Online Domain Adaptation of a Pre-Trained Cascade of Classifiers

Vidit Jain
Yahoo! Labs Bangalore
viditj@yahoo-inc.com

Erik Learned-Miller
University of Massachusetts Amherst
elm@cs.umass.edu

Abstract

Many classifiers are trained with massive training sets only to be applied at test time on data from a different distribution. How can we rapidly and simply adapt a classifier to a new test distribution, even when we do not have access to the original training data? We present an on-line approach for rapidly adapting a "black box" classifier to a new test data set without retraining the classifier or examining the original optimization criterion. Assuming the original classifier outputs a continuous number for which a threshold gives the class, we reclassify points near the original boundary using a Gaussian process regression scheme. We show how this general procedure can be used in the context of a classifier cascade, demonstrating performance that far exceeds state-of-the-art results in face detection on a standard data set. We also draw connections to work in semi-supervised learning, domain adaptation, and information regularization.

1. Introduction

Supervised learning relies on the assumption of similarity between the distribution of training and test instances. However in practice there are often significant differences between these distributions. These differences arise due to the cost of collecting large training data sets and also to the difficulties in obtaining training instances from a particular target test domain. In *face detection*, for example, where the goal is to determine the position and size of all of the faces appearing in a given image, it may be infeasible to collect training data for the enormous variety of domains in which face detection is useful. In realistic applications, then, we expect to encounter domains at test time for which we have seen little training data. Furthermore, even when doing face detection in domains for which we do have significant training data, we may be able to perform significantly better classification by adapting our classifier's output for individual images from these domains. That is, the classification boundary learned locally for each image may lead to better performance than a global classification boundary for the entire domain.

The main contribution of this paper is a method for adapting pre-trained classifiers to a new test domain to improve performance. We demonstrate a dramatic increase in the state-of-the-art performance on a standard face detection data set. While other work has addressed the case in which a small number of labeled examples are available from the target domain, we focus on the extreme case in which no labeled data are available for the new domain. We also assume, as described below, that we do not have access to the training data from which the original classifier was derived. Furthermore, we assume that there is no known relationship among our *test* images. That is, we assume that each new test image may represent a new domain for the face detection problem. This means that there is only limited information to share across images. Hence our method re-adapts a pre-existing classifier to each new image it encounters. As shown below, there is a surprisingly large amount to be gained by adapting a classifier, even using the information in just a single image.

Our domain adaptation approach exploits the structure in the appearance of the face regions in an image and the image regions very unlikely to correspond to faces, to predict the detection label collectively for all the candidate regions in an image. This differs from the typical approach of applying a classifier to each of these regions *independently* [12, 13, 15, 17].[1] The independence assumption of previous face detection systems is made for computational and statistical reasons, facilitating simpler models for appearance distributions and avoiding the issue of a lack of samples to learn complex dependencies. This assumption, however, may limit the performance of such detectors in complex scenes.

Consider the image shown in Figure 1. A detector is likely to fail on the face of the person sitting in the bottom-left corner because of a strong shadow on the left half of the face. The shadow is weaker on the two faces in the right half of the image, and a good face detector may detect them successfully. Being part of a common scene, all

[1] The reader is referred to Zhang et al.'s survey [19] for a discussion of different approaches to face detection.

four faces appearing in this image share a common illumination source. The two easy-to-detect faces could subsequently be used to infer common structure in the appearance of all the faces in this image, allowing us to normalize the difficult-to-detect candidate face regions and ultimately classify them correctly. This same reasoning can be applied to background patches, reducing both false negatives and false positives.

Figure 1. *Easy-to-detect faces can help resolve difficult-to-detect faces.* There is a shadow in the left half of all four faces. The two faces on the left are difficult to detect because each of them has a strong shadow. Being part of the same scene, these faces share the common illumination source with the other two weakly shadowed faces, which are easier to detect. A detector that can adapt itself to a particular scene using such easy-to-detect faces can normalize other co-occurring faces and thus reduce their difficulty of detection.

One naïve way to implement the above intuition is to scan the image for high-confidence faces, and then retrain the detection model according to the high-confidence face and non-face regions. This two-stage process has two problems. First, the substantial increase in computation required for re-training the detection model for each image makes this approach infeasible in real-time environments. Second, this approach may lead to overfitting to the first stage predictions. We address both of these issues by adapting our model using Gaussian process regression (GPR). For GPR, the parameters can be analytically computed in an efficient manner. Also, a term for the prior probability of detection parameters prevents overfitting to the new training examples.

