Face Detection by Aggregating Visible Components

Jiali Duan[1(✉)], Shengcai Liao[2], Xiaoyuan Guo[3], and Stan Z. Li[2]

[1] School of Electronic, Electrical and Communication Engineering,
University of Chinese Academy of Sciences, Beijing, China
jli.duan@gmail.com

[2] Center for Biometrics and Security Research
and National Laboratory of Pattern Recognition,
Institute of Automation, Chinese Academy of Sciences, Beijing, China
{scliao,szli}@nlpr.ia.ac.cn

[3] School of Engineering Science, University of Chinese Academy of Sciences,
Beijing, China
xiaoyuanguo.ucas@gmail.com

Abstract. Pose variations and occlusions are two major challenges for unconstrained face detection. Many approaches have been proposed to handle pose variations and occlusions in face detection, however, few of them addresses the two challenges in a model explicitly and simultaneously. In this paper, we propose a novel face detection method called Aggregating Visible Components (AVC), which addresses pose variations and occlusions simultaneously in a single framework with low complexity. The main contributions of this paper are: (1) By aggregating visible components which have inherent advantages in occasions of occlusions, the proposed method achieves state-of-the-art performance using only hand-crafted feature; (2) Mapped from meanshape through component-invariant mapping, the proposed component detector is more robust to pose-variations (3) A local to global aggregation strategy that involves region competition helps alleviate false alarms while enhancing localization accuracy.

1 Introduction

Unconstrained face detection is challenging due to pose and illumination variations, occlusions, blur, etc. While illumination variations are handled relatively better due to many physical models, pose variations and occlusions are the most commonly encountered problems in practice[1]. Many approaches have been specifically proposed to solve pose variations [2–4] and occlusions [5–9], however, few of them addresses pose variations and occlusions in a model explicitly and simultaneously.

[1] Blur or low resolution is a challenging problem mainly in surveillance. Though many blur face images exist in current benchmark databases (e.g. FDDB [1]), they are intentionally made out of focus in background while the main focus is the center figures in news photography.

C.-S. Chen et al. (Eds.): ACCV 2016 Workshops, Part II, LNCS 10117, pp. 319–333, 2017.
DOI: 10.1007/978-3-319-54427-4_24

Recently, a number of Convolutional Neutral Network (CNN) [10] based face detection methods [11–15] have been proposed due to the power of CNN in dealing with computer vision problems. However, CNN models generally deal with problems in face detection by learning from a large number of diverse training samples. Such data driven solutions may be good in dealing with various face variations, however, they usually result in very complex models that run slowly, which limits their application in practice, especially in embedding devices. On the other hand, Yang et al. [13] proposed a specific architecture called Faceness-Net, which considers facial component based scoring and their spatial configuration to explicitly deal with occluded face detection. This work inspires that explicit modeling of challenges in face detection is still required and more effective than pure data driven, though the fixed spatial configuration in Faceness-Net is still an issue, and the model is still expensive to apply.

Putting occlusions and large pose variations together, a common issue is that some facial components are invisible under either condition. This motivates us to only detect visible components that share some pose invariance property, and adaptively aggregate them together to form the whole face detection. Therefore, in this paper we propose a novel face detection method called Aggregating Visible Components (AVC), which addresses pose variations and occlusions simultaneously in a single framework.

Specifically, to handle pose variations, we define two pose-invariant (or pose-robust) components by considering half facial view, and a regression based local landmark alignment. Such a consistent component definition helps to reduce the model complexity. Accordingly, we train two component detectors, mirror them to detect the other half view, and introduce a local region competition strategy to alleviate false detections. To handle facial occlusions, we only detect visible facial components, and build a local to global aggregation strategy to detect the whole face adaptively. Experiments on the FDDB and AFW databases show that the proposed method is robust in handling pose variations and occlusions, achieving much better performance but lower model complexity compared to the corresponding holistic face detector.

The remaining parts of this paper are organized as follows. Section 2 gives a concise review of related works. Section 3 gives an overview of the proposed AVC detector. Section 4 introduces the pose-invariant component definition and the detector training. In Sect. 5, we present the local region competition strategy and the adaptive local to global aggregation strategy. Experimental results on AFW and FDDB are shown and discussed in Sect. 6 and we conclude the paper in Sect. 7.

