of a trained classifier is typically expected to be adaptive to different unlabelled data in different applications even with the same (often) very limited labeled training samples. Consequently, the contribution of SLC roots in the novel strategy to adaptively label the unlabelled samples given in an application, which is theoretically proven to be optimal to achieve the maximum accuracy.

Based on the developed theory, we have proposed a context-based aerial imagery object detection methodology, called CONTEXT in the rest of the paper. Context-based image understanding has been studied extensively in the literature [17,23,26]. Specifically, considering the application of aerial imagery object detection, it is well observed that typically an object is surrounded by a relatively homogeneous "background" region (e.g., an aircraft or a vehicle in the parking lot). CONTEXT takes the advantage of the availability of this specific context. The similar idea has been used in other efforts (e.g., [27]). The main difference is that context information is mostly used to improve the accuracy in the previous methods while it is also used to improve the efficiency in CONTEXT.

The paper is organized as follows. SLC theory and the method are presented in Sec. 2. CONTEXT is described in Sec. 3. The empirical evaluation focusing on aircraft detection in real aerial imagery is reported in Sec. 4. Finally, the paper is concluded in Sec. 5.

2. SLC theory and method

A typical semi-supervised learning method consists of three steps: 1) to train a classifier using the labeled training samples 2) to label the unlabelled training samples using the current classifier 3) to train the classifier using the labeled training samples and the current status of the unlabelled training samples, including their estimated labels and/or their probabilities. Steps 2 and 3 are iterated until some stop criteria are met.

The difference between the supervised learning and the semi-supervised learning is the existence and the use of the unlabelled samples. Consequently, the key step of a semi-supervised learning method is how to make a good use of the unlabelled samples. In the literature, there are two strategies developed to make use of the unlabelled data. One strategy considers that the unlabeled samples have hard labels and the learning procedure gradually updates the labels of the unlabelled samples until convergence (e.g., [2,9]). The other strategy considers that the unlabelled samples have soft (fuzzy) labels (i.e., labels with probabilities) and the learning procedure gradually updates the probabilities of the unlabelled samples until convergence (e.g., [15,19]). SLC theory follows the first strategy.

As shown below by SLC theory, an optimal classification may be achieved by an iterative process of two major steps: (1) labeling and (2) training. We first give the related theory developed for the two steps, and then list the overall learning procedure.

2.1 Labeling strategy

The labeling strategy assumes the following information as the given input: two labeled sample sets P (positive) and N (negative) and an unlabelled sample set $U=\{s_i\}$. The probability of being positive, $\{p_i\}$, for each unlabelled sample is also part of the input. The output of the labeling strategy is the estimated label $\{d_i\}$ for each unlabelled sample. The goal of the labeling strategy is to find the optimal label assignment for all the unlabelled samples, which leads to the maximum classifier accuracy. Assume that U can be further potentially decomposed into two mutually exclusive, arbitrary sets U_1 (positive sample set) and U_2 (negative sample set) and $U=U_1+U_2$. Assume that UP_1 and UN_1 are, respectively, the ground truth positive sample set and the ground truth negative sample set in U_1. Similar definitions apply to UP_2 and UN_2. Denote the correctly classified sample numbers in $P+U_1$ and $N+U_2$ as CP and CN, respectively. Then the classifier's true positive and true negative are:

$$TP=|CP|/(|P|+|U_1|) \quad (1)$$

$$TN=|CN|/(|N|+|U_2|) \quad (2)$$

Note that TP and TN are determined by considering a correct label assignment, i.e., $U_1=UP_1$ and $U_2=UN_2$. Denote the number of the correctly classified samples in UP_1, UP_2, UN_1, and UN_2 as CP_1, CP_2, CN_1, and CN_2, respectively. The *expected true positive* and the *expected true negative* are defined as:

$$RTP=(|P|\times TP+|CP_1+UP_2-CP_2|)/(|P+UP_1+UP_2|) \quad (3)$$

$$RTN=(|N|\times TN+|CN_2+UN_1-CN_1|)/(|P+UN_1+UN_2|) \quad (4)$$

Though RTP and RTN are estimated using the training samples, they are also correct for the test samples as we will show later. We use subscripts to denote the values of different classifiers (e.g., TP_1 is the TP value of classifier I). Assume that the trained classifier has $TP+TN>1$, i.e., it can correctly classify at least half of the training samples.

