Efficient Boosted Exemplar-based Face Detection

Haoxiang Li[†], Zhe Lin[‡], Jonathan Brandt[‡], Xiaohui Shen[‡], Gang Hua[†]

[†]Stevens Institute of Technology
Hoboken, NJ 07030
{hli18, ghua}@stevens.edu

[‡]Adobe Research
San Jose, CA 95110
{zlin, jbrandt, xshen}@adobe.com

Abstract

Despite the fact that face detection has been studied intensively over the past several decades, the problem is still not completely solved. Challenging conditions, such as extreme pose, lighting, and occlusion, have historically hampered traditional, model-based methods. In contrast, exemplar-based face detection has been shown to be effective, even under these challenging conditions, primarily because a large exemplar database is leveraged to cover all possible visual variations. However, relying heavily on a large exemplar database to deal with the face appearance variations makes the detector impractical due to the high space and time complexity. We construct an efficient boosted exemplar-based face detector which overcomes the defect of the previous work by being faster, more memory efficient, and more accurate. In our method, exemplars as weak detectors are discriminatively trained and selectively assembled in the boosting framework which largely reduces the number of required exemplars. Notably, we propose to include non-face images as negative exemplars to actively suppress false detections to further improve the detection accuracy. We verify our approach over two public face detection benchmarks and one personal photo album, and achieve significant improvement over the state-of-the-art algorithms in terms of both accuracy and efficiency.

1. Introduction

Researchers in the area of face detection have made great progress over the past decades. However, under uncontrolled environment, large variations in pose, illumination, expression and other appearance factors tend to severely degrade the accuracy of the state-of-the-art detectors, such as faces in the images in Figure 1 [1].

An ensemble of feature-based weak classifiers trained using the Adaboost framework has become the *de facto* standard paradigm for face detection [21]. Combined with the attentional cascade, the feature-based method supports very fast face detection. Notwithstanding its great success, the limited descriptive power of the feature-based weak classifiers hinders it to handle all the challenging visual variations in unconstrained face detection.

Figure 1. Hard cases under uncontrolled environment: extreme poses could lead to missed detections while cluttered background could lead to false detections.

Recent work [25, 4] proposed to use part based models to address this problem, which benefits from the fact that the part of an object often presents less visual variations, and the global appearance variation of the object may be conveniently modeled by the flexible configuration of different parts. However, the modeling of each part is still based on the low level descriptors, which faces the same issue as the previous feature-based holistic face detectors [21].

Recently an exemplar-based method combining fast image retrieval techniques demonstrated great potential to overcome these challenges in face detection [19]. In contrast to the low-level feature-based approach, exemplars themselves, serve as high-level "features" that are more discriminative for face detection.

Unlike the standard Viola-Jones Adaboost [21] paradigm, in which either a feature or a model works as a scanning-window classifier, the exemplar-based approach works with a generalized Hough voting based method [9, 18] to produce a voting map to locate faces. Hence the exemplar-based method bypasses the scanning window paradigm and avoids the training of the attentional cascade which is a non-trivial process [24].

[1]The results are from the Zhu et al. [25] detector (left) http://www.ics.uci.edu/~xzhu/face/ and Shen et al. [19] detector (right) http://users.eecs.northwestern.edu/~xsh835/CVPR13Sup.zip.

The exemplar-based method [19] supports a more convenient training procedure and results in state-of-the-art face detection performance. Nevertheless, we argue that it is still impractical in terms of memory efficiency and detection speed. The main challenges of applying the exemplar-based method in real-world face detection applications are 1) it requires a large face database for detection, the size of which is the bottleneck in terms of space and time; 2) having more face images in the exemplar database introduces redundancy and potentially tends to produce false alarms in the presence of highly cluttered backgrounds.

To overcome these hurdles we propose an efficient boosted exemplar-based face detection approach which uses the RealAdaboost framework [16] to select the most discriminative exemplars. Hence the size of the exemplar database can be largely reduced by eliminating unnecessary redundancy while the detection accuracy is further improved. Additionally, we generalize the concept of exemplar to include non-face images as *negative exemplars*. The discriminatively selected negative exemplars assign negative confidence scores to the cluttered backgrounds which effectively suppress many of the false alarms.

In summary, our contributions in this paper are threefold: 1) we present an efficient boosted exemplar-based face detector by combining the flexibility of exemplar-based face detection and the discriminative power of the RealAdaboost algorithm with two additional efficiency enhancements; 2) we propose to build negative exemplars to actively suppress potential false detections; 3) we achieve new state-of-the-art detection performances on three datasets, including two challenging public face detection benchmarks and a personal photo album.