2 Related Works

Given that the original Viola-Jones face detector [16] is limited to multi-view face detection, various cascade structures have been proposed to handle pose variations [2–4]. Today multi-view face detection by partitioning poses into discrete ranges and training independently is still a popular way to handle pose variations, for example, in recent works [12,17]. Zhu and Ramanan [18] proposed to jointly detect a face, estimate its pose, and localize face landmarks in the wild by a Deformable Parts-based Model (DPM), which was further improved in [19,20]. Ranjian et al. [21] proposed to combine deep pyramid features and DPM to handle faces with various sizes and poses in unconstrained settings. Chen et al. [22] proposed to combine the face detection and landmark estimation tasks in a joint cascade framework to refine face detection by precise landmark detections. Liao et al. [23] proposed to learn features in deep quadratic trees, where different views could be automatically partitioned. These methods are effective in dealing with pose variations, however, not occlusions simultaneously.

Face detection under occlusions is also an important issue but has received less attention compared to multi-view face detection, partly due to the difficulty of classifying arbitrary occlusions into predefined categories. Component-based face detector is a promising way in handling occlusions. For example, Chen et al. [8] proposed a modified Viola-Jones face detector, where the trained detector was divided into sub-classifiers related to several predefined local patches, and the outputs of sub-classifiers were re-weighted. Goldmann et al. [24] proposed to connect facial parts using topology graph. Recently, Yang et al. [13] proposed a specific architecture called Faceness-Net, which considers faceness scoring in generic object proposal windows based on facial component responses and their spatial configuration, so that face detection with occlusions can be explicitly handled. However, none of the above methods considered face detection with both occlusions and pose variations simultaneously in unconstrained scenarios.

Our work is also different from other part-based methods like [25–29] in that [25] describes an object by a non-rigid constellation of parts and jointly optimize parameters whereas we learn component detectors independently and apply an aggregation strategy to constitute a global representation. On the other hand, AVC define parts via component-invariant mapping, in contrast to [26] which defines parts by a search procedure while [27–29] deploy CNN structures.

Recently, the Convolutional Neutral Network (CNN) [10] based methods [11–15] have been proposed for face detection due to the power of CNN in dealing with computer vision problems. For example, Li et al. [11] proposed a cascade architecture based on CNN and the performance was improved by alternating between the detection net and calibration net. Most recently Zhang et al. [14] and Ranjan et al. [15] combined face detection with other vision tasks such as face alignment and involved multi-task loss into CNN cascade.

3 Overview of the Proposed Method

Figure 1 is an overview of the proposed AVC face detection method. It includes three main steps in the detection phase: visible component detection step, local region competition step, and the local to global aggregation step. AVC works by detecting only the visible components which would be later aggregated to represent the whole face. Two half-view facial component detectors are trained, and for this we introduce a pose-invariant component definition via a regression based local landmark alignment, which is crucial for training sample cropping and pose-invariant component detection. Then the two learned detectors are mirrored to detect the other half view of the facial components. Next, the detected visible facial components go through a local region competition module to alleviate false detections, and finally a local to global aggregation strategy is applied to detect the whole face adaptively.

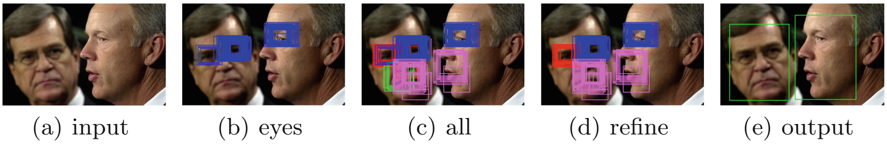

(a) input (b) eyes (c) all (d) refine (e) output

Fig. 1. The processing steps of the proposed AVC face detection method. (a) Input image. (b) Visible eye detection. (c) Detection of all visible components (Red: left eye; Blue: right eye; Green: left mouth; Pink: right mouth). (d) Refinement after local region competition. (e) Aggregated whole face detection. (Color figure online)

The intuition behind our component-based design is the fact that face images in real-world applications are often with large pose variations and occlusions. Consider for example, a face turning left over 60 degrees (see Fig. 2(a)), where the holistic face detector unavoidably includes unwanted backgrounds (see Fig. 2(b)).

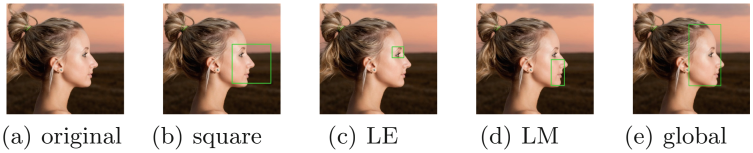

(a) original (b) square (c) LE (d) LM (e) global

Fig. 2. Illustration of holistic face detection and component-based face detection. (a) Input image. (b) Typical holistic face detection. (c) Left eye (LE) detection. (d) Left mouth (LM) detection. (e) Aggregating LE and LM to get a global detection.