In Section 2, we elaborate on this formulation of face detection as a regression problem. In particular, we will use GPR, the basics of which are explained in Section 3. In Section 4, we present our method for online domain adaptation using GPR. In Section 5, we describe an algorithm to incorporate this domain adaptation procedure in a cascade of classifiers. This classification cascade is used to achieve state-of-the-art performance in face detection. The details of these experiments are included in Section 6. The proposed work is similar to work in a variety of fields. We discuss these related approaches and distinguish our method from the previous work in Section 7. Finally, we conclude in Section 8.

2. Face Detection As Regression

The image quality of a face region depends on several factors including the pose of the person, the distance of the face from the camera, and the occlusion of the face by other objects present in the scene. For instance, in Figure 2, the resolution of the faces of people in the audience is much lower than the resolution of the faces of the players. Thus, there may be little common structure between these two classes of faces. Furthermore, there may be more than two such modes in the distribution of face appearances in a single scene.

Figure 2. *Multiple modes in appearance quality.* The resolution of the faces of people in the audience is much lower than the resolution of the faces of the players. Although the faces may be similar in appearance within each of these two types of faces, it is likely that there is little commonality in the structure of appearance across different types within a single image.

As described in the previous section, our approach exploits the common appearance structure among the face regions in an image. Due to the potential presence of multiple modes in a single image, we formulate face detection as solving a regression problem rather than a classification problem. In this formulation, our approach predicts similar detection scores for image regions that are similar in appearance. In other words, the detection scores for the faces in the audience are encouraged to be similar to each other but may be different from (or similar to) the detection scores for the regions corresponding to the faces of players.

Having similar detection scores assigned to image regions with similar appearance has an additional advantage. A visual analysis of a collection of faces sorted by their detection scores provides a natural way to select the threshold value appropriate for the application domain at hand. For instance, we may want to reject low-resolution faces while organizing personal photo collections, but would want to include them for surveillance applications.

In this work, we use Gaussian process regression with an appropriate similarity kernel to update the detection scores

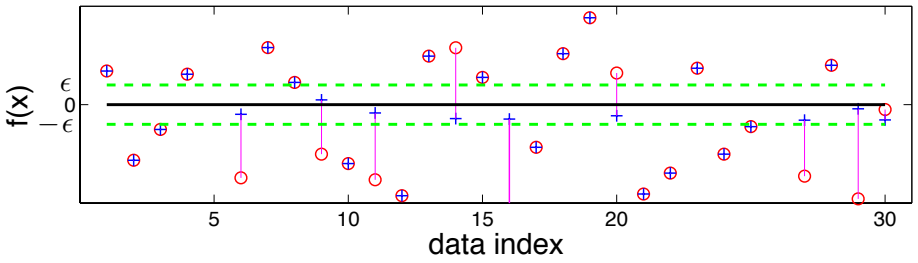


Figure 3. *Illustration of the proposed approach for online domain adaptation*. Let $f(x)$ denote the output of a classifier on a data point x. Consider an ϵ margin (green dotted line) around the classification boundary (black solid line). For points lying in the margin the classifier is not very certain about the predictive label. The proposed method updates the scores for the points in this margin based on their similarity to the other points for which the classifier is relatively more confident about the classification label. The original classification output is shown using 'blue +', whereas the updated output is shown using 'red o'.

for candidate image regions. In the next section, the notation used in the rest of the paper is introduced, and the basics of Gaussian process regression are briefly summarized.

3. Gaussian Process Regression

A Gaussian process refers to a stochastic process for which every finite set of samples is jointly Gaussian. When a Gaussian process prior is used in a Bayesian regression model for inferring continuous valued output, the resulting regression is called Gaussian process regression (GPR) (see Rasmussen and Williams [11] for further details). Consider the standard regression model with Gaussian noise:

$$y = \mathbf{x}^T\mathbf{w} + \eta, \tag{1}$$

where y is the target variable, $\mathbf{x}$ is the input vector, and $\eta \sim \mathcal{N}(0, \sigma_n^2)$. Let us assume a zero mean Gaussian prior on $\mathbf{w}$, i.e., $\mathbf{w} \sim \mathcal{N}(0, \Sigma_p)$. Using $\mathbf{X}$ to denote the collection of all input vectors, the conditional likelihood and the posterior distribution are respectively given by

$$p(\mathbf{y}|\mathbf{X}, \mathbf{w}) = \mathcal{N}(\mathbf{X}^T\mathbf{w}, \sigma^2 I), \tag{2}$$

$$p(\mathbf{w}|\mathbf{X}, \mathbf{y}) = \mathcal{N}(\frac{1}{\sigma_n^2}A^{-1}\mathbf{X}\mathbf{y}, A^{-1}), \tag{3}$$

where $A = \sigma_n^{-2}\mathbf{X}\mathbf{X}^T + \Sigma_p^{-1}$. Subsequently, the prediction for a new example $\mathbf{x}_*$ is given by

