Probabilistic Elastic Part Model for Unsupervised Face Detector Adaptation

Haoxiang Li, Gang Hua
Stevens Institute of Technology
Hoboken, NJ 07030
{hli18, ghua}@stevens.edu

Zhe Lin, Jonathan Brandt, Jianchao Yang
Adobe Systems Inc.
San Jose, CA 95110
{zlin, jbrandt, jiayang}@adobe.com

Abstract

We propose an unsupervised detector adaptation algorithm to adapt any offline trained face detector to a specific collection of images, and hence achieve better accuracy. The core of our detector adaptation algorithm is a probabilistic elastic part (PEP) model, which is offline trained with a set of face examples. It produces a statistically-aligned part based face representation, namely the PEP representation. To adapt a general face detector to a collection of images, we compute the PEP representations of the candidate detections from the general face detector, and then train a discriminative classifier with the top positives and negatives. Then we re-rank all the candidate detections with this classifier. This way, a face detector tailored to the statistics of the specific image collection is adapted from the original detector. We present extensive results on three datasets with two state-of-the-art face detectors. The significant improvement of detection accuracy over these state-of-the-art face detectors strongly demonstrates the efficacy of the proposed face detector adaptation algorithm.

1. Introduction

There are now a number of practical face detection solutions as well as publicly available face detectors [25, 26, 18, 20]. In general, appearance-based detectors outperform other peer solutions. In the appearance-based face detector, models are usually learned from a large set of positive and negative training images to capture the variations of faces while keeping the model discriminative. As a result, state-of-the-art face detectors are practical in a number of general scenarios.

However, a drawback of current methods is that they may not produce good enough results for a specific task, such as detecting faces in a specific collection of images. Since factors affecting face detection such as scale, location, pose, face expression, lighting conditions, etc. vary from task to task, it requires non-trivial engineering efforts to make a face detector work for extensive scenarios even with the cutting-edge algorithms. While the design of a general face detector is an important and attractive problem, to the end users how a face detector performs on their photos matters most. In this sense, how to adapt a general face detector to a specific task (*e.g.* to a specific image collection) is an important and realistic problem.

It can be straightforward to fit this detector adaptation problem into a domain transfer learning framework. For most of domain adaptation/transfer learning methods [6, 11, 14, 24, 19, 15, 21], it is assumed that there are few or no labels on the target domain while a large amount of labeled data exist in the source domain. However, for the face detector adaptation task it may not be possible to access the training data since state-of-the-art detectors are usually trained with massive data that cannot be easily transferred, or is not publicly available. In this paper, we treat the general face detector as a black box without access to its training data and approach to this problem with an online trained classifier with a novel probabilistic elastic part (PEP) representation.

Our approach consists of an offline phase and an online phase. In the offline stage, we train a PEP model, which is a local spatial-appearance feature based Gaussian mixture model, from a set of face examples. Then we utilize the PEP model to build the PEP representation for every candidate detection extracted by the face detector. We argue that the PEP representation statistically aligns faces and confines the comparison between two faces within locally corresponding face structures. Using the detection confidence scores from the general face detector, we pick up the detections with high confidence scores as positive candidates and regard detections with low confidence scores as negative candidates. Finally, over these positive and negative examples, with the PEP representation, we train a discriminative classifier online to induce a probability output to predict how likely a candidate detection is right. We set the probability output as the detection confidence score of the adapted detector.

Under this framework, we turn the detector adaptation problem into a binary classification problem. As in any image classification problem, the first step is to build an appropriate image representation [12, 1, 23]. Since human faces could show varied appearances due to factors such as expression, pose etc., we anticipate that an aligned visual representation for face images would be robust to these visual variations, as it enables local comparison within the same face structures. We approach this by building the PEP representation as a statistically aligned visual representation for face image with the help of an offline trained PEP model.

In our framework, we augment every local feature extracted from face image with its spatial location in the image. The PEP model formulated as a special Gaussian mixture model with spherical Gaussian components is trained over the spatial-augmented local features from the offline training faces. As a result, a weighted mixture component in the PEP model captures the appearance of a face structure (e.g. the nose) as well as the location of the structure on the face (e.g. center of the face). In this sense, given a face image as a set of spatial-augmented local features, to align the features to the PEP model is to align features to the face structures described by the mixture components.

We find this process in our framework is well-aligned to the working model of receptive field in the area of biology, by regarding a Gaussian component as a statistical receptive field (See Section 4.1). In brief, given a face image as a set of spatial-augmented local features, one weighted Gaussian mixture component picks one feature which *activates* it. And we concatenate all the *activation features* to build PEP representation of the face image. After that we train the discriminative classifier on these PEP representations.

In short, our contributions are three-fold: (1) we introduce a novel PEP representation for face image; (2) we apply the PEP representation in an unsupervised detector adaptation framework and demonstrate its effectiveness in improving the face detection performance; (3) empirically, we show that the proposed adaptation method improves two state-of-the-art face detectors on two personal photo albums and a standard face detection benchmark by a large margin.