2. Related Work

Face detection has been an active research area for decades. We refer readers to a more comprehensive review of previous work in [24], and focus on the methods technically relevant and conceptually similar to ours.

Ever since Viola-Jones's seminal work [21], Adaboost is widely adopted in the area of face detection following a standard cascade paradigm. While most previous work along this line of research combined Adaboost with variations of Haar-like features [21] or more advanced features [22, 6], we rarely see literature applying Adaboost with exemplar-based method for face detection.

In related research areas, the work presented in [17, 23, 13] are conceptually related to our approach as they combined the boosting framework with discriminatively trained exemplar classifiers. In [17, 23, 13], the discriminative exemplar classifier is trained using holistic descriptors and the classifier responses on the detection windows are used as features in the boosting framework. With the sliding window paradigm where the responses need to be exhaustively computed between every detection window and every exemplar, this design can be computationally expensive especially when training with a large candidate exemplar pool or combining with a higher dimensional descriptor.

Comparing to this scheme, the main uniqueness of our work is the different nature of exemplars. In contrast to the holistic descriptors based exemplar, we use bag of localized visual words as the exemplar which works with a generalized Hough voting based method. Besides the fact that the voting based way avoids the sliding window paradigm, it further enables us to make use of fast retrieval techniques to jointly compute the similarities between one detection window and all the exemplars with the inverted file. Hence our approach is more efficient and scalable.

Technically, the most relevant work to ours is Shen et al.[19] which proposed the exemplar-based method with an efficient image retrieval technique for face detection. Their detector achieved state-of-the-art accuracy. However, the highly accurate detection performance required a large database of over 18,000 exemplars, the size of which hinders it to be a practical face detector in a lot of real-world applications. Furthermore, exemplars in their work are not discriminatively selected to optimize the detection accuracy and the similarity scores generated by different exemplars are actually in different ranges. As a result, the aggregated similarity scores over the exemplar database are suboptimal to differentiate faces from the background.

We follow the exemplar-based face detection paradigm. Nevertheless, in our framework, exemplars are trained discriminatively and assembled into a strong face detector through the RealAdaboost framework. Only the most effective exemplars are kept for detection. Hence, the detection performance is largely improved over the previous work while keeping a much more compact exemplar database. In our experiments, we show that on the Face Detection Data Set and Benchmark (FDDB) our detector with only 500 exemplars outperformed the state-of-the-art by a large margin. In contrast, the number of exemplars used in the previous exemplar-based face detector [19] is 18,486. More importantly, since our framework is more flexible, exemplars in our framework are no longer limited to be face images. Including non-face images as negative exemplars can further help differentiate faces from the cluttered background. As far as we know, no previous work proposed to incorporate negative exemplars in face detection.

3. Boosted Exemplar-based Face Detection

In this section, we will firstly discuss how a single exemplar works in our framework. We then describe how to train the boosted exemplar-based face detector through the RealAdaboost algorithm. Finally, we discuss the motivation of negative exemplars and two enhancements to help improve the detection efficiency.

3.1. Single Exemplar as Weak Detector

The exemplar in our framework refers to an image represented as a bag of localized visual words and a domain-partitioned weak classifier. The bag of visual words with the spatial locations produce a voting map over the target image while the domain-partitioned weak classifier gives real-value confidence ratings over every pixel location on the voting map. This procedure is shown in Figure 2.

In the training stage, given an image as a candidate exemplar, we resize the image into a fixed size and process them into the bag-of-visual-words representations by densely extracting feature descriptors and quantizing the descriptors over a pre-trained vocabulary. In the meantime, the spatial locations of the quantized descriptors are also kept to conduct the generalized Hough voting. Once the image is selected into the database, the bag-of-visual-words and associated spatial locations will be built into the inverted file index for efficient retrieval in the testing stage. We refer the readers to [9, 18, 19] for more details.

In the testing stage, the target image is processed through the same feature extraction and descriptor quantization pipeline and the exemplar is used to conduct generalized Hough voting over the testing image. The result will be a voting map in which each point represents the similarity score between the exemplar and the image subregion centered at that pixel location. Since the voting map locates the occurrences of this particular exemplar instead of the actual faces, it is not completely aligned with the goal of face detection if directly using the voting map to locate faces. Hence, the domain-partitioned weak classifier is further applied to the voting map to generate the confidence map. The domain-partitioned classifiers are trained discriminatively to differentiate faces from non-faces. As a result, the confidence map implies potential face centers and is more suitable for face detection.