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\frac{1}{\sigma_n^2}\mathbf{x}_*^T A^{-1}\mathbf{X}\mathbf{y}, \mathbf{x}_*^T A^{-1}\mathbf{x}_*). \tag{4}$$

Instead of using the original representation for data $\mathbf{x}$, a function $\phi(\cdot)$ can be used to project $\mathbf{x}$ into a (potentially) higher-dimensional space. After rearranging a few terms in the expression for the prediction for a new example $\mathbf{x}_*$, we have

$$p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = \mathcal{N}(\phi_*^T\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\mathbf{y}, \phi_*^T\Sigma_p\phi_* - \phi_*^T\Sigma_p\Phi(K + \sigma_n^2 I)^{-1}\Phi^T\Sigma_p\phi_*), \tag{5}$$

where $\phi_* = \phi(\mathbf{x}_*)$, $\Phi = \Phi(\mathbf{X})$, and $K = \Phi^T\Sigma_p\Phi$. Since all the terms involving the projection function ϕ occur in the form $\phi(\mathbf{x})\Sigma_p\phi(\mathbf{x}')$, an appropriate covariance function or *kernel* $K_{\mathbf{x},\mathbf{x}'}$ can be used to avoid an explicit representation of the feature space. Thus, the resulting predictive distribution $p(y_*|\mathbf{x}_*, \mathbf{X}, \mathbf{y})$ is a Gaussian distribution with mean and covariance defined as

$$\mu(\mathbf{x}_*, \mathbf{X}, \mathbf{y}) = K_{\mathbf{x}_*,\mathbf{X}}(K_{\mathbf{X},\mathbf{X}} + \sigma_n^2 I)^{-1}\mathbf{y}, \tag{6}$$

$$\sigma^2(\mathbf{x}_*, \mathbf{X}) = K_{\mathbf{x}_*,\mathbf{x}_*} - K_{\mathbf{x}_*,\mathbf{X}}(K_{\mathbf{X},\mathbf{X}} + \sigma_n^2 I)^{-1}K^T_{\mathbf{x}_*,\mathbf{X}}. \tag{7}$$

These mean and covariance terms are used in the next section to re-compute the predictions for instances near the classification boundary.

4. Online Domain Adaptation

Let S denote a classifier based on the sign of the prediction value from a function $f(\cdot)$, i.e.,

$$S(\mathbf{x}|f) := \text{sgn}(f(\mathbf{x})). \tag{8}$$

Let us assume that the probability of error of f is monotonically non-increasing with $|f(\mathbf{x})|$. In other words, the large prediction values (both positive and negative) are more likely to be correct than the small prediction values. This assumption further suggests that the classification label obtained by the classifier S for the points near the classification boundary (i.e., 0) may not be reliable. For the face detection problem, this assumption suggests that the pre-trained classifier can confidently accept unoccluded, in-focus, or "easy-to-detect" faces, and reject many non-face regions from a given image. This classifier is assumed to generate large prediction values for these easy acceptances and rejections. Consequently, the decision for the faces with low prediction values is assumed to be more difficult as compared to the decision for the regions with high prediction values.

Here we propose to update the scores for the data instances with low prediction values from the pre-trained classifier by encouraging the smoothness in the final prediction values. In other words, if two data points $\mathbf{x}_1$ and $\mathbf{x}_2$ are similar to each other, then the corresponding predictions $f'(\mathbf{x}_1)$ and $f'(\mathbf{x}_2)$ are encouraged to be similar to each other. To this end, we define a small margin around the classification boundary. The data points with prediction values outside this margin are used to learn a GPR model, which is used to reclassify the data points with prediction values lying inside the margin. Figure 3 illustrates the intuition for this classifier adaptation, and the formal description is included below.