Lemma 1: *Assume that s_i and s_j are two unlabelled samples with $p_i>p_j$. For any two classifiers using the same training samples with the only difference that in classifier I s_i is considered as positive and s_j is considered as negative while in classifier II s_i is considered as negative and s_j is considered as positive, classifier I has both higher RTP and RTN than those of classifier II for the training samples.*

Proof: Since the two classifiers have almost identical training sample set, we assume that when the training sample set is sufficiently large, $TP_1=TP_2=TP$ and $TN_1=TN_2=TN$. Then we have:

$$RTP_1-RTP_2=(p_i-p_j)(TP+TN-1)/(|P|+|U_P|) \quad (5)$$

$$RTN_1-RTN_2=(p_i-p_j)(TP+TN-1)/(|P|+|U_N|) \quad (6)$$

Since $p_i>p_j$ and $TP+TN>1$ by the assumption, we have $RTP_1>RTP_2$ and $RTN_1>RTN_2$. □

From Lemma 1, it is clear that if the maximum positive probability of all the unlabelled samples in the negative sample set is higher than the minimum positive probability of all the unlabelled samples in the positive sample set, the accuracy can be increased by changing the labels of the two corresponding unlabelled samples. By iteratively applying this conclusion, we have:

Lemma 2: *Assume that we sort the unlabelled samples by the ascending order of their positive probabilities. Then the optimal label assignment of the unlabelled samples satisfies $p_i<p_j$ for any unlabelled sample s_i in the negative sample set and for any unlabelled sample s_j in the positive sample set.*

From Lemma 2, it is clear that the optimal label assignment problem is reduced to the problem of finding the optimal split threshold in the probability space. Theorem 1 provides an elegant solution to this problem without the exhaustive search of the threshold.

Theorem 1: *Given an arbitrary λ, for an accuracy function $\lambda\times RTP+(1-\lambda)\times RTN$, the label assignment which assigns the negative labels to the samples with their positive probabilities less than $1-\lambda$ and the positive labels to other samples is optimal.*

Proof: Assume that the samples in U are sorted by the ascending order of their positive probabilities p_i. Assume $U=U_1+U_2$ is the optimal assignment where U_1

is the negative sample set and U_2 is the positive sample set. Assume U_1 contains H samples. Based on Lemma 2, we know that U_1 contains the first H samples in U while U_2 contains the remaining samples. Denote the true positive and the true negative of the classifier as TP_H and TN_H. Denote the expected true positive and the expected true negative as RTP_H and RTN_H. We have:

$$RTP_H = (|P| + \sum_{j>H} p_j) \times TP_H + \sum_{j \le H} p_j \times (1 - TN_H) / (|P| + |U_P|) \quad (7)$$

$$RTN_H = (|N| + \sum_{j \le H} (1 - p_j)) \times TN_H + \sum_{j>H} (1 - p_j) \times (1 - TP_H) / (|N| + |U_N|) \quad (8)$$

Denote the accuracy function as $\Psi(H)$, we have:

$$\Psi(H) = \lambda \times \frac{TP_H \times (|P| + \sum_{j>H} p_j) + \sum_{j \le H} p_j \times (1 - TN_H)}{|P| + |U_P|} + (1-\lambda) \times \frac{TN_H \times (|N| + \sum_{j \le H} (1 - p_j)) + \sum_{j>H} (1 - p_j)(1 - TP_H}{|N| + |U_N|} \quad (9)$$

Based on the same argument used in the proof of Lemma 1, we assume that $TP_H = TP_{H+1} = TP_{H-1} = TP$ and that $TN_H = TN_{H+1} = TN_{H-1} = TN$. Since the label assignment is the optimal one, we have:

$$\Psi(H+1) - \Psi(H) \le 0 \text{ and } \Psi(H) - \Psi(H-1) \ge 0 \quad (10)$$

Replace H with $H+1$ and $H-1$, respectively, in (9) and substitute them accordingly in (10), we have

$$p_H \le \frac{(1-\lambda) \times (|P| + |U_P|)}{(1-\lambda) \times (|P| + |U_P|) + \lambda \times (|N| + |U_N|)} \le p_{H+1} \quad (11)$$