However, a robust face detector should not only predict the number of faces but also give bounding boxes as tight as possible. The criteria on this performance was first introduced by FDDB [1], a face benchmark that employs both discrete metric and continuous metric for evaluation. While a typical face detector may fail to bound a profile face tightly and miss faces under occlusions, we discover however, that pose variations and occlusions can be jointly solved by locating and aggregating facial components. We trained two facial component detectors respectively for the detection of left eyebrow + left eye (denoted as LE Fig. 2(c)) and left nose + left mouth (denoted as LM Fig. 2(d)).

It's observed that although a face with large rotation towards left may lead to left eye invisible, we can still, under this circumstance, locate the right eye or mouth and nose etc. It also applies to occlusions where for example, the left half face is occluded by another person's shoulder, we can still locate the whole face by the detection of right eye. Furthermore, we only consider training two half facial view components, and mirror them to detect the other half view. This strategy not only reduces the training effort, but also enables us to deal with larger pose variations because for example, the left eye component appears to be invariant under 0–60° pose changes, and beyond this range the right eye or other component is usually detectable.

4 Pose-Invariant Component Detection

4.1 Pose-Invariant Component Mapping

As was indicated in AFLW [30], although there is largely an agreement on how to define anchor points and extents of rectangle for frontal faces, it's not so obvious for profile and semi-profile views, which makes it harder to get consistently annotated samples for training. Unlike the training input of a holistic face detector, facial part detector requires uniform eye patches and mouth patches as training set. This would not be made possible without pose-invariant component mapping.

Samples in AFLW consist of 21 landmarks. We first calculate the mean shape of the whole database with samples normalized and missing coordinates excluded. Region in the mean shape which we want to map i.e. left eyebrow and left eye for LE component is mapped directly to a new input sample by applying the transformation

$$a\bar{\mathbf{x}} + x_0 = \mathbf{x} \tag{1}$$

$$a\bar{\mathbf{y}} + y_0 = \mathbf{y} \tag{2}$$

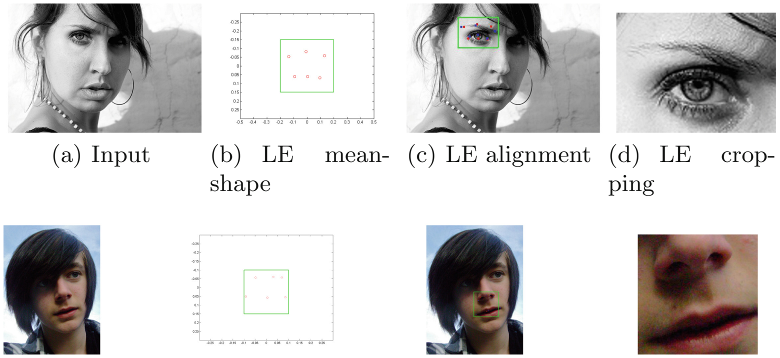
Note that in (1) and (2) $\bar{\mathbf{x}}$ and $\bar{\mathbf{y}}$ are vectors representing x coordinates and y coordinates of mean shape while $\mathbf{x}$ and $\mathbf{y}$ representing those of a new sample. $\mathbf{E}$ is a nx1 vector with all elements being 1, x_0, y_0 are scalars that denote offsets and n is the number of landmarks used for regression. Closed form solution can be derived as the following

$$a = \frac{\bar{\mathbf{x}}^T \cdot \mathbf{x} + \bar{\mathbf{y}}^T \cdot \mathbf{y} - \frac{1}{n} \cdot (\bar{\mathbf{x}}^T \cdot \mathbf{E})(\mathbf{x}^T \cdot \mathbf{E}) - \frac{1}{n} \cdot (\bar{\mathbf{y}}^T \cdot \mathbf{E})(\mathbf{y}^T \cdot \mathbf{E})}{\bar{\mathbf{x}}^T \cdot \mathbf{x} + \bar{\mathbf{y}}^T \cdot \mathbf{y} - \frac{1}{n} \cdot (\bar{\mathbf{x}}^T \cdot \mathbf{E})^2 - \frac{1}{n} \cdot (\bar{\mathbf{y}}^T \cdot \mathbf{E})^2} \tag{3}$$

$$x_0 = \frac{1}{n} \cdot \mathbf{x}^T \cdot \mathbf{E} - a\frac{1}{n} \cdot \bar{\mathbf{x}}^T \cdot \mathbf{E} \tag{4}$$

$$y_0 = \frac{1}{n} \cdot \mathbf{y}^T \cdot \mathbf{E} - a\frac{1}{n} \cdot \bar{\mathbf{y}}^T \cdot \mathbf{E} \tag{5}$$