Given $\epsilon > 0$, define the *in-margin set* $\mathbf{X}_m \subseteq \mathbf{X}$ such that $|f(\mathbf{X}_m)| < \epsilon$. Similarly, define the *out-of-margin set* $\mathbf{X}_o = \mathbf{X} \setminus \mathbf{X}_m$. For each element in the in-margin set, we compute the expected mean and covariance terms for the distribution of its prediction values using its similarity with the elements of the out-of margin set. Under the GPR model, this distribution of prediction values follow a Gaussian distribution. Hence we can compute the probability of the "true" function value being greater than a fixed value (the acceptance threshold in the classification setting). For instance, if the threshold is more than three standard deviation away from the mean, the probability that the prediction label is correct is greater than 99.7%. Therefore, we define the score updating function $f'(\mathbf{x})$ as

$$f'(\mathbf{x}) = \begin{cases} f(\mathbf{x}) & \text{if } |f(\mathbf{x})| > \epsilon \\ \mu(\mathbf{x}, \mathbf{X}_o, f(\mathbf{X}_o)) - c\sigma(\mathbf{x}, \mathbf{X}_o) & \text{otherwise,} \end{cases} \tag{9}$$

where c is a positive constant, and μ and σ are the mean (Equation 6) and covariance (Equation 7) terms of the predictive distribution. Figure 4 illustrates the shape of the predictive distribution $f'(\mathbf{x})$ for a toy example with one-dimensional $\mathbf{x}$. The value of c is set to 3 in our experiments.

In this formulation of the score updating function, including the term $\mu - c\sigma$ prevents accepting image patches with low confidence on the prediction labels. Similarly, we can include a symmetric term $\mu + c\sigma$ to prevent rejecting image patches with uncertain predictions. However, we exclude this term in our formulation for the following reason. The face detector implementation used in this work examines multiple overlapping image regions with small translations and scale changes. As a result, the rejection of some of these regions with uncertain predictions has little effect on the overall performance. On the positive side, the exclusion of this terms helps keeping the size of the out-of-margin set manageable for adaptation.

Finally, the classifier is defined as

$$S'(\mathbf{x}|f, \epsilon) \quad := \quad \text{sgn}(f'(x)). \tag{10}$$

In our face detection experiments, we use the noisy squared-exponential function as the covariance function,

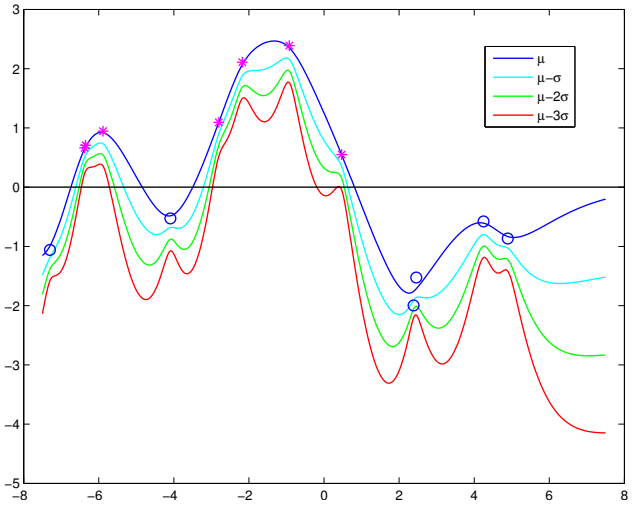


Figure 4. *Illustration of the score updating function (Equation 9).* The original scoring function $f(\mathbf{x})$ is shown as $*$ and o for the positive and negative examples respectively. Also shown are four curves $f'(\mathbf{x})$ for different choices of $c \in \{0, 1, 2, 3\}$. Thresholding these non-parametric curves with higher values of c results in a stricter acceptance criterion based on a higher minimum for the confidence in predictions.

i.e., $K^{\theta}_{\mathbf{x}_i,\mathbf{x}} = \nu^2 \exp\left(-\frac{\|\mathbf{x}_i - \mathbf{x}\|^2}{2l^2}\right) + \sigma^2_{gpn}\delta_{\mathbf{x}_i,\mathbf{x}}$, where ν and l refer to the weight and scale-length parameters of the squared-exponential function, and σ^2_{gpn} is the variance of the added noise when $\mathbf{x}_i$ and $\mathbf{x}_j$ are identical. Also, $\delta_{\mathbf{x}_i,\mathbf{x}}$ is a Kronecker delta. Hereafter, we use $\theta^T = [\nu, l, \sigma]^T$ to refer to the set of all of these three parameters of the covariance function.

In the next section, we present the details of incorporating this domain adaptation into a pre-trained cascade of classifiers. The resulting adaptive cascade of classifiers is used in the face detection experiments described in Section 6.

5. Adaptive Cascade of Classifiers

Viola and Jones [17] developed a cascade of AdaBoost classifiers to efficiently detect faces in an image. Their system examines image patches at different locations, sizes, and scales as follows. Given an image patch P, a set of features Φ are computed and fed into an AdaBoost-based binary classifier. Their detector uses n such classifiers to define a cascade that instantaneously rejects a patch that is rejected by any of the n classifiers. As a result, an image region is classified as a face region if and only if it is accepted by all of the classifiers in the cascade.