Based on the above argument, it is clear that $\Psi(H)$ increases when H increases until $p_H > T = \frac{(1-\lambda) \times (|P| + |U_P|)}{(1-\lambda) \times (|P| + |U_P|) + \lambda \times (|N| + |U_N|)}$ and then decreases when H increases. Consequently, by assigning positive labels to the unlabelled samples with a positive probability higher than T and negative labels to other unlabelled samples, the accuracy function reaches its maximum value, i.e., it is an optimal label assignment. □

We can select the λ based on the relative importance of the *RTPs* and the *RTPs*. This shows that SLC is adaptive to different applications with different foci of expected true positive and expected true negative combinations.

Theorem 1 identifies the optimal split threshold based on the positive probability of the unlabelled samples. Since the probabilities are generally unknown, in practice, an iterative procedure is used to estimate them by considering them as unknown values and using the EM algorithm [5] to solve for the problem.

2.2 Classifier training

After the labeling step, the input to the training step includes: the labeled samples, the unlabelled samples with their estimated labels and their positive probabilities p_i. Intuitively, an unlabelled sample with a large p_i has a large probability that its estimated label is correct. Similarly, an unlabelled sample with a low p_i has a large probability that its estimated label is correct. On the other hand, an unlabelled sample with a moderate p_i has a low probability that its estimated label is correct. Consequently, from Theorem 1, the *certainty* h_i is defined to represent the probability that the estimated label for an unlabelled sample s_i is correct:

$$h_i = \begin{cases} e^{|p_i - (1-\lambda)|/(1-\lambda) - 1}, & p_i < 1 - \lambda \\ e^{|p_i - (1-\lambda)|/\lambda - 1}, & p_i \ge 1 - \lambda \end{cases} \quad (12)$$

All the labeled samples have certainty 1. The reason to introduce the certainty is that we expect the unlabelled samples with a high certainty value to contribute to the learning more than the unlabelled samples with a low certainty value. This fact is incorporated into the training error that is used to evaluate the classifier. The certainty here is similar to the weight in [19] with the difference that a certainty value is dynamically determined while a weight value is a constant, as different samples may have different certainty values but the same weight.

The following summaries the learning algorithm:

1. Estimate the split threshold T using (11)
2. Learn classifier Ω_0 using L
3. Iteration number i set to *1*
4. While stop criteria do not meet
   For j=*1* to $|U|$
   a. Input s_j to Ω_{i-1} to determine p_j
   b. Determine h_j using equation (12)
   c. Set d_j=1 if $p_j > T$ and d_j=0 if $p_j \le T$
   Learn classifier Ω_i using L, U, d_j's, and h_j's
   Increase iteration number by *1*
5. Output Ω_{i-1}

3. Context-based object detection

We first define several terminologies before we present the methodology. In the subsequent text, the context regions in an image are referred to as the *surrounding regions* (SRs), and the foreground regions surrounded by the SRs are referred to as the *enclosed regions* (ERs). Though there are differences among the SRs in different images, in general, such differences are much less than the possible differences among the objects in different images. Consequently, it is easier to build a classifier for the SRs with a high true positive, which is the percentage of the SRs that are correctly classified, and an acceptable true negative, which is the percentage of the non-SRs that are correctly classified.

CONTEXT works as follows. An image is first segmented by a conservative segmentation algorithm. An SR classifier (called SRC) is applied to identify all the background regions to generate all the potential SRs.

Finally, all the ERs and combination of ERs are identified to form the object candidate set, which is in turn classified by an object verifier (OV).

