The Fourteenth International Symposium on Artificial Life and Robotics 2009 (AROB 14th '09),
B-Con Plaza, Beppu, Oita, Japan, February 5 - 7, 2009

Covariance-based Recognition Using Incremental Learning Approach

HASSAB ELGAWI Osman
Image Science and Engineering Laboratory, Tokyo Institute of Technology, Japan
Email: osman@isl.titech.ac.jp

Abstract

We propose an on-line machine learning approach for object recognition, where new images are continuously added and the recognition decision is made without delay. Random forest (RF) classifier has been extensively used as a generative model for classification and regression applications. We extend this technique for the task of building incremental component-based detector. First we employ object descriptor model based on bag of covariance matrices, to represent an object region then run our on-line RF learner to select object descriptors and to learn an object classifier. Experiments of the object recognition are provided to verify the effectiveness of the proposed approach. Results demonstrate that the propose model yields in object recognition performance comparable to the benchmark standard RF, AdaBoost, and SVM classifiers.

1 Introduction

Object recognition is one of the core problems in computer vision, and it turns out to be extremely difficult for reproduce in artificial devices, simulated or real. Specifically, an object recognition system must be able to detect the presence or absence of an object, under different illuminations, scales, pose, and under differing amounts of background clutter. In addition, the computational complexity is required to be kept minimum, in order for those algorithms to be applicable for real-life applications. Based on "strongly supervised" approach and "weakly supervised" method (*without using any ground truth information or bounding box during the training*), considerable progress has been made for detection of objects. Several studies also have shown that supervised component-based approach is more robust to natural pose variations, than

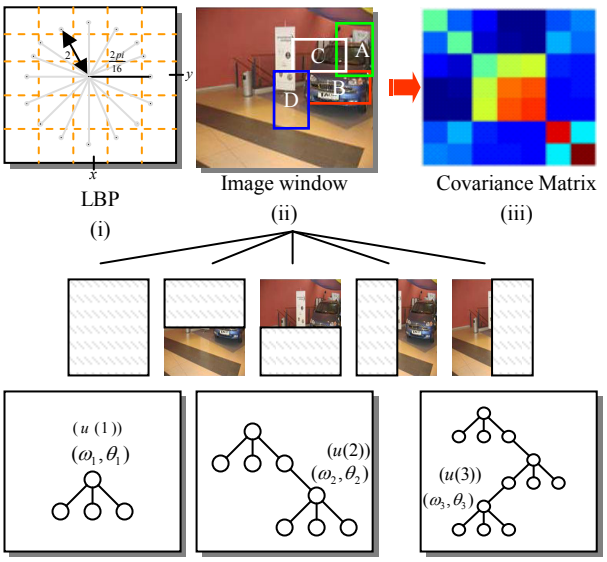

Figure 1: (i) Points sampled to calculate the LBP around a point (x, y). (ii) Rectangles [A, B, C and D] are examples of possible regions for histogram features. (iii)Any region can be represented by a covariance matrix. Size of the covariance matrix is proportional to the number of features used

the traditional global holistic approach. However, supervised learning is usually carried out batch on the entire training set, often is not optimal in a dynamic recognition tasks. In this paper we consider instead how machine learning models for object recognition categories, can be build 'incrementally' or 'on-line' so that new images are continuously added and the recognition decision is made without delay. The process consists of two stages. First we employ object descriptor model based on bag of covariance matrices, to represent an image window then run our online random forest (RF) learning algorithm [3]. RF technique has been extend in this paper for the task

of building incremental component-based detector, for attacking the problem of recognizing generic object categories, such as bikes, cars or persons purely from object descriptors that combines histograms and appearance model.

1.1 Our Object Descriptor Approach

We have used bag of covariance matrices, to represent an object region. Let I be an input color image. Let F be the $W \times H \times d$ dimensional feature image extracted from I

$$F_{W,H,d}(x,y) = \phi(I,x,y) \tag{1}$$

where the function ϕ can be any feature maps (such as intensity, color, etc). For a given region $R \subset F$, let $\{z_k\}_{k=1\cdots n}$ be the d dimensional feature points inside R. We represent the region R with the $d \times d$ covariance matrix C_R of feature points.

$$C_R = \frac{1}{n-1}\sum_{k=1}^{n}(z_k - \mu)(z_k - \mu)^T \tag{2}$$

where μ is the mean of the point. Fig. 1 (i) depicts the points that must be sampled around a particular point (x,y) in order to calculate the LBP at (x,y). In our implementation, each sample point lies at a distance of 2 pixels from (x,y), instead of the traditional 3×3 rectangular neighborhood, we sample neighborhood circularly with two different radii (1 and 3). The resulting operators are denoted by $LBP_{8,1}$ and $LBP_{8,1+8,3}$, where subscripts tell the number of samples and the neighborhood radii. In Fig. 1 (ii), different regions of an object may have different descriptive power and hence, difference impact on the learning and recognition. We follow [4] and represent an object with five covariance matrices $C_{i=1\cdots5}$ of the feature computed inside the object region, as shown in the second row of Fig.1. A bag of covariance which is necessary a combination of Ohta color space histogram ($I_1 = R+G+B/3$, $I_2 = R-B$, $I_3 = (2G-R-B)/2$), LBP and appearance model of different features of an image window is presented in Fig.1 (iii). We use this representation to automatically detect any target in images. We then apply on-line RF learner to select object descriptors and to learn an object classifier.