This detector has remained a top contender in face detection both in terms of accuracy and speed since its introduction in 2004. In this work, we use this system as the base

face detector. We apply the domain adaptation algorithm presented in the previous section to each of the classifiers in the Viola-Jones cascade (see Algorithm 1 for details).

Algorithm 1 Cascade of adaptive classifiers

Require: input X, classifier cascade $\{S\}_{1\ldots n}$, margin $\epsilon \geq 0$, covariance function $k(\cdot,\cdot)$

1: **for** $n = 1$ to N **do**
2: Let the stage classifier $S_n := \text{sgn}(f_n(x))$
3: $X_m \leftarrow \{x \in X | \, |f_n(x)| < \epsilon\}$
4: $X_o \leftarrow X \setminus X_m$
5: $Y_o \leftarrow f_n(X_o)$
6: $\theta^* \leftarrow \underset{\theta}{\text{argmax}} \; \log p(Y_o|X_o, \theta, k)$, where θ are the parameters of k
7: $\forall x \in X_o$, $f'_n(x) \leftarrow f_n(x)$
8: $\forall x \in X_m$, compute $f'_n(x)$ using Equation 9
9: $X \leftarrow \{x \in X | f'_n(x) > 0\}$
10: **end for**

For each test image, the parameters θ of the covariance function ($K^{\theta}_{\mathbf{x}_i,\mathbf{x}}$) are estimated by minimizing the negative log-likelihood of observing the data and prediction values for elements in the out-of-margin set X_o (step 6 of Algorithm 1). The minimization is performed using a conjugate gradient approach. In our experiments, we observed little dependency on the initialization of this minimization step.

The objective of the classifiers in the first few stages of the Viola-Jones cascade is to weed out only the image patches (e.g., regions with uniform intensity) that are highly unlikely to be face regions. The acceptance criteria for these stages are very relaxed, leading to large size of the out-of-margin sets for these stages. Since the computational cost of GPR varies as $O(n^3)$ with n being the size of the out-of-margin set, performing the proposed adaptation for these stages is expensive. Also, we found that the improvements in performance of these initial stages has little effect on the overall performance of the entire cascade. Therefore we skipped the adaptation steps for the initial few (ten) stages of the cascade in our experiments. For later stages of the cascade, the size of the out-of-margin set was found to be usually less than 50 (and less than 10 for the last few stages). Our unoptimized code for GPR took only a couple of milliseconds to process sets of such small size on a standard desktop computer. As a result, the running times of our cascade are similar to those of the original cascade.

6. Face Detection Experiments

Jain et al. [6] recently developed the FDDB benchmark[2] for evaluating the performance of face detection algorithms. This data set contains photographs from several news sources, and includes images of faces under very challenging, unconstrained environments. This collection has a total of 5171 faces in 2845 images. Jain et al. also specified an evaluation scheme based on computing two ROC curves using: (a) discrete score, and (b) continuous score. The discrete score is similar to the previous evaluations where each detection gets a binary match/non-match label. The continuous score associates a real-valued score with each detection based on the overlap between the detected and the annotated regions. As specified by Jain et al., we report the results for 10-fold cross-validation experiments, and use the FDDB evaluation software to generate the performance curves.

We use the OpenCV implementation of the Viola-Jones face detector as the base face detection algorithm, making it the natural baseline in our comparisons. In our comparisons, we also include Mikolajczyk et al.'s [10] variant of Schneiderman et al.'s method [13] for parts-based detection. These two systems have been considered the best-performing public implementations of face detection algorithms. We also include the curves for a recent approach by Subburaman et al. [14] since they show improvement in performance for a range of false positives. Kienzle et al.'s face detector [8] is also included in our experiments, although its performance was found to be very low. The performance curves for all of these approaches are shown in Figure 5.

To further study the effect of the choice of margin width ϵ, we compare the performance curves for different choices of ϵ in Figure 6. We found the performance to be stable around $\epsilon = 3$. We also experimented with multiple iterations of score updates for each of the classifiers in the cascade, where the updated scores obtained using our method are fed into the next iteration as the predictions from the pre-trained classifiers. For these experiments, we observed that the predictions converged within a few iterations.