An intuitive visual interpretation is shown in Fig. 3. In Fig. 3(c), blue points are annotated landmarks while red points are mapped from meanshape. Positive samples extracted in this way retain excellent uniformity, which would be used for training LE and LM component detector. The pose-invariant component mapping method is also used for preparing negative samples for bootstrapping (see Fig. 4).



(a) Input (b) LE meanshape (c) LE alignment (d) LE cropping

(e) Input (f) LM meanshape (g) LM alignment (h) LM cropping

Fig. 3. Pose-invariant component mapping and cropping. (a) Input. (b) Meanshape of the LE component. (c) Regression based local landmark alignment of LE component. (d) Cropping of the LE component. (e) Input. (f) Meanshape of the LM component. (g) Regression based local landmark alignment of LM component. (d) Cropping of the LM component. (Color figure online)

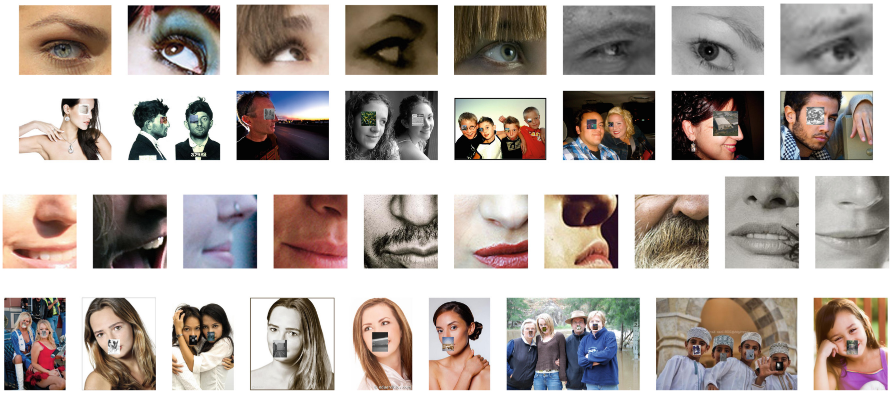


Fig. 4. Positive and negative examples for components. The first and third rows show positive training samples of the LE and LM components respectively, while the second and forth rows show images for bootstrapping negative LE and LM samples respectively.

4.2 Why LE and LM?

In our paper, we trained two part-based detectors, namely LE (left eyebrow and left eye) and LM (left nose and left mouth) and Fig. 4 displays some positive and hard-negative training samples obtained using method of the last subsection. But why not eyes, noses or other patches? Our motivations are: (1) These patches are not defined arbitrarily or conceptually but based on the regression of local landmarks. As in Fig. 3, these landmarks are derived by LE/LM meanshape of AFLW to ensure that they retain invariance throughout the database

(2) Why 6 landmarks instead of 3 or 9? According to AFLW, a nose is defined by 3 landmarks, the width/height of these patches would then be too small for training and testing. While 9 landmarks would result with a facial area too broad thus vulnerable for occlusions.

4.3 Training Procedure

In this subsection, we give a brief introduction about the feature employed for facial representation as well as the work flow of the training algorithm.

Feature: We choose NPD [23] as our feature mainly for its two properties: illumination invariant and fast in speed because each computation involves only two pixels. For an image with size $p = w \times h$, the number of features computed is C_p^2 which can be computed beforehand, leading to superiority in speed for real world applications. With the scale-invariance property of NPD, the facial component detector is expected to be robust against illumination changes which is important in practice.

Training Framework: The Deep Quadratic Tree (DQT) [23] is used as weak classifier which learns two thresholds and is deeper compared to typical tree classifiers. Soft-Cascade [31] as well as hard-negative mining are applied for cascade training. While individual NPD [32] features may be "weak", the Gentle AdaBoost algorithm is utilized to learn a subset of NPD features organized in DQT for stronger discriminative ability.