This paper is organized as follows. In Section 2, we review related methods. In Section 3, we explain the unsupervised detector adaptation work-flow and we introduce the probabilistic elastic part model in Section 4. We then present our experimental results in Section 5 and conclude in Section 6.

2. Related Work

Our problem can be viewed as an instance of the domain adaptation problem [6, 11, 14, 24, 19, 15, 21]. Specially, we can formulate the target problem as one of adapting the general face detector pre-trained on a source domain to the testing photos which are from a different distribution as the target domain. One common approach to handle unlabeled data in unsupervised domain adaptation is to utilize the training labels in the source domain [19, 6]. Gopalan *et al.* [6] propose to learn the transformation between the source and target domain subspace to train a discriminative classifier from the projected training data in the target domain. Tang *et al.* [19] mix the confident detections in the target domain with training data in the source domain to train the classifier. Another line of related methods turn to estimate the labels of unlabeled data. It can be done by exploiting the spatio-temporal information in the video detector adaptation setting[15, 14]. In a surveillance camera environment labels can be estimated by building a generative model for the foreground and background through a long time adaptation [17]. These approaches are not applicable to our scenario because the pre-trained detector is treated as a black box without access to its massive training data and we aim at adapting it to a specific collection of photos, which do not have dense temporal correlation as in videos.

The more relevant works to ours is Jain *et al.* [11] and Wang *et al.* [21]. Jain *et al.* [11] reclassify detections near the decision boundary using a Gaussian process regression scheme in an unsupervised framework. A disadvantage of their work is that the Gaussian process regression model is trained repeatedly which can be computationally expensive, while in our method, the classifier is trained only once. Wang *et al.* [21] approach the pedestrian detector adaptation with an online non-parametric classifier on binary representations of candidate detections. Both theirs and our method try to explore the correlation within the testing photos to improve the detection. On one hand their method is proposed for pedestrian detector adaptation while ours are for face detector adaptation. On the other hand, our method with the PEP representation is stronger in addressing large intra-class (between different faces) variations. Considering its similar work-flow to ours, we compare with this method in our experiments and the results support this interpretation.

From another perspective, our work can be partially explained as a self-taught learning framework. As introduced in [16], a self-taught learning algorithm learns a higher-level representation from low cost unlabeled data and reinterprets images in the higher-level representation, resulting in improved classification accuracy. Similarly, in this paper a set of faces is used to learn the PEP model, which is subsequently leveraged to build a new representation, the PEP representation, for candidate detections and ultimately improve the final classification accuracy.

3. Unsupervised Detector Adaptation

General face detectors are typically not perfect, so they may have false positives and/or missed detections. By setting the detection threshold to a low level, we can achieve reasonable recall but the precision will be low. So, starting with a candidate set of detections with high recall but possibly low precision, our goal is to find a subset of de-

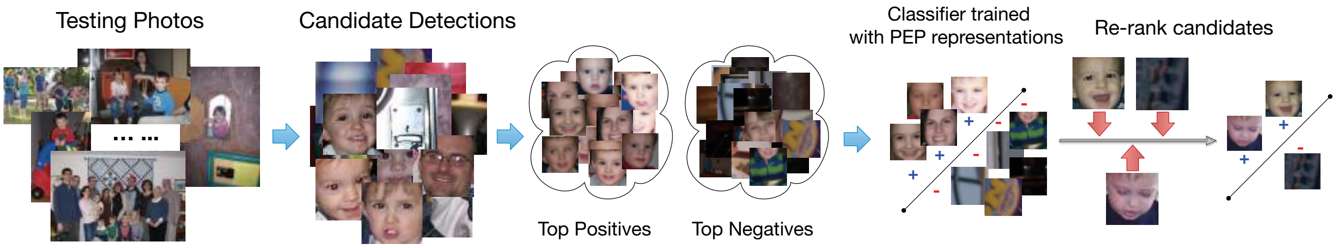


Figure 1. Our unsupervised detector adaptation work-flow

tections with high recall and precision. Formally, given a set of test images, we denote the extracted candidate detections from a general face detector as $T = \{t_1, t_2, \ldots, t_M\}$. Here, we set the detection threshold lower to retrieve many detections and thereby ensure a relatively high recall rate. We assume the face detector generates confidence scores for candidate detections, otherwise we will resort to the *activation score* (See Section 4.3) of the candidate detections over the probabilistic model trained offline as the detection confidence scores. For now, we assume each $t_i, i = 1, \ldots, M$ is associated with a detection confidence score $S = \{s_i, i = 1, \ldots, M\}$. Without losing generality, we assume that T and S are already sorted in descending order that $s_i > s_{i+1}, i = 1, \ldots, M-1$. Two thresholds are manually chosen here: λ_h and λ_l (these thresholds are stable properties of a given detector so they can be automatically estimated through cross-validation on a third dataset). λ_h denotes the threshold that detections with confidence score higher then λ_h is expected to be faces with high confidence (top positives). Similarly, candidates with confidence score lower than λ_l are supposed to be false alarms confidently (top negatives). Here we denote H as the top positives and N as the top negatives, i.e.,

$$\begin{aligned} H &= \{t_i : s_i > \lambda_h, i = 1, \ldots, M\} \\ N &= \{t_i : s_i < \lambda_l, i = 1, \ldots, M\}. \end{aligned}$$