3.1 SR detection

In order to detect SR, a segmentation algorithm is applied to an image to generate regions. Since we do not expect an accurate segmentation, a simple edge-based segmentation algorithm is used: first, three edge images based on, respectively, R, G, and B color components are generated using the gradient edge detector followed by a thresholding; second, the three edge images are combined and morphological operations are applied; third, the connected component algorithm is applied to the non-edge areas to generate the SR candidates.

Each region is denoted as a 7 dimensional vector $(x_1,x_2,..x_7)^T$, where x_1, x_2, and x_3 are, respectively, the means of the R, G, and B values of the region; x_4 is the intensity standard variance of the region. Each background region is divided into four sub-regions which are, respectively, the left-up, left-down, right-up, and right-down sub-regions, w.r.t. the center of the region. x_5 represents the standard variance of the intensity means of the four sub-regions; x_6 represents the standard variance of the intensity standard variances of the four sub-regions; and finally, x_7 represents the mean of the intensity standard variances of the four sub-regions. A linear discriminant analysis (LDA) model is selected as the base classifier of SRC where the unlabelled samples are exploited.

A classic LDA using only the labeled samples learns a model through maximizing the ratio of the between-class matrix (S_B) determinant to the within-class matrix (S_W) determinant. Now we extend the classic LDA to the LDA using the unlabelled samples. The differences are:

$$\mu_P = (\sum_{j\in P} s_j + \sum_{d_j=1} h_j \times s_j)/(|P| + \sum_{d_j=1} h_j) \quad (15)$$

$$\mu_N = (\sum_{j\in N} s_j + \sum_{d_j=0} h_j \times s_j)/(|N| + \sum_{d_j=0} h_j) \quad (16)$$

$$\mu = (\sum_{j\in L} s_j + \sum_{j\in U} h_j \times s_j)/(|L| + \sum_{j\in U} h_j) \quad (17)$$

$$S_W = \sum_{j\in P}(s_j-\mu_P)(s_j-\mu_P)^T + \sum_{d_j=1} h_j(s_j-\mu_P)(s_j-\mu_P)^T + \sum_{j\in N}(s_j-\mu_N)(s_j-\mu_N)^T + \sum_{d_j=0} h_j(s_j-\mu_N)(s_j-\mu_N)^T \quad (18)$$

$$S_B = (\mu_p-\mu)(\mu_p-\mu)^T + (\mu_N-\mu)(\mu_N-\mu)^T \quad (19)$$

Assuming that W is the estimated LDA projection matrix, the positive probability p_j is defined as:

$$p_j = |W(s_j-\mu_P)|/(|W(s_j-\mu_P)| + |W(s_j-\mu_N)|) \quad (20)$$

3.2 Object verification

After the SRs are detected, what is surrounded by the SRs may be either a single ER or a group of ERs. In the former case, the ER is considered as an object candidate immediately. The latter case is likely to be an over-segmentation scenario, and the ERs are merged together to form a single object candidate using the repeatedly subsampling approach [20]. Once an object candidate is generated, it is classified using OV. An object is detected if the candidate is classified as positive.

Two sets of features are extracted from a binarized edge image. The first set includes the 7 invariant moments in the order up to 3 while the second set is denoted as $(y_1,y_2..y_K)$, where y_i, $i = 1, \ldots, K$, is the number of the edge pixels whose distances to the gravity center of edge pixels in the object candidate region are less than i/K of the maximum distance between any edge pixel to the center. Note that both sets of features are translation and rotation invariant. To address the scale-invariant property, we explicitly apply a scale transformation to the training samples.

We argue that using the edge information only, as we proposed here, is sufficient for the object verification in the context of object detection in aerial imagery. This is due to the fact that for many objects in aerial imagery the shape features are typically preferred to other features such as color, as there are typically a very limited number of object shapes for these objects (e.g., aircraft) as compared with a substantially larger number of variations for other features such as object color. Since edge image is a good representation of object shape feature, we elect to use edge information as the only feature for OV. This observation is also true in many object detection efforts beyond the aerial imagery context from the state-of-the-art literature (e.g., the EOH features for the face detection work in [13]).