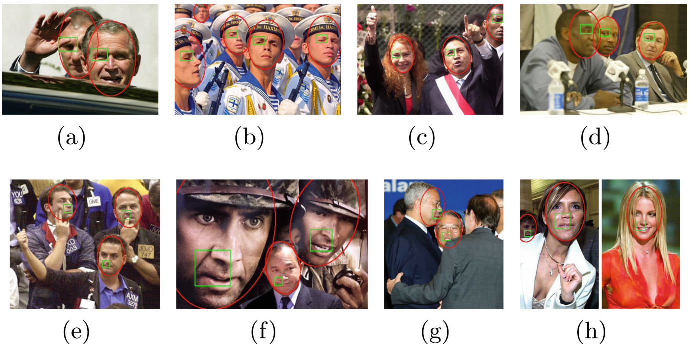
5 Local to Global Aggregation

5.1 Symmetric Component Detection

Figure 5 shows some example outputs by LE and LM detector respectively. As can be seen, our component-based detector has the inherent advantages under occasions of occlusions (Fig. 5(a,h)) and pose-variations (Fig. 5(c,g)), where a holistic detector would normally fail. The detection of right eyebrow + right eye (RE) and right mouth + right nose (RM) can be achieved by deploying the detector of their left counterpart. Figure 6(a) to (d) illustrates how we locate RM and RE using the same detectors as LM and LE.

5.2 Local Region Competition

Adopting facial part detection also brings about many troublesome issues. If handled improperly, the performance will vary greatly. First, LE, LM, RE, RM detector for different facial parts will each produce a set of candidate positive windows with a set of confidence scores. But the goal for face detection is to locate faces each with a bounding box as tight as possible, so we need to merge these detections from different facial part detectors and remove duplicated windows. A common solution is Non-Maximum Suppression (NMS) [33] but issue


(a) (b) (c) (d)

(e) (f) (g) (h)

Fig. 5. Some example component detections by the proposed LE (upper row) and LM facial component detector.

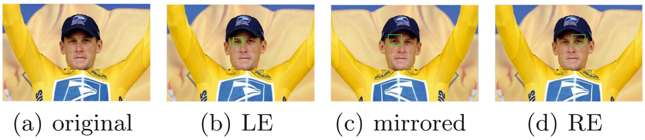

(a) original (b) LE (c) mirrored (d) RE

Fig. 6. (a): Input image; (b): Left eye detection; (c): Left eye detection in mirrored image; (d): Right eye detection mapped back to the original image.

arises on how to do window merging with a good trade-off between high precision rate and high detection rate. Second, different benchmarks with different annotation styles could lead to biased evaluation. Noted in [20], this diversity becomes more prominent for profile faces. In this section, we address the above issues by exploiting the advantage of a competitive strategy.

Figure 1 illustrates the idea of the proposed local region competition. The core idea is to reject false alarms during merging (compete) while improving localization accuracy during aggregation (collaborate). In Algorithm 1 line 6 to line 11 first obtains candidate outputs of a specific facial region by LE, RE, LM, RM facial part detectors denoted as region_rects, see Fig. 1(c) which shows detection results of all components and Fig. 1(d) after competition as an example. In this example, left eye region may well contain the outputs of other facial part detectors such as RE (false alarms) other than LE and vice versa. It is through this competitive strategy that we ensure candidate windows of only one facial part detector are reserved for each region, rooting out the possibility of using false alarms for aggregation.

5.3 Aggregation Strategy

After deploying competitive strategy to exclude possible false positives, the task now is to ensure accurate localization of detection outputs. This is achieved by taking the full use of information from rects of different regions. We use rectangle as facial representation. Note that our proposed pipeline also applies to elliptical representation as the aforementioned workflow remains unchanged.

Algorithm 1. Detection framework with local region competition strategy

```
Require:
    Input and Model: LE and LM model; RGB or gray image I
    Options: eyeHeight, eyeWidth, mouthHeight, mouthWidth for scanning windows;
    overlap for IOU; minEyeSize; minMouthSize; numThreads for parallel computing
Ensure: outRect
 1: [LE,LM]=Scan(eyeModel,mouthModel,I,eyeHeight,eyeWidth,mouthHeight,
    mouthWidth,minEyeSize,minMouthSize,numThreads)
 2: Symmetrically detect RE and RM
 3: $LE \cup LM \cup RE \cup RM \subseteq \mathrm{R}$
 4: predicate(i,j)=1 if IOU between R pair (i,j) > overlap
 5: [label,numCandi]=Partition(predicate)
 6: for i=1:numCandi do
 7:    Get region_rects{i} with rects labelled i
 8:    for r in region_rects{i} do
 9:       categorize r to LE, LM, RE, RM detector
10:    end for
11:    Reserve rects from the detector with the highest score
12:    Fitting rects to bounding boxes for the whole face
13: end for
14: predicate(i,j)=1 if IOU between rectangle pair (i,j) > overlap
15: [label,numCandi]=Partition(predicate)
16: for i=1:numCandi do
17:    Weight adjustment
18: end for
19: Elimination
20: return outRect
```