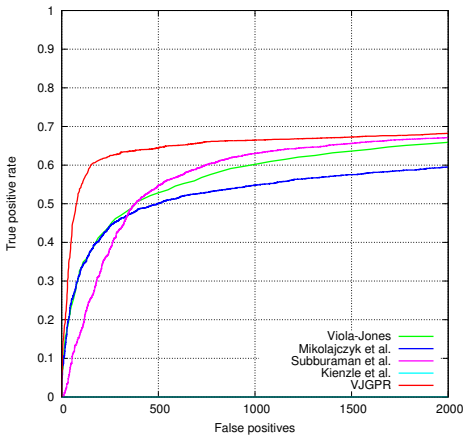
We also identified the situations where VJGPR consistently improves the detection results obtained by the Viola-Jones detector. Some examples of these situations are shown in Figure 7. Figure 8 shows some examples where VJGPR failed to improve detection performance.

To summarize the results, we have shown that simply by adapting a black-box classifier so that its outputs are smooth with respect to a new test set, we can dramatically improve its performance. Our proposal for adaptation is similar to work in a variety of fields. Next, we discuss these related approaches and distinguish our method from previous work.

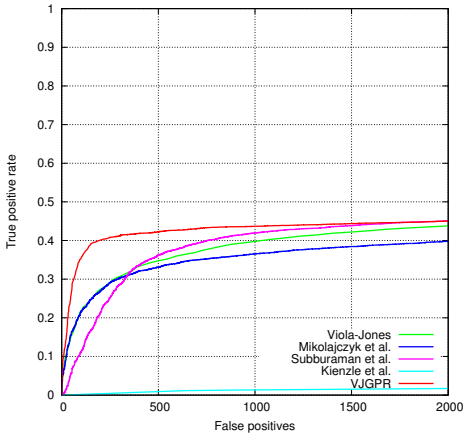
7. Other Related Work

Semi-supervised learning refers to the problem of learning from both labeled and unlabeled data. One common approach for handling unlabeled data is to construct a graph using pairwise similarities between both labeled and unlabeled training instances. The nodes corresponding to the

[2] http://vis-www.cs.umass.edu/fddb



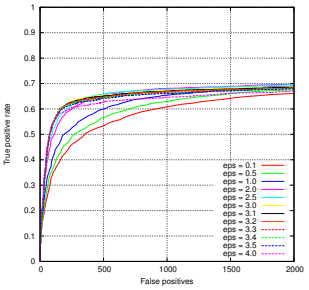
(a) Using discrete score



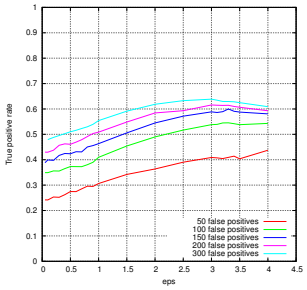
(b) Using continuous score

Figure 5. *Performance curves for different approaches*. The ROC curve for the proposed online adaptation approach (VJGPR) outperforms the curves for all of the other methods for the entire range of the choice of false positives. The improvement in performance is greater than 50% for small (< 200) number of false positives over the state-of-the-art on a very challenging data set. Note that the improvement based on the continuous score also suggests that the VJGPR detector is more accurate in determining the extent of the face regions as well.

labeled instances are annotated with the original labels, and *label propagation* [20] is performed to estimate the labels for unlabeled instances. In Zhu's *self-training* [20], a classifier is first trained using a small set of labeled examples and then the predictions from this trained model are used to re-train the classifier. However, in our approach, the original classifier is treated as a black-box and is *not* trained again.



(a) ROC curves



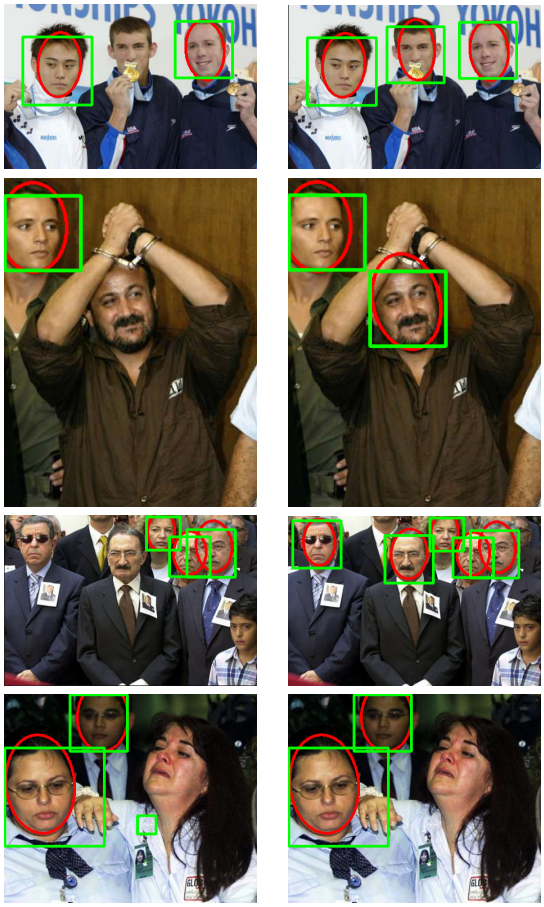
(b) Slices of ROC curves for different number of false positives

Figure 6. *Effect of the choice of the margin parameter* ϵ. As we increase the margin width ϵ from zero, the scores are updated for more data points resulting in an improvement in performance. Although after reaching a certain width, the performance starts to become worse since fewer instances are left in the confident set X_o to learn an effective model for online adaptation.