We expect to have a balanced setting of H and N. So we keep $C = min(|H|, |N|)$ candidates from these two sets as the final top positives and negatives, i.e.,

$$\begin{aligned} H &= \{t_1, t_2, \ldots, t_C\}, \\ N &= \{t_{M-C+1}, t_{M-C+2}, \ldots, t_M\}, \\ C &< M/2 \end{aligned}$$

Intuitively, H are typical faces and N are typical false alarms. The better the face detector is the more reliable is this assumption. In our scenario, the face detector is treated as a black-box. In order to improve the performance of detection, we re-evaluate the detections by assigning a new confidence score to each candidate detection $\hat{S} = \{\hat{s}_i\}, i = 1, \ldots, M$. In the sense that the new confidence score $\hat{s}_i$ for t_i is defined as the probability that t_i is a face (positive) and we have H and N as estimated positive and negative examples. We can re-interpret the re-evaluation problem as binary classification with probability estimation. We can easily train a discriminative classifier online given the H and N as training data to predict the probability that t_i is positive.

The high-level work-flow is shown in Figure 1. In practice, detections, as regions on the original image, are extracted as "face" images, processed into sets of features and built into PEP representations.

4. Probabilistic Elastic Part Model

The effectiveness of the online discriminative classifier depends directly on the descriptive power of the PEP representations. In this paper, we firstly train a probabilistic model offline to assist the online representations building process.

We use a set of general faces to train the offline probabilistic model — a Gaussian Mixture Model (GMM) with spatial information encoded. The training corpus here can be any set of general faces which are very easy to collect from the Internet or existing face datasets. Faces are roughly aligned with the funneling method [7] to reduce pose variations. Then face images are processed into spatial-augmented features through a feature extraction pipeline. In this step, we extract dense patches over multiple scales and augment the local image patch descriptor with the spatial location of the image patch to build a set of spatial-appearance features for every face image as in [13].

To balance the strength of the spatial and appearance constraints, the Gaussian components are confined to be spherical [13]. So an offline trained GMM with K components is denoted as

$$P(\mathbf{f}|\Theta) = \sum_{k=1}^{K} \omega_k \mathcal{G}(\mathbf{f}|\mu_k, \sigma_k^2 \mathbf{I}), \tag{1}$$

where $\Theta = (\omega_1, \mu_1, \sigma_1, \ldots, \omega_K, \mu_K, \sigma_K)$; $\mathbf{I}$ is an identity matrix; ω_k is the mixture weight and $\mathcal{G}(\mu_k, \sigma_k^2 \mathbf{I})$ is a Gaussian distribution with mean μ_k and variance $\sigma_k^2 \mathbf{I}$.

Given a set of features extracted from training faces, we use Expectation-Maximization (EM) algorithm to learn the GMM parameters Θ [5]. Since the features have spatial information encoded, the GMM captures both the appearance and spatial distribution of face structures such as nose, left eye corner, right eye corner, etc. Intuitively, a weighted Gaussian mixture component describes the average appearance of a certain face structure and its location jointly.

4.1. PEP representation

As in the classical image classification pipeline, we first build a more descriptive and structured PEP representation for each face image. In our framework, the PEP representation statistically aligns a face image to the mixture components. In this sense, comparison between PEP representations consists of a set of local comparisons. Each local comparison is conducted between features aligned to a mixture component, or in other words, between features that describe the same face structure (See Figure 3).

We can explain the feature alignment process in a biological manner. In biology, a strong response of the neuron could suggest that a stimuli hit on the center receptive field [9]. In our framework, we define the response of a feature over a statistical receptive field as the probability of the feature over the mixture component since this value measures how likely the feature describes the expected face structure. Similar to the biological model, a strong response of a feature over the statistical receptive field means the feature "hit" the mixture component. We treat the one draws the highest probability over a mixture component as the *activation feature* of the component and say the feature *activates* the component. To align a face image (a feature set $\mathcal{F}$) to the GMM, we make each Gaussian mixture component find its *activation feature* from $\mathcal{F}$. Technically, given a set of features $\mathcal{F} = \{\mathbf{f}_i\}_1^{|F|}$, the k-th mixture model picked $\mathbf{f}_{g_k(\mathcal{F})}$ if

$$g_k(\mathcal{F}) = \arg\max_i \omega_k \mathcal{G}(\mathbf{f}_i|\mu_k, \sigma_k^2\mathbf{I}). \quad (2)$$

This process is illustrated in Figure 2. In Figure 3 we can easily observe how features are aligned to the mixture components. The effect of the spatial constraint is that face structures similar in appearance but different in location can be differentiated from each other. For example, in row (2) and (k) in Figure 3, while the left and right brows look similar in appearance, the spatial constraint prevents them from mixing together and the elasticity makes sure that they are not compared to other nearby structures. Furthermore, details of faces that describe the same face structure with a slight offset can be captured, as depicted in row (3) and (4).