2 Machine Learning Approach

In the following we introduce the on-line random forests learning algorithm [3] for object recognition based on Breiman's random forest (RF) [1]. Details discussion of Breiman's random forest learning algorithm is beyond the scope of this paper, however, in order to simplify the further discussion, we will need to define some fundamental terms:

Random Forests (RF) is a tree-based ensemble prediction technique combining properties of an efficient classifier and feature selection [1]. Briefly, it is an ensemble of two sources of randomness to generate base decision trees; bootstrap replication of instances for each tree and sampling a random subset of features at each node.

Decision tree. For the k-th tree, a random vector C_k is generated, independent of the past random vectors $C_1, \ldots, C_{k-1}$, and a tree is grown using the training set positive and negative image I and covariance feature C_k. The decision generated by a decision tree corresponds to a covariance feature selected by learning algorithm. Each tree casts a unit vote for a single matrix from the bag of covariance matrices.

Base classifier. Given a set of M decision trees, a base classifier selects exactly one decision tree classifier from this set, resulting in a classifier $h(I, C_k)$.

Forest Given a set of N base classifiers, a forest is computed as ensemble of these tree-generated base classifiers $h(I, C_k)$, $k = 1, \ldots, n$. Finally, a forest detector is computed as a majority vote.

2.1 On-line Learning Random forest (RF)

To obtain an on-line algorithm, each of the steps described above must be on-line, where the current classifier is updated whenever a new sample arrives. In particular on-line RF works as follows: First, the fixed set tree K is initialized. In contrast to off-line random forests, where the root node always represents the object class in on-line mode, for each training sample, the tree adapts the decision at each intermediate node (nonterminal) from the response of the leaf nodes, which characterized by a vector (w_i, θ_i) with $\|w_i\| = 1$. Root node numbered as 1, the activation of two child nodes $2i$ and $2i+1$ of node i is given as

$$u_{2i} = u_i.f(w_i^{'}I + \theta_i) \tag{3}$$

$$u_{2i+1} = u_i.f(-w_i^{'}I + \theta_i) \tag{4}$$

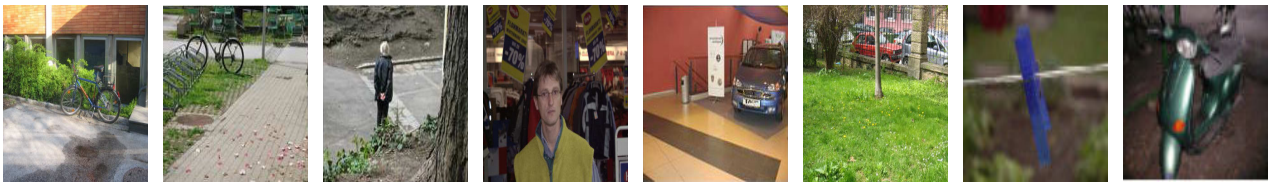

Figure 2: Examples from GRAZ02 dataset [2] for four different categories: bikes (1st pair), people (2nd pair), cars (3rd pair), and background (4th pair).

where I is the input image, u_i represents the activation of node i, and $f(.)$ is chosen as a sigmoidal function. Consider a sigmoidal activation function $f(.)$, the sum of the activation of all leaf nodes is always unity provided that the root node has unit activation. The forest consist of fully grown trees of a certain depth l. The general performance of the on-line forests depends on the depth of the tree. However, we found that the number of trees one needs for good performance eventually tails off as new data vectors are considered. Since after a certain depth, the performance of on-line forest does not vary to a great extent, the user may choose K (the number of trees in forest) to be some fixed value or may allow it to grow up to the maximum possible which is at most $|T|/N_k$, where N_k the tree size chosen by the user.