To learn OV, similar to SRC learning, a small number of labeled samples and a large number of unlabelled samples are selected. The differences are the λ selection and the base classifier selection. Since now there is no preference on *RTP* or *RTN*, λ is set to *0.5*. Since it is unlikely that the object feature space and the non-object feature space are linearly separable, a non-linear classifier is used to achieve the expected accuracy. A standard multi-layer perceptron (MLP) with Back Propagation (BP) training is used due to its simplicity to accommodate the certainty. Assuming that the original updating weight in a standard BP training is ΔW incurred by the training sample s_i with certainty h_i, the modified updating weight here is $h_i \Delta W$.

Due to the expected imperfect segmentation and the typical complexity of the object appearance in aerial imagery, occasionally an object candidate may be over-segmented or under-segmented. The over-segmentation problem is resolved by repeatedly subsumpling [20] the ERs. An object candidate failing to be verified in a

higher resolution may be verified in a lower resolution. The under-segmentation scenario typically occurs, in the context of aerial imagery, when two objects are located very close to each other. In this case, we consider this type of under-segmentation as a correct segmentation as long as OV correctly detects the ER as the object. In Fig. 3 (c), the four aircrafts in the left-middle of the image are detected as one "aircraft" object.

We note that for aerial imagery, CONTEXT is more efficient and effective than the "direct" methods that the majority of the existing object detection work in the literature is categorized to. This is due to the fact that for a typical aerial image, the "background" regions are typically homogeneous and there are only a very limited number of variations, which means that it is easy to train SRC. With the detected SR context, the number of possible variations of an object candidate in terms of a feature space (e.g., the shape feature) is rather limited, which means that it is much easier to train OV than to train a typical object classifier using a "direct" method. Consequently, the training complexity of CONTEXT is much lower than that of a typical "direct" method. Therefore, with the same training complexity, CONTEXT would result in higher detection accuracy than a typical "direct" method. This observation is supported in the experiments in Sec. 4.

4. Evaluation

We evaluate CONTEXT by focusing on aircraft detection. The evaluation data set consists of 42 color aerial images of airport scenes. The images vary substantially in scale, with the maximum resolution as 10496×21618 and the minimum resolution as 326×291. The number of aircrafts in each image varies from 13 to 47. The minimum aircraft is in a 17×17 bounding box and the maximum aircraft is in a 348×348 bounding box.

24 images are randomly selected to build the training samples. The segmentation algorithm described in Sec. 3.1 is first applied to the 24 images. 15 positive SRs and 15 negative SRs are manually selected. In addition, 270 unlabelled SRs are also randomly selected. These 300 regions form the training sample set for SRC.

After the learning of SRC, the SRs in the 24 images are detected. 25 positive aircraft samples and 25 negative aircraft samples are manually selected and extracted from the ERs. For each sample, scale transformations of 1.2, 1.5, and 1.8 are applied. Thus, we have 100 positive samples and 100 negative samples to train OV. In addition, 800 unlabelled samples are randomly selected to add into the training set.

In order to evaluate the effect of the number of the training samples vs. the classifier accuracy, the following experiment is conducted. Denote the number of the training samples for a classifier, either SRC or OV, as *A*, where *B* percent of the samples are labeled samples. The classifier accuracy with different *A* and *B* values is reported in Table 1 and Table 2. In both Tables, the first column represents *A* values while the first row represents *B* values. From the Tables, it is clear that in general the accuracy increases with the increase of *A* or *B* except when *A* or *B* is small. For SRC, typically *RTP* is higher than *RTN*, which is consistent with the higher λ value selected for SRC.

Table 1: SRC evaluation results

	2%	5%	10%	15%	20%
160	0.76/0.52	0.83/0.52	0.78/0.68	0.83/0.67	0.82/0.78
180	0.82/0.64	0.87/0.71	0.89/0.69	0.92/0.73	0.93/0.78
200	0.87/0.73	0.90/0.77	0.96/0.83	0.96/0.85	0.96/0.85
220	0.89/0.79	0.91/0.81	0.99/0.85	0.99/0.87	0.99/0.87