Following Equation 2, K Gaussian components pick K *activation features* given a face image. As shown in Figure 3, the appearance part of the K *activation features* is concatenated in the sequence of Gaussian mixture model as the PEP representation of the face image $\mathcal{F}$, i.e., representation of $\mathcal{F}$ is

$$f^a(\mathcal{F}) = [\mathbf{f}_{g_1}\ \mathbf{f}_{g_2}\ \dots\ \mathbf{f}_{g_K}]. \quad (3)$$

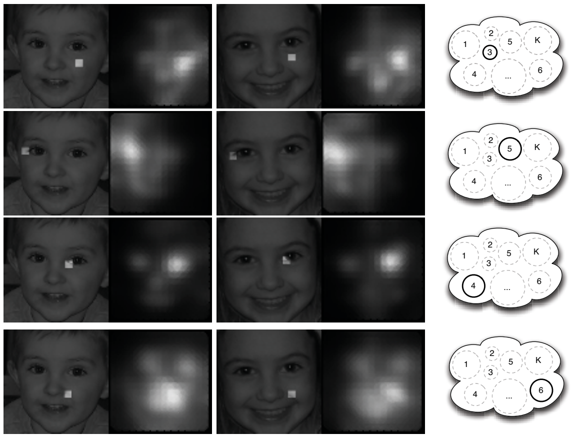


Figure 2. *Activation features* of two faces over some mixture components. Each row shows the activation process for a mixture component (illustrated in the right-most column); the highlighted part on the face represents the location of the *activation feature*, right to the face image we visualize the response of all features of a face over the mixture component; similar to a receptive field the lightest point on the response visualization locates the *activation feature* on the face image.

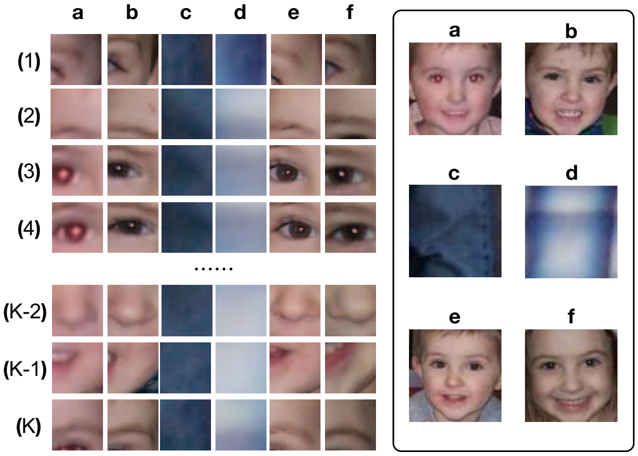


Figure 3. PEP representations of six images (we use a larger patch size for better visualization) built by a K components GMM. Each column is the PEP representation for a face (showing the appearance part), each row shows the *activation features*. We can see (across the row) element-wise comparison between PEP representations actually consist of local comparison within the same face structure.

Note that we removed the encoded spatial information from features $\{\mathbf{f}_{g_i}\}$ in formulating the PEP representation and the number of *activation features* is always the same to the size of GMM. That is, the same GMM produces unified representation for feature sets of different size.

4.2. Online Classifier

Once we build PEP representations for all candidate detections, consider H and N as sets of positive and negative

(a) *E-Album* (b) *G-Album* (c) *FDDB*

Figure 4. Example photos in the *E-Album*, *G-Album* and *FDDB*

examples respectively, we train a Support Vector Machine (SVM) as an online discriminative classifier. The SVM with probability estimation support [22] is trained with Gaussian Radial Basis Function (RBF) kernel. Applying the trained classifier to every candidate detection, a new detection confidence score $\hat{s}_i$ is computed for t_i.

4.3. Activation Score

An optional step in the proposed framework is to calculate the *activation score* as the initial confidence score in case the general face detector does not provide a detection confidence score. We re-use the offline trained GMM in a generative manner to calculate the *activation score* $S^a(t_i)$ for a candidate detection t_i, which is defined as the total response of all the *activation features*, i.e.,

$$S^a(t_i) = \sum_{k=1}^{K} log \max_{i} \omega_k \mathcal{G}(f_i|\mu_k, \sigma_k^2\mathbf{I}), \tag{4}$$

where $\mathcal{F} = \{\mathbf{f}_i\}_1^{|F|}$ is the set of features of detection t_i.