Formally, given a testing image, for the i-th exemplar e_i, at point p on the voting map, the similarity score is

$$S(p, e_i) = \sum_k \sum_{\substack{f \in R(p), g \in e_i \\ \omega(f)=\omega(g)=k \\ ||\mathbf{T}(L(f))-L(g)||<\epsilon}} \frac{idf^2(k)}{tf_{R(p)}(k) tf_{e_i}(k)}, \quad (1)$$

where $R(p)$ is the image subregion centered at p, f and g are local features quantized into the k-th visual word in a pre-trained vocabulary. $idf(k)$ is the inverse document frequency of k and $tf_{R(p)}(k)$, $tf_{e_i}(k)$ is the term frequencies of k in $R(p)$ and e_i, respectively. A spatial constraint is enforced over the spatial locations $L(f)$ and $L(g)$ of f and g respectively through a transformation $\mathbf{T}$. In this paper, $\mathbf{T}$ represents only translation and scale change (please refer to [19] for details).

The voting map is then processed by the associated domain-partitioned classifier $h_i(R(p))$. At point p we have

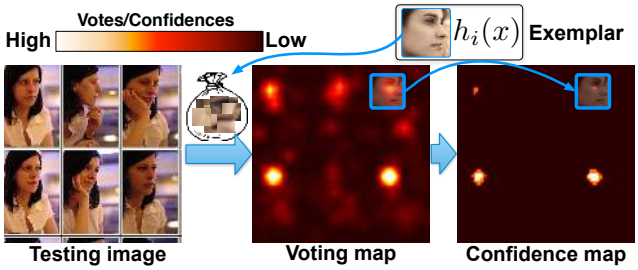


Figure 2. A single exemplar as a weak face detector; voting map and confidence map are shown as heat-maps.

confidence score

$$h_i(R(p)) = c_{i,m} \text{ if } S(p, e_i) \in \mathcal{X}_m, \quad (2)$$

where $\{c_{i,m}\}_{m=1}^{M} \in \mathcal{R}$ are the confidence ratings, $\{\mathcal{X}_m\}_{m=1}^{M}$ are disjoint partitions of the range of the similarity scores, i.e., $\mathcal{X}_1 \cup \mathcal{X}_2 \cup \cdots \cup \mathcal{X}_M = \mathcal{R}$. In this paper, we adopted a fixed partition for training convenience through empirical evaluation. The confidence rating for each partition is decided statistically in the training stage as described in the following section.

3.2. Ensemble Exemplars

Our goal is to select and assemble

$$H(x) = \sum_{t=1}^{T} \hat{h}_t(x), \quad (3)$$

where $\hat{h}_t(x)$ is the t-th selected exemplar and $H(x)$ gives the confidence of image subregion x being a face.

To minimize the error of $H(x)$ over the training data, we take an iterative process to build $H(x)$ following the RealAdaboost framework[16],

Given a set of training data $\{(x_n, y_n), n = 1, \ldots, N\}$, where $\{x_n\}$ denotes subregions of training images, label $y_n \in \{+1, -1\}$ denotes whether x_n is a face (positive) or non-face (negative) respectively. N^+ and N^- are the number of positive and negative training samples respectively. We have K exemplars $e_1, e_2, \ldots, e_K$ for selection and each e_i is associated with a domain-partitioned weak classifier $h_i(x)$ of M_i partitions, where $i = 1, \ldots, K$.

Initially, the training samples are assigned weights as

$$\omega_n^{(1)} = (1 + e^{y_n log(\frac{N^+}{N^-})})^{-1}. \quad (4)$$

Then T iterations are conducted to train a T stages boosted exemplar-based face detector.

At the iteration t, the training error for each exemplar e_i given the current training sample weights is calculated as $Z_i = 2\sum_m \sqrt{W_{i,m}^+ W_{i,m}^-}$, where $W_{i,m}^+$ and $W_{i,m}^-$ are the sum of weights of samples falling into the m-th partition of $h_i(x)$, *i.e.* let $\Omega_{i,m}^+ = \{n | S(x_n, e_i) \in \mathcal{X}_{i,m} \wedge y_n = +1\}$,