In Algorithm 1 line 12, winning rectangles from each region as illustrated in Fig. 5 are regressed directly to bounding boxes. Note that we only learn two sets of regression parameters (linear regression), because during inference the coordinates of RE/RM component are first mirrored, regressed and then mirrored back using the same parameters of their left counterparts. This is a local to global bottom up strategy because rects of different facial regions are mapped to global facial representations. In Algorithm 1 Line 15 to Line 18, these rects are then concatenated for partitioning using disjoint-set algorithm. Then the locations of partitioned rects are translated and adjusted by tuning their widths and heights according to their confidence scores (weights). Through this process, information of different regions are collaborated to get a more accurate localization of the whole face. Finally, NMS [33] is deployed to eliminate interior rects.

6 Experiments

6.1 Training Parameters:

Annotated Facial Landmarks in the Wild (AFLW) [1] is an unconstrained face benchmark that contains 25993 face annotations in 21997 real world images with large pose variations, occlusions, illumination changes as well as a diversity of ages, genders, and ethnicity. In total, we use 43994 images from AFLW together with its flipped counterpart as positive samples and 300000 background images for training. And an additional 12300 images of natural scenes are scraped from the Internet to mask face components for hard-negative mining. In training AVC, images of 15x20 pixels are assigned to LE component while images of 20x20 pixels are used for LM. Pose-invariant component mapping is deployed to crop positive training patches and prepare bootstrapping samples.

6.2 AFW Results:

Annotated Faces in the Wild (AFW) [18] contains 205 images collected from Flickr that contain images of cluttered scenes and different viewpoints.

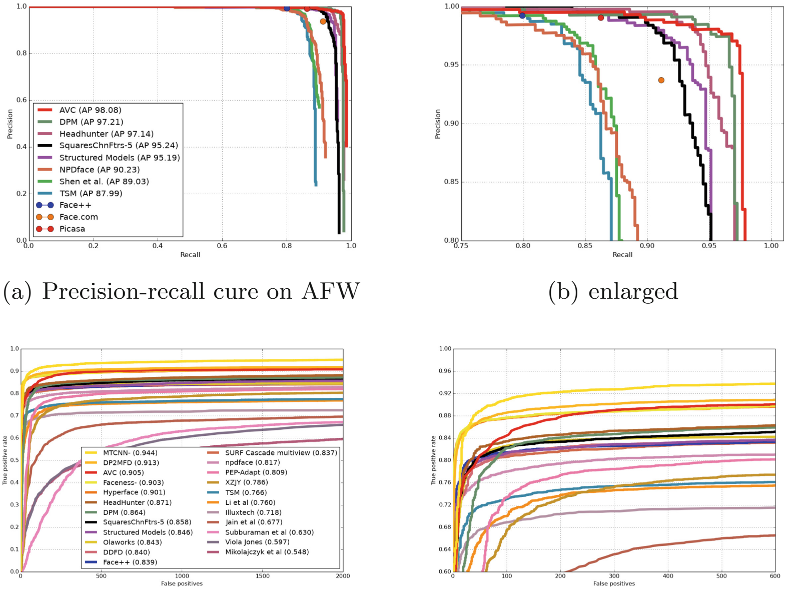
To evaluate on AFW, we fit winning rects from local component detectors to rectangle representations of the whole face, which would be used for further aggregation. The fitting parameters are learned on AFLW using 10-cross validation and this also applies to the learning of elliptical fitting parameters for testing on FDDB.

We use the evaluation toolbox provided by [20]. The comparison of Precision-Recall curves generated by different methods is shown in Fig. 7(a). We compare AVC with both academic methods like DPM, HeadHunter, Structured Models and commercial systems like Face++ and Picasa. As can be seen from the figure, AVC outperforms DPM and is superior or equal to Face++ and Google Picasa. The precision of AVC is 98.68% with a recall of 97.13%, and the AP of AVC is 98.08%, which is comparable with the state-of-the-art methods. Example detection results are shown in the first row of Fig. 8, note that we output rectangle for evaluation on AFW.