Another related line of research uses the unlabeled data to estimate the underlying data density and move the classification boundary out of regions with high data density. For example, Lawrence and Jordan [9] presented a null-category noise model to push the unlabeled data out of the margin; and Szummer et al. [16] included a regularization term based on a local estimate of the mutual information between the data and the label distributions to move the classification boundary out of the high density data regions. The latter approach, also referred to as *information regularization*, was extended by Corduneanu et al. [4] for semi-supervised learning.

Our approach uses the similarity between the data points to update the detection score of the data points for which the predicted score from the pre-trained detector is near the classification boundary. This update effectively sparsifies the distribution of data around the original classification boundary. While the intuition behind this procedure is similar to those of the above-mentioned methods, these semi-supervised learning approaches assume that both of the labeled and unlabeled data are sampled from an identical underlying distribution. As described in Section 1, this assumption does not hold true for our problem setting.

The problem formulation used in this paper is similar to the work in *domain adaptation*. In domain adaptation, labeled data from one or multiple "source" domains is used to train models to perform well on a different yet related "target" domain. Daumé and Marcu [5] approach this problem by modeling the data distribution for each of these domains as a mixture of a global and a domain-specific component. This global component is inferred from the data of



Viola-Jones detector Our adaptive detector

Figure 7. *Example images where our approach improves over the base detection algorithm.* (The detections are denoted by green rectangles, whereas the matched ground truth face annotation is denoted by red ellipses. These face detection results are obtained using systems with identical false positive rate.) The proposed algorithm adapts the base detector using the appearances of the detected faces in a given image. Without increasing the false positive rate, the resulting detector is made more robust to: occlusions on lips (the face in the middle in row 1), facial hair and self occlusions (the face on the right in row 2), and the presence of dark glasses and thick mustache (the two faces on the left in row 3). Note that due to the absence of similar easy-to-detect examples, the boy on the right-bottom corner in row 3 remains undetected by the adaptive detector as well. In the image shown in row 4, the adaptive detector got rid of the spurious detection on the body of the person on the right.

the source domain(s) and applied to the data of the target domain. Another approach to the domain adaptation problem employs models trained on the data from the source domain to label a subset of the unlabeled data from the unlabeled target domain, and re-trains the classifier on the combined labeled data set [18].

Viola-Jones detector Our adaptive detector

Figure 8. *Challenging examples.* (The detections are denoted by green rectangles, whereas the matched ground truth face annotation is denoted by red ellipses. These face detection results are obtained using systems with identical false positive rate.) The faces appearing in the image shown in the top row display extreme emotions, occlusions, and large variation in the head pose. Thus, there is little similarity in appearance among different faces in this image. As a result, our adaptive detector failed to detect several faces in this image. The middle row shows photographs of two different people acquired separately, but placed together to form a single image. Although our model is capable of handling multiple types of face appearance in a single image, if one of these types has only a few detections with low detection scores, the detections of this type may be removed as spurious detections by our detector. The bottom row shows an example image where there are no easy-to-detect faces present in the image. Clearly, the adaptive detector also fails to improve detection rate on these images.

Most of the work in domain adaptation (including the above two) suggests minimizing a convex combination of source and target empirical risk [3]. Thus the classifier needs to be re-trained (repeatedly) from scratch for every new domain. For face detection, we argue that the distribution of face appearances varies significantly from one image to another. Hence, in the domain adaptation setting, every image represents a new domain. Applying existing techniques for domain adaptation would therefore be prohibitively slow for our problem formulation.

Our work could also be interpreted as regularizing the output of a face detection algorithm on the data manifold. Belkin et al. [1] proposed smoothing a discriminative function by controlling the complexity of the learned classifier through the norm of the desired function in the correspond-

ing reproducing kernel Hilbert spaces. Although this *manifold regularization* framework provides useful insights into the relation between the hypotheses for the original detector and the adapted detector, the infeasibility of re-training the classifier for a new test image prevents us from building on this work as well. A similar argument holds true for the relevance of the previous work related to the analysis of *covariate shift* [2].