3 Object Recognition

Given a feature set and a sample set of positive (contains the object relevant to the class) and negative (does not contain the object) images, to detect a specific object, e.g. human, in a given image, we train a random forests learner (detector) offline using covariance descriptors of positive and negative samples. We start by evaluation feature from input image I after the detector is scanned over it at multiple locations and scales. This has to be done for each object. Then for feature in I, we want to find corresponding covariance matrix for estimating a decision tree. Each decision tree learner may explore any feature f, we keep continuously accepting or rejecting potential covariance matrices. We then apply the on-line random forests at each candidate image window to determine whether the window depicts the target object or not. The on-line RF detector was defined as a 2 stage problem, with 2 possible outputs in each stage: In the first one, we build a detector that can decide if the image contains an object, and thus must be recognized, or if the image does not contain objects, and can be discarded, saving processing time. In the second stage, based on selected features the detector must decide which object descriptor should be used. There are two parameters controlling the learning recognition process: The depth of the tree, and the least node. It is not clear how to select the depth of the on-line forests. One alternative is to create a growing on-line forests where we first start with an on-line forest of depth one. Once it converges to a local optimum, we increase the depth. Thus, we create our on-line forest by iteratively increasing its depth.

Table 1: Number of images and objects in each class in the GRAZ02 dataset.

Dataset	Images	Objects
Bikes	373	511
Cars	420	770
Persons	460	785
Total	1253	2066

4 Experiments and Evaluation

To evaluate and validate our approach we used data derived from the GRAZ02[1] dataset [2], a collection of 640×480 24-bit color images and illustrated in Figure 2. As can be seen in Table 1, this dataset has three object classes, bikes (373 images), cars (420 images) and persons (460 images), and a background class (270 images).

4.1 Experimental settings

For testing our framework we used the datasets described above and run it against three state of the art

[1] available at *htt://www.emt.tugraz.at/pinz/data/*

Table 2: Mean AUC performance of four classifiers on the Bikes vs. Background dataset, by amount of training data. Performance of on-line RF is reported for different Depths

	On-line RF					RF	AdaB	SVM
	D3	D4	D5	D6	D7			
10%	0.85	0.86	0.81	0.85	0. 85	0.86	0.81	0.82
50%	0.91	0.90	0.89	0.91	0.92	0.90	0.89	0.90
90%	0.92	0.90	0.91	0.92	0. 92	0.91	0.90	0.91

Table 3: Mean AUC performance of four classifiers on the Cars vs. Background dataset, by amount of training data. Performance of on-line RF is reported for different Depths

	On-line RF					RF	AdaB	SVM
	D3	D4	D5	D6	D7			
10%	0.77	0.79	0.75	0.78	0.73	0.79	0.75	0.73
50%	0.85	0.84	0.82	0.82	0.84	0.85	0.82	0.80
90%	0.86	0.82	0.83	0.85	0.86	0.85	0.83	0.82

Table 4: Mean AUC performance of four classifiers on the Persons vs. Background dataset, by amount of training data. Performance of on-line RF is reported for different Depths

	On-line RF					RF	AdaB	SVM
	D3	D4	D5	D6	D7			
10%	0.84	0.84	0.83	0.80	0.83	0.84	0.77	0.80
50%	0.88	0.86	0.88	0.88	0.88	0.88	0.84	0.86
90%	0.90	0.86	0.89	0.90	0.90	0.90	0.86	0.89

classifiers (offline RF, AdaBoost, and SVM). Each of the classifiers used in our experimentation were trained with varying amounts (10%, 50% and 90% respectively) of randomly selected training data. All image not selected for the training split were put into the test split.

5 Experimental Results

GRAZ02 images contain only one object category per image so the recognition task can be seen as a binary classification problem: bikes vs. background, people vs. background, and car vs. background. The well known statistic measure; the Area Under the ROC Curve (AUC) is used to measure the classifiers performance in these object recognition experiments.

5.1 Mean AUC Performance

Tables 2, 3, and 4 give the mean AUC values across all runs to 2 decimal places for each of the classifier and training data amount combinations, for the bikes, cars ad people datasets respectively. For on-line RF we report the results for different depths of the tree. As can be seen, our algorithm always performs significantly better than the offline RF. We found that the differences in performance are (avg. $= 1.2 \pm 15\%$).The improvement when we varying the tree depth are relatively small. This makes intuitive sense: when an image is characterized by high geometric variability, it is difficult to find useful global features.

6 Conclusions

In this paper we have presented an on-line learning framework for object recognition categories that avoids hand labeling of training data. We have demonstrated that on-line learning obtain comparable results to offline learning. Moreover, the proposed framework is quite general (i.e, it can be used to learn completely different objects) and can be extended in several ways.

References

[1] Leo Breiman, "Random Forests," *Machine Learning*, 45(1):5.32, 2001.

[2] Oplet A., Fussenegger M., Pinz A. and Auer P. "Generic object recognition with boosting," *IEEE Transactions on Pattern Analysis and Machine Intelligence* 28(3) pp. 416-431, 2006.

[3] Hassab Elgawi Osman, "Online Random Forests based on CorrFS and CorrBE," *In Proc.IEEE workshop on online classification, CVPR*, 2008.

[4] O. Tuzel, F. Porikli, and P. Meer. "Region covariance: A fast descriptor for detection and classification," *In Proc. ECCV*, 2006.