Table 2: OV evaluation results

	5%	10%	15%	20%	25%
800	0.69/0.72	0.73/0.77	0.77/0.82	0.77/0.84	0.81/0.88
900	0.81/0.79	0.84/0.82	0.88/0.89	0.91/0.90	0.92/0.93
1000	0.83/0.84	0.87/0.84	0.91/0.89	0.93/0.92	0.94/0.95
1100	0.85/0.84	0.86/0.80	0.90/0.89	0.93/0.92	0.94/0.96

We apply CONTEXT to all the 42 images. Examples of the detection are displayed in Fig. 4. The processing time is very promising. For images with a resolution about 2000×2000, the detection time is less than 1 second under the platform of Pentium IV 2GHz CPU with 512MB memory. From the figure, it is clear that almost all the aircrafts, including the helicopters, are successfully detected given the complex and varied appearances of the object and the background.

In order to demonstrate the strength of SLC theory and the method, we evaluate the classification performance of SLC against the classic semi-supervised learning method by Nigam et al [19] (for the reference purpose, we call it NMTM). To ensure a fair comparison, we use the same features and the comparison is observed for SRC, OV, and the combination. Fig. 3 documents the performance comparison between the two semi-supervised learning methods using the same aircraft detection data set. From the figure, it is clear that SLC performs substantially better than this classic method. In addition, it appears that SRC accuracy difference between the two methods is in general larger than OV accuracy difference between the two methods. This is due to the fact that *RTP* is more important than *RTN* in SRC.

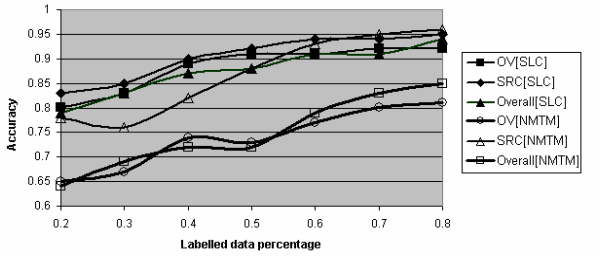

Fig. 3: Performance comparison between SLC and NMTM.

In order to experimentally justify that CONTEXT is superior to the conventional "direct" methods for aerial imagery, we disabled SRC and used only OV for object detection. We call this method as DIRECT for the reference purpose. To ensure a fair comparison, we used the number of samples to train DIRECT as the total number of the samples to train SRC and OV in CONTEXT. The detection performance is measured in terms of the *detection rate*, which is defined as the percentage of the number of correctly detected objects from the ground-truthed number of objects in the data set, and the *false alarm rate*, which is defined as the percentage of the number of incorrectly detected objects from the number of the detected objects in the data set. Table 3 documents the performance comparison between CONTEXT and DIRECT. Clearly, CONTEXT outperforms DIRECT. It is also noted that in addition to the better detection effectiveness, CONTEXT also outperforms DIRECT in detection efficiency.

Finally, in order to demonstrate the strength of CONTEXT, we evaluate the performance of CONTEXT against that of a state-of-the-art object detection method by Viola&Jones [29] (called VJ here) using all the test images for aircraft detection. To ensure a reasonable performance for VJ, we used twice the number of the labeled samples as the total number of the labeled samples used in SRC and OV in CONTEXT. Table 3 reports the comparison. Clearly, CONTEXT outperforms VJ. It is also noticed that VJ is substantially slower in terms of the detection time than both DIRECT and CONTEXT.

Table 3: Detection performance comparison

	CONTEXT	DIRECT	VJ
Detection rate	95%	91%	89%
False alarm rate	7%	27%	17%
Detection time	0.27sec.	0.51sec.	3.63min.

5. Conclusion

In this paper, we have made two contributions. First, we have developed a theoretic foundation for aerial imagery object detection using semi-supervised learning. Second, based on this theory, we have proposed a context-based object detection methodology. Both theoretic analyses and experimental evaluations have successfully demonstrated the great promise of the developed theory and the related detection methodology.