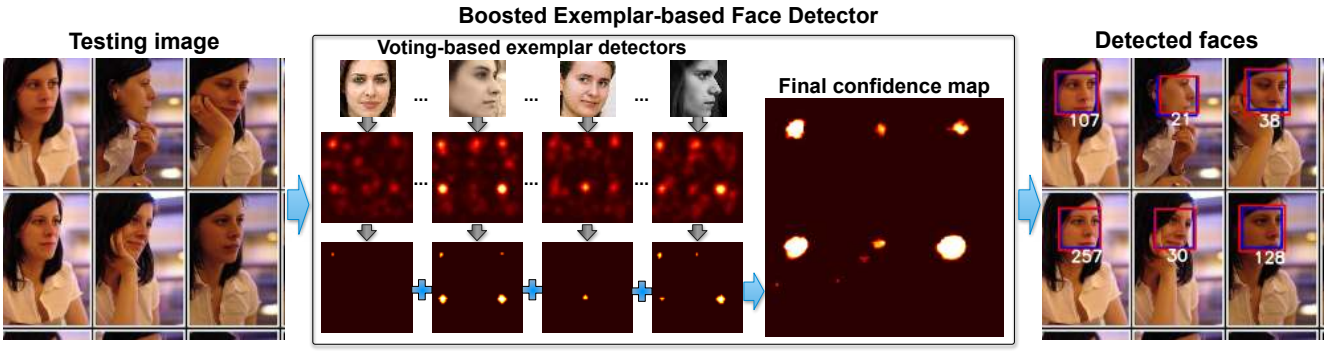


Figure 3. Detection work-flow of the boosted exemplar-based face detector: each exemplar processes the testing image into a voting map to obtain the confidence map; exemplar-level confidence maps are summed up to build the final confidence map; faces could then be located referring to the peaks on the final confidence map. Detection confidence scores are shown in integers below each detection bounding box. The blue and red squares are detection bounding boxes before and after calibration, respectively.

Figure 4. Top face exemplars selected by the RealAdaboost: different appearance variations are nicely covered, such as pose, expression and illumination.

$$\Omega^-_{i,m} = \{n | S(x_n, e_i) \in \mathcal{X}_{i,m} \wedge y_n = -1\}$$

$$W^+_{i,m} = \sum_{n \in \Omega^+_{i,m}} \omega^{(t)}_n, \quad W^-_{i,m} = \sum_{n \in \Omega^-_{i,m}} \omega^{(t)}_n, \qquad (5)$$

where $\bigcup_{m=1}^{M_i} \mathcal{X}_{i,m} = \mathcal{R}$, $\mathcal{X}_{i,m} \cap \mathcal{X}_{i,m'} = \emptyset, \forall m \neq m'$, $S(x_n, e_i)$ is the similarity score from the center of the voting map built by applying the exemplar e_i over the sample x_n as in Equation 1 (see Figure 2).

Let $i^* = \arg\min_i Z_i$, the most discriminative exemplar at the t-th iteration is then selected to be $\hat{e}_t = e_{i^*}$ associated with the domain-partitioned classifier $\hat{h}_t(x) = h_{i^*}(x)$ and the confidence ratings

$$c_{i^*,m} = \frac{1}{2} log(\frac{W^+_{i^*,m} + \epsilon}{W^-_{i^*,m} + \epsilon}), \qquad (6)$$

where ϵ is a small positive value for smoothness. The training sample weights are then updated to emphasis the hard training samples,

$$\omega^{(t+1)}_n \leftarrow (1 + ({\omega^{(t)}_n}^{-1} - 1) e^{y_n \hat{h}_t(x_n)})^{-1}. \qquad (7)$$

After T iterations, we have the T stages boosted exemplar-based face detector $H(x)$ as in Equation 3. In Figure 4, we show exemplars selected in the early stages in training the RealAdaboost framework.

In the testing stage, given a target image, each $\hat{h}_t(x)$ generates a confidence map for the image which is then aggregated to build the final confidence map to locate faces. Note

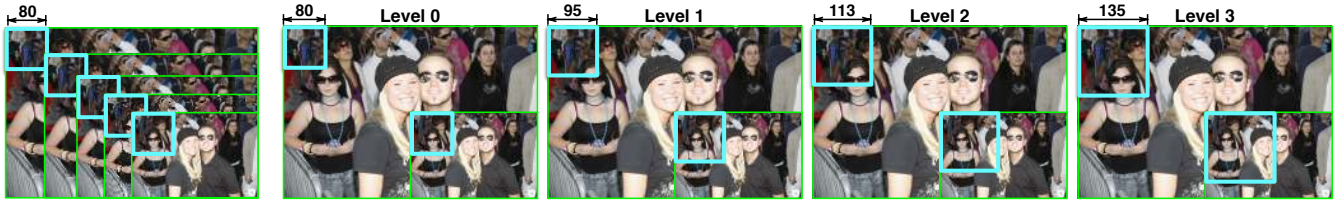
Figure 5. Some non-face negative exemplars selected by the RealAdaboost: this coincides with the intuition that negative exemplars should be discriminative to face and non-face classification, such as highly repeated patterns (scenes) and frequent false detections (round shape and human body).

that without the necessity of early stage rejection, the process of generating the T exemplar-level confidence maps could be parallelized freely, which could be another technical advantage of our algorithm.

The testing work-flow is shown in Figure 3. The rightmost image shows the detection result of the top 500 exemplars over the leftmost testing image.