6.3 FDDB Results:

Face Detection Data Set and Benchmark (FDDB) [1] contains 2845 images with 5171 faces, with a wide range of arbitrary poses, occlusions, illumination changes and resolutions. FDDB uses elliptical annotations and two types of evaluation metrics are applied. One is the discrete score metric which counts the number of detected faces versus the number of false alarms. A detected bounding box is considered true positive if it has an IoU of over 0.5 with ground truth. The other is the continuous score metric that measures the IoU ratio as the indicator for performance.

As FDDB uses ellipse for annotations, we fit the output rectangles to elliptical representations of the whole face. We use the evaluation code provided by Jain and Learned-Miller [1] and the results using discrete score metric are shown in Fig. 7. We compare our results with the latest published methods on FDDB including MTCNN, DP2MFD, Faceness-Net and Hyperface. Ours performs worse than MTCNN and DP2MFD which resort to powerful yet complex CNN features but is better than Faceness-Net, which is also component-based but with the help of CNN structure. AVC gets 84.4% detection rate at $\text{FP} = 100$,



(a) Precision-recall cure on AFW

(b) enlarged

(c) Curve of discrete score on FDDB

(d) enlarged

Fig. 7. Experimental results on AFW and FDDB database. Best viewed in color.

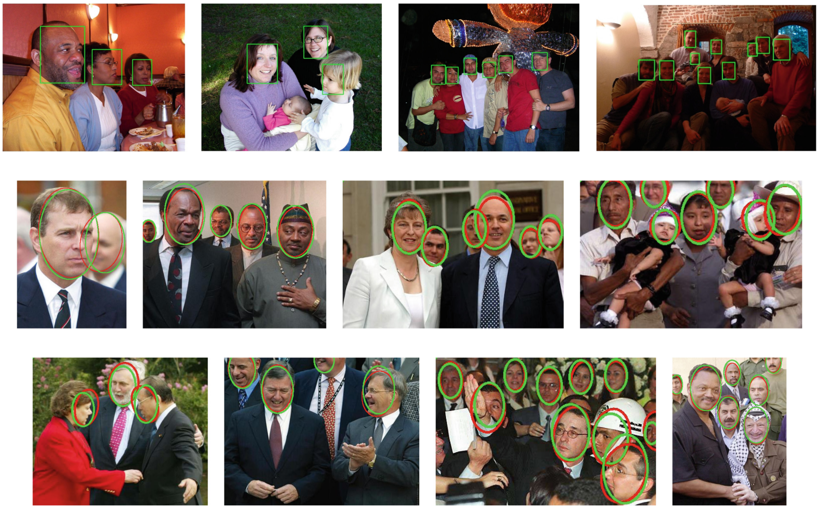


Fig. 8. Qualitive results of AVC on AFW (first row using rectangle representations) and FDDB (second and third row using elliptical representations).

and a detection rate of 89.0% at FP = 300. Example detection results are shown in the second and third row of Fig. 8, where faces under poses changes and occlusions have been successfully located.

6.4 Does Component-Invariant Mapping Help?

We have tried two other methods when preparing facial-component patches for training component detectors. One is to define anchor points and extents of rectangle, the other is to project 3D landmarks back to 2D plane. However, unlike training holistic face detector that gets by with ordinary methods, the uniformity of component training-set under profile or semi-profile views deteriorates notably compared to those under frontal views. The resulting detectors that we have trained achieve at best 81% AP on FDDB. To the best of our knowledge, it remains a tricky issue on how to achieve consistency under profile views [30]. This motivates us to make new attempts and explore component-invariant mapping, whose performance is further boosted with the help of symmetric component detection because, when a face only exposes RE/RM component, LE/LM component detector would fail. Second, its likely that symmetric component detection presents a symmetric but unblocked or simpler view for detector. Third, symmetric detection obviates the need to train another two more detectors and regression parameters. Experiment shows that trained part-detectors using conventional cropped patches will decrease AP by about 8.2% on FDDB.

6.5 Model Complexity

As is shown in Table 1, different tree levels for training have been evaluated, leading to different training stages and number of weak classifiers. Training FAR indicates to what extent AVC has converged, but it can not reflect the performance of the model on test set. The complexity of the model is measured by aveEval, which means the average number of NPD features evaluated per detection window. The lower the value of aveEval, the faster the detector. For the sake of speed, this index is important for the choices of our component models.