5. Experiments

We verified the proposed method with the Viola-Jones (VJ) detector [20] and the XZJY detector [18], over the *E-Album* [3], *G-Album*[4] and *FDDB* face detection benchmarks [10]. We also tested the Wang *et al.* [21] detector adaptation algorithm over the albums for comparison. *E-Album* contains 108 photos with 145 labeled faces; *G-Album* contains 512 photos with 873 labeled faces; *FDDB* contains 5171 faces in 2845 images taken from unconstrained environment.

We explore the factors which could influence the final performance such as, the top positive and negative thresholds and number of mixture components K. To demonstrate the statistical alignment introduced by PEP representation contributes to performance improvement, we design a baseline experiment by concatenating the local features directly as the face representation. The experimental results strongly support the proposed method and suggest that the PEP representation is a very effective visual representation for face image.

One of our motivations is to exploit within dataset correlation to help improve face detection. In this sense, our method is more suitable for detector adaptation on personal albums, since the same person might appear in many photos within an album. *E-Album* has stronger in-dataset correlation since the number of different individuals in this album is limited, in other words face of the same person appear many times in *E-Album*. Compared to the albums, the *FDDB* benchmark contains far fewer repeated faces of the same person, so that it shows larger in-class variations. However we observe that we can still boost the face detector performance by a large margin since the PEP representation enlarges the gap between faces and non-faces.

In practice, our method is quite efficient. Comparing to the detection time of state-of-the-art detectors such as the XZJY detector, which can take 10 seconds per image on the albums, the time taken by our adaptation process is minor. We extract SIFT features on the candidate detections instead of the whole image, and it takes 0.8 seconds for one candidate detection. After the feature extraction step, on our machine [1], it takes 0.01 seconds for one image on average. For example, it takes 14 seconds for *G-Album* including the PEP representation building, online classifier training and re-ranking of 1428 candidate detections. On the *E-Album*, it takes 3.5 seconds for the 271 candidates.

5.1. Settings

We use the OpenCV implementation of Viola-Jones face detector [2] and the XZJY face detector [18] which achieved the state of the art results on the FDDB and UCI datasets.We train the PEP offline with face images from the face dataset Labeled Face in the Wild [8], in which faces are roughly aligned with the funneling method. In the online classifier training stage, we use the LibSVM [2] implementation [3]. We evaluate the results with the Receiver operating characteristic (ROC) curves, where the y-axis denotes the true positive rate or recall rate, the x-axis is the total number of false alarms ($number\ of\ detections * (1 - precision)$). A good face detector should have high true positive rate at a low false alarm level. For the specific parameters, we resize the detections into 150 × 150 images and densely extract 128-dimensional SIFT features in a 8 × 8 sliding window with 8-pixel spacing in 3 scales with a scaling factor 0.9. In the sliding window its two-dimensional locations are scaled by 2 and concatenated to the extracted SIFT feature as the spatial-augmented feature, resulting in a total of 130 dimensions.

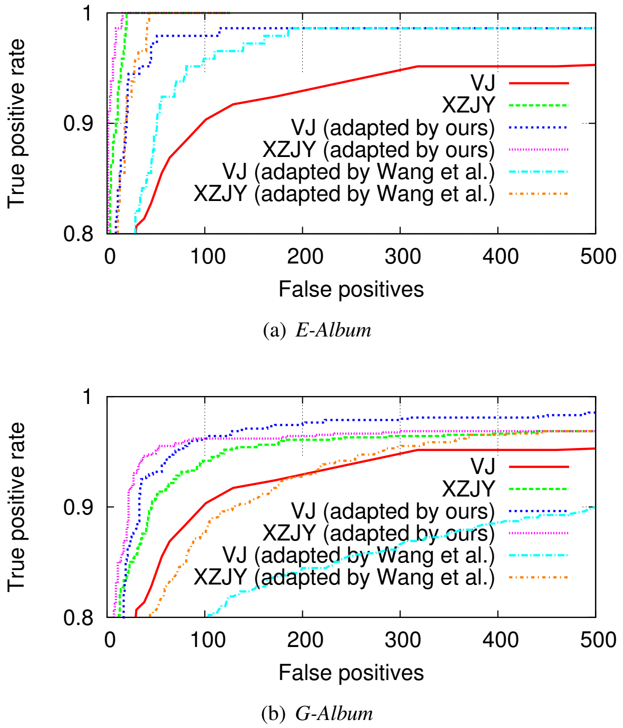
5.2. Results on person photo albums

As shown in Figure 5 the proposed method improves the performance of the Viola-Jones (VJ) detector and the XZJY detector by a large margin and outperforms the Wang *et al.* detector adaptation algorithm significantly. On *E-Album*,

[1] Intel(R) Xeon(R) E5645 12 Cores

[2] We simply choose the number of neighbor rectangles around a candidate detection as the detection confidence score.