We apply non-maximum suppression to locate faces from the final confidence map. This method produces decent results as the blue rectangles shown in Figure 3. However, due to the fact that the annotated face centers of face exemplars are not always aligned well with the center of the detected faces, there could be small differences in face center locations and face sizes between the detection bounding boxes and target faces. For example, a frontal face exemplar could vote to a shifted center for a profile testing face. The correct geometric normalization of face annotations is unknown as discussed in [20], we apply a calibration step to improve the quality of bounding boxes.

In the calibration, we make the image subregion within the detection bounding box ϕ work as an exemplar while the face exemplars in the database act as testing images. Therefore ϕ predicts a detection rectangle on each face exemplar. We calculate the offset and scale difference between the detection rectangle and the annotation for each exemplar, and average the offsets and scale differences over all face



(a) Standard Image Pyramid (b) Simplified Image Pyramid

Figure 6. Comparison between the standard image pyramid and the simplified image pyramid we adopted: across the levels of the simplified image pyramid, base face size varies but feature descriptors are shared to save computational resource.

exemplars as the final calibration offset and scale change which is applied to adjust the detection bounding box ϕ. Since the voting maps and visual words of ϕ are already obtained in the detection stage, the overhead of this step is very small.

This calibration work-flow is different from the validation step of [19]. While they recalculated the detection confidence score in their validation step, we only use the calibration step to improve the bounding box quality as an optional step. In the experiments, we observed the calibration step brings improvement (see Figure 11).

3.3. Negative Exemplars

In terms of the similarity scores, positive and negative samples are highly mixed in the low score range. As a result, among the trained exemplars, face exemplars will assign high positive confidence ratings to face regions and generally have nearly zero confidence ratings over non-face regions following Equation 6. Adding face exemplars could expand the appearance spanning of the face detector, but it also increases the potential of accumulating confidence over face-like non-face regions unexpectedly. Motivated by the fact that throughout the training stage, we are not relying on the assumption that the exemplars are face images, we add non-face images as candidate negative exemplars into the training exemplar pool for discriminative selection.

In this paper, the negative exemplar has a domain-partitioned weak classifier which assigns negative confidence to its highly similar image subregions. In the Figure 5, we show some selected negative exemplars. They could actively suppress the confidence ratings of potential false detections from cluttered backgrounds.

Randomly sampling from non-face images could be a straightforward way to collect large number of negative exemplar candidates. Empirically, hard negative samples can be potentially better choices based on the loose assumption that they are correlated to the appearance distribution of faces so they are correlated to each other. More implementation details are included in Section 4.1.1.

3.4. Efficiency Enhancements

In addition to the approaches discussed above, we introduce two techniques to improve the efficiency of the proposed approach: the tile-based detection which improves the memory efficiency and a simplified image pyramid which helps speed up the detection.

Figure 7. Large image is divided into overlapped tiles for memory constrained detection

3.4.1 Tile-based detection

Although the voting-based method could bypass the exhaustive sliding window process, it needs to maintain the voting maps and confidence maps in memory. This could be a defect for real-world application in detecting faces on a large image, since the required memory is proportional to the image size. To overcome this drawback, we propose to use a tile-based detection strategy. Specifically, we divide the image into tiles with overlapped regions and process each tile one by one independently. In this straightforward way, without degrading the performance, we confine the memory footprint to be a constant independent to the testing image size. As illustrated in Figure 7, the left image of size 1000×667 is divided into 4 tiles of sizes not larger than 640×640 which leads to a 40% smaller memory footprint.

3.4.2 Simplified Image Pyramid

In the standard face detection pipeline, the base size of face is fixed throughout the detecting procedure while an image pyramid is built from the testing image to detect faces at different scales. A defect of applying this approach in our work is that we need to extract features repeatedly over each level of the image pyramid. In fact, when the scaling factor of image pyramid is relatively small and the feature descriptor is scale invariant, the extracted descriptors could be reused without affecting the detection performance. Under this assumption, we set the scaling factor of image pyramid to be 2. Within two levels of the image pyramid, we keep using the same feature descriptors to detect face with 4 different base sizes with a scaling factor $\sqrt[4]{2}$. As illustrated in Figure 6, for a 2-level image pyramid we can reduce the

Figure 8. Annotation mismatch between our face exemplar and FDDB ground-truth: We are using square bounding boxes while FDDB uses ellipses.

number of extracted features by 55.47% which could make a significant difference in terms of detection speed.

4. Experiments

We verify the proposed work over two published and widely recognized datasets, Annotated Faces in-the-Wild (AFW) [25] and the Face Detection Data Set and Benchmark (FDDB) [7] and a personal photo album *G-Album* [5, 12]. On all the datasets, the proposed method utilizes 3,000 exemplars and significantly outperforms the state-of-the-art performances, especially on FDDB. The detection performance of a faster detector with 500 exemplars is reported as well to show the proposed approach could have decent performance with a more practical setting. In comparison, previous exemplar-based face detector [19] requires a database contains 18,486 exemplars in the detecting stage.