Reference

[2] K.P.Bennett, A.Demiriz, and R.Maclin, "Exploiting Unlabeled Data in Ensemble Methods", *KDD* 2002, 289-296

[3] P.B.Chou and C.M.Brown, "The theory and practice of bayesian image labeling", *Int. J. Comp. Vis.*, 4(1990), 185-210

[4] I.Cohen, F.G.Cozman, N.Sebe, M.C.Cirelo, and T.S.Huang, "Semisupervised Learning of Classifiers: Theory, Algorithms, and Their Application to Human-Computer Interaction", *PAMI* 26(12), 2004, 1553-1567

[5] A.P.Dempster, N.M.Laird, and D.B.Rubin, "Maximum-likelihood from incomplete data via the em algorithm", *J.Royal Statist. Soc.*, 39, 1977, 1-38

[8] A.Filippidis, L.C.Jain, and N.Martin, "Fusion of intelligent agents for the detection of aircraft in SAR images", *PAMI* 22(4), 2000, 378-383

[9] S.A.Goldman and Y.Zhou, "Enhancing Supervised Learning with Unlabeled Data", *ICML* 2000

[10] B.Kamgar-Parsi, B.Kamgar-Parsi, A.K.Jain, and J.E.Dayhoff, "Aircraft Detection: A case study in using human similarity measure", *PAMI* 23(12), 2001, 1404-1414

[11] M.A.Khabou and P.D.Gader, "Automatic target detection using entropy optimized shared-weight neural networks", *IEEE. Trans. Neural Network*, 11(1), 2000, 186-193

[12] Z.Kim and J.Malik, "Fast Vehicle Detection with Probabilistic Feature Grouping and Its Application to Vehicle Tracking", *ICCV* 2003

[13] K.Levi and Y.Weiss, "Learning Object Detection from a Small Number of Examples: the Importance of Good Features", *CVPR* 2004, 53-60

[14] J.Li, R.Nevatia, and S.Nornoha, "User Assisted Modeling of Buildings from Aerial Images", *CVPR*, 1999

[15] B.Liu, W-S.Lee, P.S.Yu, and X-L.Li, "Partially supervised classification of text documents", *ICML* 2002

[17] J.L.Mundy and T.M.Strat, *Proc. IEEE Workshop on Context Based Vision*, 1995

[19] K.Nigam, A.K.McCallum, S.Thrun, & T.M.Mitchell, "Text classification from labeled and unlabelled data using EM", *Machine Learning*, 2000 (39), 103-134

[20] H.A.Rowley, S.Baluja, and T.Kanade, "Neural network-based face detection", *PAMI,* 20(1):23–38, 1998

[21] H.Schneiderman, "Feature-centric evaluation for efficient cascaded object detection", *CVPR* 2004, 29-36

[23] T.M.Strat and M.A.Fischler, "Context-Based Vision: Recognizing Objects Using Information from Both 2D and 3D Imagery", *PAMI*, 13(10), 1991, 1050-1065

[24] J.Sun, J.M.Rehg, and A.Bobick, "Automatic Cascade Training with Perturbation Bias", *CVPR* 2004, 276-283

[25] A.Torralba, K.P.Murphy, and W.T.Freeman, "Sharing visual features for multiclass and multiview object detection", *CVPR* 2004, 762-769

[26] Anatonio Torralba and Pawan Sinha, "Statistical context priming for object detection", *ICCV* 2001, 763-770

[27] Z.Tu, X.Chen, A.L.Yuille, and S-C.Zhu, "Image parsing: unifying segmentation, detection, and recognition", *ICCV* 2003, 18-25

[28] V.Venkateswar and R.Chellappa, "A Framework for Interpretation of Aerial Images", *ICPR*, 1990, 204-206

[29] P.Viola and M.Jones, "Rapid object detection using a boosted cascade of simple features", *CVPR* 2001, 511-518

[30] Ying Wu, Thomas S. Huang, and Kentaro Toyama, "Self-supervised learning for object recognition based on kernel discriminant-EM algorithm", *ICCV* 2001, 275-280

[32] T.Zhang, F.J.Oles, "A probability analysis on the value of unlabeled data for classification problems", *ICML* 2000

[33] Tao Zhao and Ram Nevatia, "Car detection in low resolution aerial image", *ICCV* 2001, 710-71