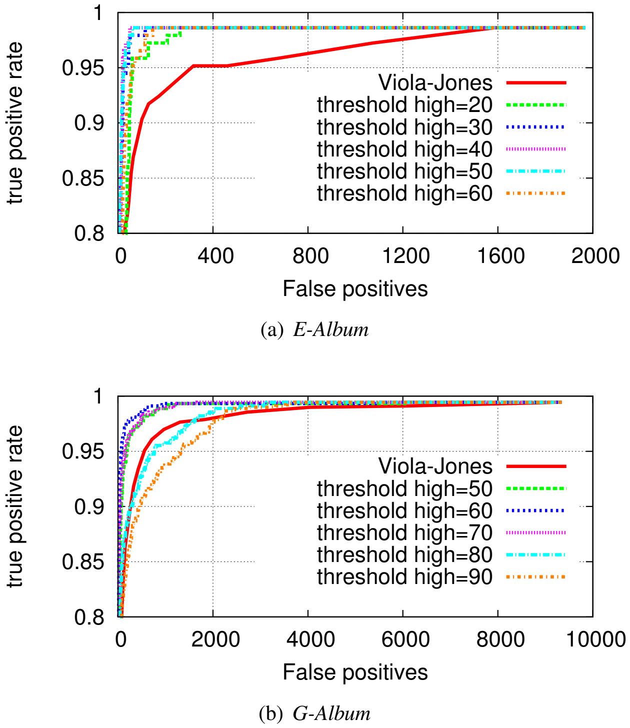
[3] The LibSVM provides the implementation of probability estimation introduced in [22]



(a) *E-Album*

(b) *G-Album*

Figure 5. ROC Curves over *E-Album*, *G-Album*



(a) *E-Album*

(b) *G-Album*

Figure 6. ROC Curves over *E-Album* and *G-Album* for different λ_h

Wang *et al.*'s method is able to effectively exploit intra-dataset correlations to improve accuracy. However, our proposed method consistently improves accuracy by a much greater margin. On G-Album, the performance of Wang *et al.*'s method actually drops relative to the non-adapted detector. However in this case, our proposed method improves performance again significantly. One possible explanation of the drop of Wang *et al.*'s performance in this case is that their representation is unable to enlarge inter-class variance in the presence of greater intra-class variance.

Influence of thresholds The thresholds λ_h and λ_l determine the number of top positive and negative examples for online classifier training. To investigate the sensitivity of the algorithm to these threshold values, we vary the thresholds of the detector to get the H and N. The lower boundary of the Viola-Jones detector confidence score in our case is fixed as 1 by design while the upper boundary can be very large, we fix $\lambda_l = 3$ and alter λ_h to explore how our method performs with respect to different high thresholds λ_h [4].

The experimental results are shown in Figure 6. We observe our method is robust to the choice of λ_h within a reasonable range.

Influence of the PEP We tested 128, 256, 512 and 1024 mixture components for the PEP and compared the overall performance. We observe that generally more mixture components make the performance more stable. As long as we have enough mixture components (256 in our case), the performance is robust to the choice of this parameter, as shown in Figure 7.

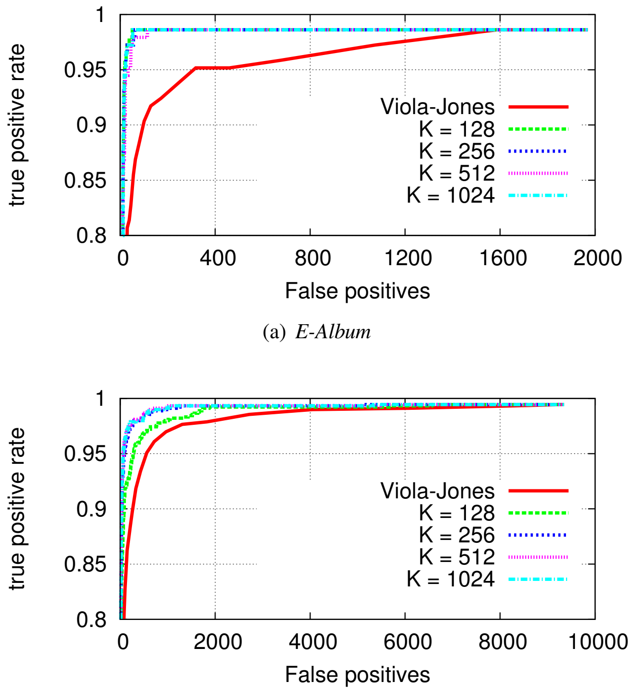
Concatenated SIFT as baseline We design this baseline experiment to show that the PEP representation is important in the performance improvement. For each face image, the baseline representation is built by concatenating all SIFT features from left to right, top to bottom, while in the PEP representation a set of features (possibly with duplicates) is selected by mixture components and concatenated. As shown in Figure 8, the contribution of PEP representation in our detector adaptation framework is significant.

5.3. Results on FDDB dataset

FDDB is a dataset designed as a benchmark for face detection algorithms [10]. Images in this dataset are extracted from news articles and display large variations in pose, background and appearance. Several existing face detection algorithms are tested on the FDDB dataset and the results are published on the website. The evaluation procedure is standardized and researchers are expected to use the same evaluation program to report the results.