4.1. Setting

We extract dense SIFT features with 3 pixels spacing in both training and testing stages. Fast approximate k-means [14] is used for training the vocabulary of 100,000 visual words and FLANN [15] is used for quantizing the extracted feature descriptors. The base size of our face detector is 80×80.

In the training stage, we collect 15,832 face images as candidate face exemplars and 12,732 non-face images. Part of them are from the AFLW dataset [8], and no testing images are included in the exemplar pool. For each face image, we expand the bounding box of the annotated face by a factor of 2, crop out the face from the image and resize it to 160×160 to make it a face exemplar. Feature descriptors extracted from the whole 160×160 images are used for calculating the inverse document frequencies and term frequencies of the visual words while only the descriptors extracted from the inner 80×80 face regions are used as the exemplar, *i.e.* for calculating the voting maps.

In building the domain-partitioned classifiers associated with the exemplars, we manually set the number of partitions to be 3 for face exemplars and 2 for non-face exemplars. For training efficiency, we make the domain-partition setting to be the same for face and non-face exemplars, respectively. We manually choose the thresholds for the partitions by observing the distribution of similarity scores of the training samples, and checking the detection performance with the help of a validation dataset.

Figure 9. Qualitative results of our detector over AFW (left), FDDB (middle) and *G-Album* (right).

In the testing stage, images smaller than 1480×1480 will be upscaled to have a maximum dimension to be 1480. In the tile-based detection (presented in Section 3.4.1), the tile size is set to be 640×640 with 140 pixels overlapping stride. The simplified image pyramid is built to be image pyramid of scaling factor 2 with 4 different base face sizes with scaling factor $\sqrt[4]{2}$.

4.1.1 Training

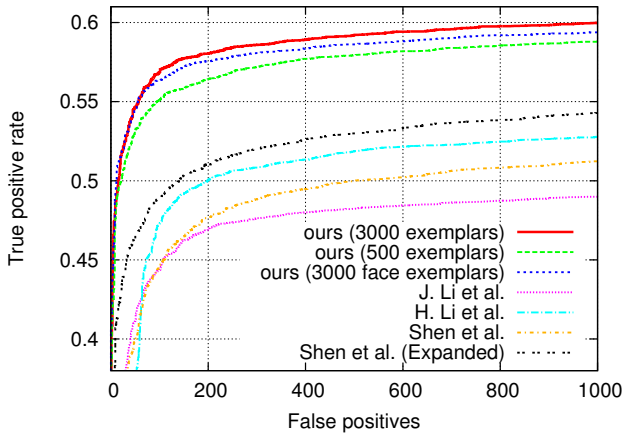
The positive training data are 80×80 face subregions and their horizontally flipped versions. To collect negative training data, we extract 80×80 negative samples from a set of non-face images uniformly. We collect 31,664 positive samples and 468,666 negative samples to train the first boosted exemplar-based face detector. Then we perform bootstrapping [3] over a larger set of negative images to detect hard negative samples. The hard negative samples are then added into both the training corpus as negative training data and the exemplar pool as potential negative exemplars to train the final stage face detector.

4.2. Face Detection Data Set and Benchmark

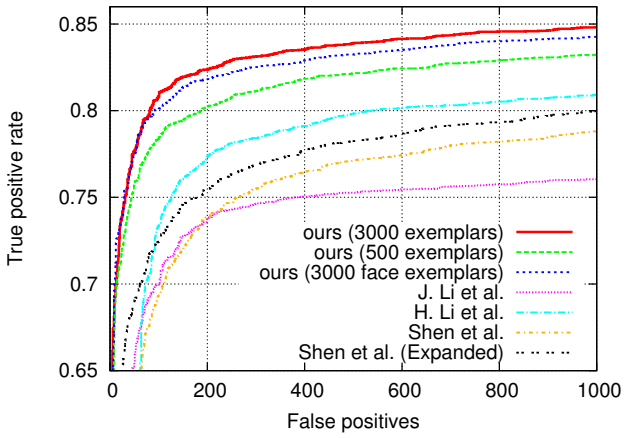
FDDB contains 5,171 annotated faces in 2,845 images. Images in this dataset are extracted from news articles which present large face appearance variations. As far as we know, [11, 10, 19] showed the state-of-the-art performances on FDDB. The evaluation procedure is standardized and researchers are expected to use the same evaluation program to report the results.