In another related work, Jain et al. [7] proposed an algorithm for query-dependent re-ranking of image search results using GPR. They used clicks associated with an image as weak supervision for learning a regression model. A similar kind of supervision is not available for face detection.

To the best of our knowledge, our work is the first to approach domain adaptation in a completely unsupervised and on-line setting. In other words, instead of training a new classifier from scratch for a new target domain, we adapt a classifier trained on a different source domain by encouraging smoothness of the output function. We present a simple yet effective method to perform this adaptation, and report state-of-the-art results in face detection using this approach.

8. Conclusion

We have shown that simply by adapting a black-box classifier so that its outputs are smooth with respect to a new test set, we can dramatically improve its performance. The performance gain we have achieved on the FDDB face detection benchmark is dramatic, especially in view of the fact that the Viola-Jones classifier has remained a top contender in face detection accuracy since its introduction in 2004. (We note that FDDB is particularly difficult, including many profile views and other faces which the best detectors currently miss.)

While it is certainly worth asking whether semi-supervised methods, despite their greater computational burden, could be applied in this scenario, several problems have kept us from pursuing this question. First, the original training data for the Viola-Jones classifier is proprietary and unavailable for us to use. Thus, we must already accept an alternative training set than the one used to train the original published classifier. Second, the details of training the original Viola-Jones classifier are not completely specified in the literature. We have had difficulty reproducing the original results on the classifier, even after contacting one of the original authors of the paper.

Without the original training data and without a clearly specified training algorithm, the task of applying a semi-supervised method is especially daunting. This perhaps makes our approach even more appealing, since there is no need for either the original data or the original algorithms. In future work, we hope to characterize the necessary and/or sufficient conditions for our approach.

9. Acknowledgments

We thank Allen Hanson and Sundararajan S. for invaluable discussions. This work was supported in part by National Science Foundation under CAREER award IIS-0546666.

References

[1] M. Belkin, P. Niyogi, and V. Sindhwani. Manifold regularization: A geometric framework for learning from labeled and unlabeled examples. *JMLR*, 2006. 583

[2] S. Bickel, M. Brückner, and T. Scheffer. Discriminative learning under covariate shift. *JMLR*, 2009. 584

[3] J. Blitzer, K. Crammer, A. Kulesza, F. Pereira, and J. Wortman. Learning bounds for domain adaptation. In *NIPS*, 2008. 583

[4] A. Corduneanu and T. Jaakkola. On information regularization. In *UAI*, 2003. 582

[5] H. Daumé III and D. Marcu. Domain adaptation for statistical classifiers. *JAIR*, 2006. 582

[6] V. Jain and E. Learned Miller. FDDB: A benchmark for face detection in unconstrained settings. Technical report, University of Massachusetts Amherst, 2010. 581

[7] V. Jain and M. Varma. Learning to re-rank: Query-dependent image re-ranking using click data. In *WWW*, 2011. 584

[8] W. Kienzle, G. H. Bakır, M. O. Franz, and B. Schölkopf. Face detection — efficient and rank deficient. In *NIPS*, 2005. 581

[9] N. D. Lawrence and M. I. Jordan. Semi-supervised learning via Gaussian processes. In *NIPS*, 2005. 582

[10] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. In *ECCV*, 2004. 581

[11] C. E. Rasmussen and C. K. I. Williams. *Gaussian Processes for Machine Learning*. The MIT Press, 2005. 579

[12] H. A. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *PAMI*, 1998. 577

[13] H. Schneiderman and T. Kanade. Probabilistic modeling of local appearance and spatial relationships for object recognition. In *CVPR*, 1998. 577, 581

[14] V. B. Subburaman and S. Marcel. Fast bounding box estimation based face detection. In *ECCV Workshop on Face Detection: Where we are, and what next?*, 2010. 581

[15] K.-K. Sung and T. Poggio. Example-based learning for view-based human face detection. *PAMI*, 1998. 577

[16] M. Szummer and T. Jaakkola. Information regularization with partially labelled data. In *NIPS*, 2002. 582

[17] P. A. Viola and M. J. Jones. Robust real-time face detection. *IJCV*, 2004. 577, 580

[18] D. Wu, W. S. Lee, N. Ye, and H. L. Chieu. Domain adaptive bootstrapping for named entity recognition. In *Empirical Methods in Natural Language Processing*, 2009. 583

[19] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical report, Microsoft Research, 2010. 577

[20] X. Zhu. *Semi-Supervised Learning with Graphs*. PhD thesis, Carnegie Mellon University, 2005. 582