Since this dataset provides images in 10 separate folds. We perform two types of adaptation on *FDDB*. The first one is the *all-folds* adaptation where all images in the 10 folds are put together and we train only one online classifier for all 10 folds. The second type is *fold-by-fold* adaptation in

[4] We make the cascade stages thresholds lower for a higher recall



(a) *E-Album*

(b) *G-Album*

Figure 7. ROC Curves over *E-Album* and *G-Album* for with different number of mixture components

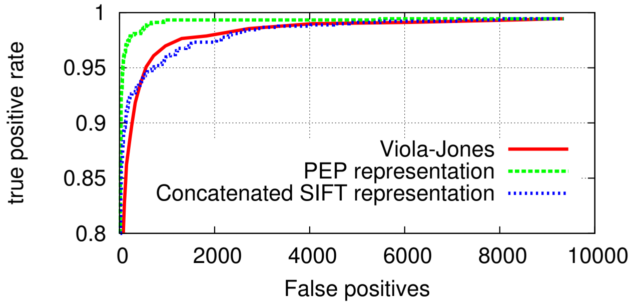
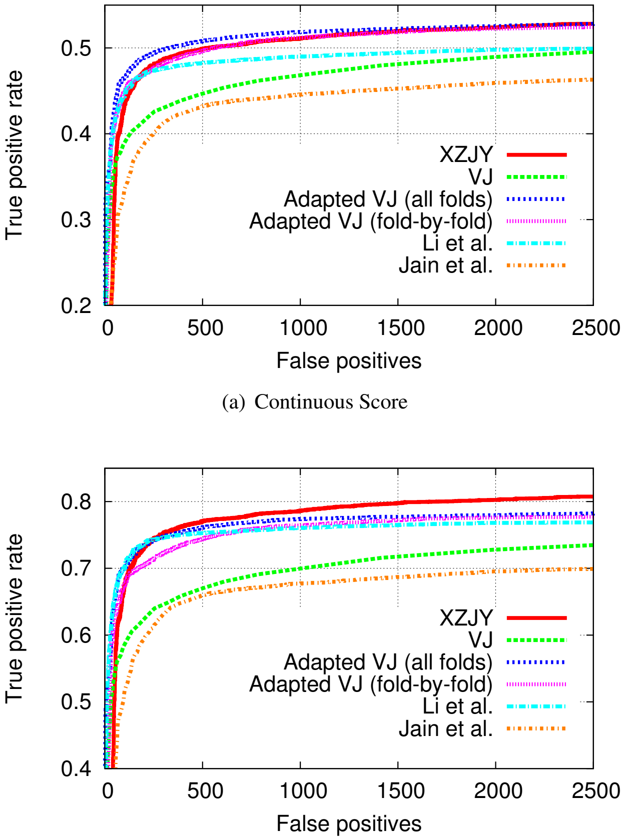


Figure 8. ROC Curves over *G-Album* for representation comparison

which we repeat the proposed method on each fold separately and have 10 different online classifiers. We observe in the experiments the *all-folds* way achieve a better performance since with the same threshold more top positive and negative examples can be obtained for the online classifier training.

As shown in Figure 9, the result from our method adapting the OpenCV's frontal face detector achieved very competitive results to the state-the-art detectors [5]. We even outperform state-of-the-arts detectors with non-trivial margin in the continuous score evaluation.



(a) Continuous Score

(b) Discrete Score

Figure 9. Performance comparison on *FDDB* by adapting Viola-Jones face detector

Another interesting question could be whether we could boost a very strong face detector on *FDDB* which is of less intra-class correlation and larger inter-class variance. As shown in Figure 10, using the XZJY detector which is the current state-of-the-art on the *FDDB* as the starting point, after adaptation, the detection performance is improved by a measurable margin.