Due to the annotation mismatch (see Figure 8), quite a few true positive detections are evaluated as false alarms because their overlapping ratio with the ground truth are just right below the threshold. To resolve this issue, we expand our square bounding boxes vertically by 20% to approximate ellipses with rectangles. Since [19] reported similar annotation mismatch issue, to make the comparison fair, we apply the same adjustment for their detection results [19] and confirmed their observation.

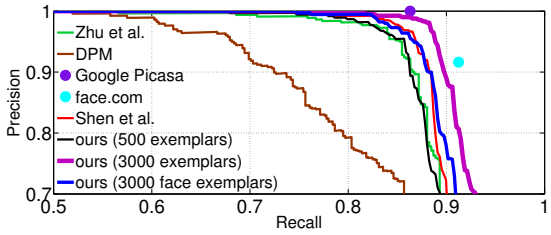
As shown in Figure 10, using 500 exemplars our method outperformed the state-of-the-art by a significant margin. With more exemplars, we achieve further improvement. We also reported the result without vertically expanding the detection bounding boxes in Figure 11.



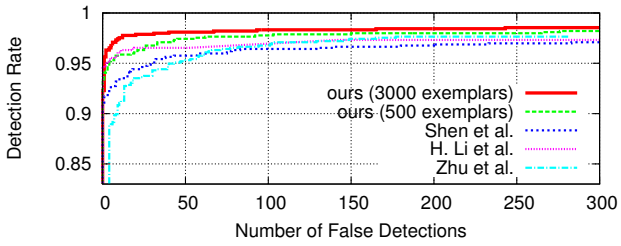
(a) FDDB - Continuous score



(b) FDDB - Discontinuous score



(c) AFW



(d) *G-Album*

Figure 10. Performance evaluation: We reported the results of our detectors with 500 and 3,000 exemplars including negative exemplars and a 3,000 exemplars detector with only face exemplars. On FDDB we compared with J. Li et al. [11], H. Li et al. [10] and Shen et al. [19]; on AFW, we compared with Zhu et al. [25], DPM [4], Face.com [1] and Google Picasa [2]; on *G-Album* we compared with Shen et al. [19], Zhu et al. [25] and Li et al. [10].

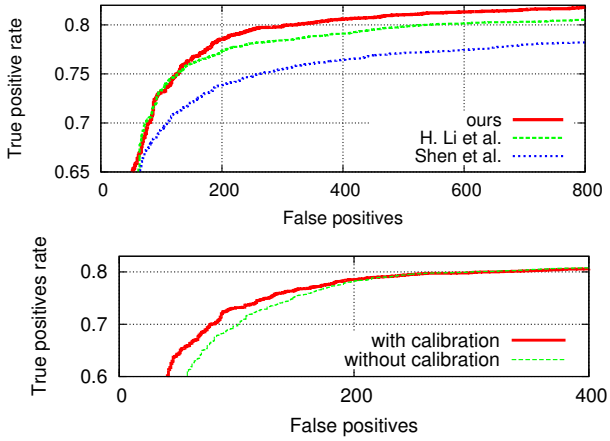


Figure 11. Evaluation on FDDB (Discontinuous score): the above one shows without vertically expanding the bounding boxes, our method still outperformed the state-of-the-art; the below one shows the calibration step helps in the low false positives part.

4.3. Annotated Faces in the Wild

Annotated Faces in the Wild (AFW) is a 205 images dataset proposed by [25]. This dataset contains cluttered backgrounds with large appearance variations intentionally. As shown in Figure 10, our detector further outperformed the state-of-the-art methods and is closing the gap to the commercial detectors. The improvement of adding negative exemplars are more obvious in AFW since the testing images have more cluttered backgrounds.

4.4. G-Album

We use 512 photos with 895 labeled faces from *G-Album* [12, 5]. Different from the previous two benchmarks, this family photo album presents a more realistic application scenario. It contains a lot of children and baby faces presenting very rich variations in terms of pose and expression. As shown in Figure 10, our detector achieved the best performance on this dataset.

4.5. Qualitative Results

In Figure 9, we present several detection results over the three datasets with a fixed detection threshold 30. With this threshold, on AFW we can achieve 99.0% precision and 87.5% recall rate; on FDDB we can achieve 82.9% recall rate with totally 254 false detections out of 2,845 images; on the *G-Album* we achieve 97.6% recall rate with 22 false detections out of 512 photos.

4.6. Efficiency

In terms of running time and memory usage, our boosted exemplar-based face detector is more efficient. With a detector of 500 exemplars with minimal face size of 80×80, the detection time of our method is 900 ms for an image of size 1480×986. This performance is nearly 33 times faster than the Shen et al. [19] detector which takes 33 seconds. Zhu et al. [25] detector takes 231 seconds for the same image.

Given the same testing image, for the memory footprint size, our detector requires around 150MB memory, the Shen et al. detector acquires 866MB memory while Zhu et al. detector could allocate up to 2GB memory. Obviously, our detector is more practical for real-world applications.

5. Conclusion

In this paper, we present an efficient boosted exemplar-based face detector utilizing exemplar-based weak detector and the RealAdaboost algorithm to approach the unconstrained face detection problem. This approach makes the exemplar-based face detection practical by largely reducing the number of required exemplars and training discriminative exemplar detectors. Furthermore by making the definition of exemplar more general to incorporate both face and non-face images, the efforts of collecting exemplars are relieved and negative exemplars can be built purposely to suppress false alarms. By enhancing the idea with a tile-based detection paradigm and simplified image pyramid, the speed and memory efficiency is further improved. The significant improvement of face detection accuracy comparing with the state-of-the-art methods over two public benchmarks and one personal photo album demonstrated the effectiveness of our proposed approach.

Acknowledgement

This work is partly supported by US National Science Foundation Grant IIS 1350763, China National Natural Science Foundation Grant 61228303, a collaborative gift grant from Adobe, GH's start-up funds from Stevens Institute of Technology, a Google Research Faculty Award, a gift grant from Microsoft Research, and a gift grant from NEC Labs America.

References

[1] http://face.com/. 7

[2] http://picasa.google.com/. 7

[3] B. Efron and R. J. Tibshirani. *An Introduction to the Bootstrap*. 6

[4] P. F. Felzenszwalb, R. B. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part based models. In *PAMI*, 2010. 1, 7

[5] A. C. Gallagher and T. Chen. Clothing cosegmentation for recognizing people. In *CVPR*, 2008. 6, 7

[6] C. Huang, H. Ai, Y. Li, and S. Lao. Learning sparse features in granular space for multi-view face detection. In *FG*, 2006. 2

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

[8] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *ICCVW*, 2011. 6

[9] B. Leibe, A. Leonardis, and B. Schiele. Combined object categorization and segmentation with an implicit shape model. In *ECCVW*, 2004. 1, 3

[10] H. Li, G. Hua, Z. Lin, J. Brandt, and J. Yang. Probabilistic elastic part model for unsupervised face detector adaptation. In *ICCV*, 2013. 6, 7

[11] J. Li, T. Wang, and Y. Zhang. Face detection using surf cascade. In *ICCVW*, 2011. 6, 7

[12] D. Lin, A. Kapoor, G. Hua, and S. Baker. Joint people, event, and location recognition in personal photo collections using cross-domain context. In *ECCV*, 2010. 6, 7

[13] T. Malisiewicz, A. Gupta, and A. A. Efros. Ensemble of exemplar-svms for object detection and beyond. In *ICCV*, 2011. 2

[14] M. Muja and D. G. Lowe. Fast approximate nearest neighbors with automatic algorithm configuration. In *In VISAPP International Conference on Computer Vision Theory and Applications*, 2009. 6

[15] J. Philbin, O. Chum, M. Isard, J. Sivic, and A. Zisserman. Object retrieval with large vocabularies and fast spatial matching. In *CVPR*, 2007. 6

[16] R. E. Schapire and Y. Singer. Improved boosting algorithms using confidence-rated predictions. In *Machine Learning*, 1999. 2, 3

[17] Y. Shan, F. Han, H. S. Sawhney, and R. Kumar. Learning exemplar-based categorization for the detection of multi-view multi-pose objects. In *CVPR*, 2006. 2

[18] X. Shen, Z. Lin, J. Brandt, S. Avidan, and Y. Wu. Object retrieval and localization with spatially-constrained similarity measure and k-nn re-ranking. In *CVPR*, 2012. 1, 3

[19] X. Shen, Z. Lin, J. Brandt, and Y. Wu. Detecting and aligning faces by image retrieval. In *CVPR*, 2013. 1, 2, 3, 5, 6, 7, 8

[20] P. Viola, J. C. Platt, and C. Zhang. Multiple instance boosting for object detection. In *NIPS*, 2006. 4

[21] P. A. Viola and M. J. Jones. Rapid object detection using a boosted cascade of simple features. In *CVPR*, 2001. 1, 2

[22] P. Wang and Q. Ji. Learning discriminant features for multi-view face and eye detection. In *CVPR*, 2005. 2

[23] J. Xu, D. Vazquez, S. Ramos, A. Lopez, and D. Ponsa. Adapting a pedestrian detector by boosting lda exemplar classifiers. In *CVPRW*, 2013. 2

[24] C. Zhang and Z. Zhang. A survey of recent advances in face detection. Technical report, 2010. 1, 2

[25] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*, 2012. 1, 6, 7, 8