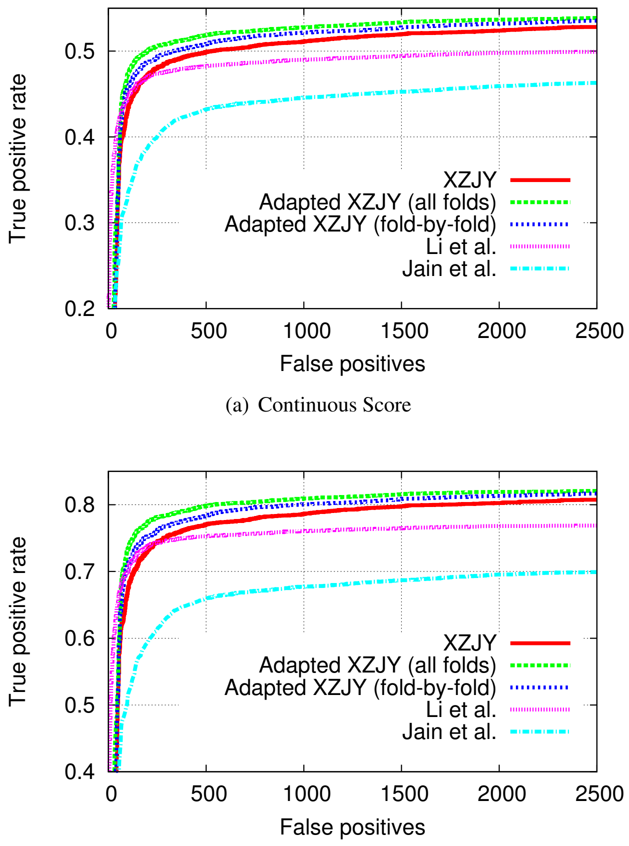
6. Conclusion

In this paper, we propose an unsupervised detector adaptation framework with an effective PEP representation for face images. We verified the proposed framework with two state-of-the-art face detectors on two person albums and a challenging unconstrained face detection benchmark, and demonstrate that the proposed method can improve the two general face detectors on a specific task by a significant margin.

Acknowledgement

This work is partly supported by US National Science Foundation Grant IIS 1350763, China National Natural Sci-

[5] We are using the up-to-date OpenCV implementation of the Viola-Jones detector which in our case outperformed what the author of FDDB reported a lot.



(a) Continuous Score

(b) Discrete Score

Figure 10. Performance comparison on *FDDB* by adapting the XZJY face detector

ence Foundation Grant 61228303, GH's start-up funds from Stevens Institute of Technology, a Collaboration Research Gift Grant from Adobe System Incorporated, a gift grant from NEC Labs American, and a Google Research Faculty Award.

References

[1] Y.-L. Boureau, F. Bach, Y. LeCun, and J. Ponce. Learning mid-level features for recognition. In *CVPR*, 2010. 1

[2] C.-C. Chang and C.-J. Lin. LIBSVM: A library for support vector machines. *ACM TIST*, 2011. 5

[3] J. Cui, F. Wen, R. Xiao, Y. Tian, and X. Tang. Easyalbum: an interactive photo annotation system based on face clustering and re-ranking. In *CHI*, 2007. 5

[4] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008. 5

[5] J.-L. Gauvain and C.-H. Lee. Maximum a posteriori estimation for multivariate gaussian mixture observations of markov chains. *IEEE Transactions on Speech and Audio Processing*, 1994. 3

[6] R. Gopalan, R. Li, and R. Chellappa. Domain adaptation for object recognition: An unsupervised approach. In *ICCV*, 2011. 1, 2

[7] G. Huang, V. Jain, and E. Learned-Miller. Unsupervised joint alignment of complex images. In *ICCV*, 2007. 3

[8] G. B. Huang, M. Ramesh, T. Berg, and E. Learned-Miller. Labeled faces in the wild: a database for studying face recognition in unconstrained environments. Technical report, University of Massachusetts, Amherst, 2007. 5

[9] D. H. Hubel. The visual cortex of the brain. In *Scientific American*, 1963. 4

[10] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, 2010. 5, 6

[11] V. Jain and E. G. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In *CVPR*, 2011. 1, 2

[12] Y. Jia, C. Huang, and T. Darrell. Beyond spatial pyramids: Receptive field learning for pooled image features. In *CVPR*, 2012. 1

[13] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic matching for pose variant face verification. In *CVPR*, 2013. 3

[14] V. Nair and J. J. Clark. An unsupervised, online learning framework for moving object detection. In *CVPR*, 2004. 1, 2

[15] A. Prest, C. Leistner, J. Civera, C. Schmid, and V. Ferrari. Learning object class detectors from weakly annotated video. In *CVPR*, 2012. 1, 2

[16] R. Raina, A. Battle, H. Lee, B. Packer, and A. Y. Ng. Self-taught learning: transfer learning from unlabeled data. In *ICML*, 2007. 2

[17] P. M. Roth, S. Sternig, H. Grabner, and H. Bischof. Classifier grids for robust adaptive object detection. In *CVPR*, 2009. 2

[18] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *CVPR*, 2013. 1, 5

[19] K. Tang, V. Ramanathan, L. Fei-Fei, and D. Koller. Shifting weights: Adapting object detectors from image to video. In *NIPS*, 2012. 1, 2

[20] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 1, 5

[21] X. Wang, G. Hua, and T. Han. Detection by detections: Non-parametric detector adaptation for a video. In *CVPR*, 2012. 1, 2, 5

[22] T.-F. Wu, C.-J. Lin, and R. C. Weng. Probability estimates for multi-class classification by pairwise coupling. *JMLR*, 2004. 5

[23] J. Yang, K. Yu, Y. Gong, and T. Huang. Linear spatial pyramid matching using sparse coding for image classification. In *CVPR*, 2009. 1

[24] C. Zhang, R. Hamid, and Z. Zhang. Taylor expansion based classifier adaptation: Application to person detection. In *CVPR*, 2008. 1, 2

[25] C. Zhang and Z. Zhang. A survey of recent advances in face detection, 2010. 1

[26] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 1