Table 1. Comparison of model complexity between AVC and NPD

Model	Weaks	Features	aveEval
LE	200	6193	24.754
RE	200	6193	24.754
LM	300	3561	26.755
RM	300	3561	26.755
LE+RE+LM+RM	1000	19508	103.018
NPD	1226	46401	114.507

The aveEval in LE and LM are 24.754 and 26.755 respectively (See Table 1). So the total number of features per detection window that AVC has to evaluate is 103.018 with symmetric detection considered, which is faster than NPD holistic face detector implemented in [23] that has 46401 weak classifiers and an aveEval of 114.507. With regard to pose-variations and occlusions, AVC also outperforms

NPD detector by a notable margin on FDDB (See Fig. 7(c)). Another advantage of AVC is that storage memory required is low compared to CNN methods, which is crucial for real-world applications. The total model size of AVC is only 2.65 MB, smaller compared to NPD (6.31 MB) or a typical CNN model.

7 Conclusion

In this paper, we proposed a new method called AVC highlighting component-based face detection, which addresses pose variations and occlusions simultaneously in a single framework with low complexity. We show a consistent component definition which helps to achieve pose-invariant component detection. To handle facial occlusions, we only detect visible facial components, and build a local to global aggregation strategy to detect the whole face adaptively. Experiments on the FDDB and AFW databases show that the proposed method is robust in handling illuminations, occlusions and pose-variations, achieving much better performance but lower model complexity compared to the corresponding holistic face detector. The proposed face detector is able to output local facial components as well as meanshape landmarks, which may be helpful in landmark detection initialization and pose estimation. We will leave it as future work for investigation.

Acknowledgement. This work was supported by the National Key Research and Development Plan (Grant No. 2016YFC0801002), the Chinese National Natural Science Foundation Projects #61672521, #61473291, #61572501, #61502491, #61572536, NVIDIA GPU donation program and AuthenMetric R&D Funds.

References

1. Jain, V., Learned-Miller, E.G.: FDDB: a benchmark for face detection in unconstrained settings. UMass Amherst Technical report (2010)
2. Wu, B., Ai, H., Huang, C., Lao, S.: Fast rotation invariant multi-view face detection based on real adaBoost. In: IEEE Conference on Automatic Face and Gesture Recognition (2004)
3. Li, S., Zhang, Z.: Floatboost learning and statistical face detection. IEEE Trans. Pattern Anal. Mach. Intell. **26**, 1112–1123 (2004)
4. Huang, C., Ai, H., Li, Y., Lao, S.: High-performance rotation invariant multiview face detection. IEEE Trans. Pattern Anal. Mach. Intell. **29**, 671–686 (2007)
5. Hotta, K.: A robust face detector under partial occlusion. In: International Conference on Image Processing (2004)
6. Lin, Y., Liu, T., Fuh, C.: Fast object detection with occlusions. In: Proceedings of the European Conference on Computer Vision, pp. 402–413 (2004)
7. Lin, Y., Liu, T.: Robust face detection with multi-class boosting (2005)
8. Chen, J., Shan, S., Yang, S., Chen, X., Gao, W.: Modification of the adaboost-based detector for partially occluded faces. In: 18th International Conference on Pattern Recognition (2006)

9. Goldmann, L., Monich, U., Sikora, T.: Components and their topology for robust face detection in the presence of partial occlusions. IEEE Trans. Inf. Forensics Secur. **2**, 559–569 (2007)
10. LeCun, Y., Bengio, Y.: Convolutional networks for images, speech, and time series. In: The Handbook of Brain Theory and Neural Networks, pp. 33–61 (1995)
11. Li, H., Lin, Z., Shen, X., Brandt, J., Hua, G.: A convolutional neural network cascade for face detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 5325–5334 (2015)
12. Farfade, S.S., Saberian, M.J., Li, L.J.: Multi-view face detection using deep convolutional neural networks. In: Proceedings of the 5th ACM on International Conference on Multimedia Retrieval, pp. 643–650. ACM (2015)
13. Yang, S., Luo, P., Loy, C.C., Tang, X.: From facial parts responses to face detection: a deep learning approach. In: Proceedings of the IEEE International Conference on Computer Vision, pp. 3676–3684 (2015)
14. Zhang, K., Zhang, Z., Li, Z., Qiao, Y.: Joint face detection and alignment using multi-task cascaded convolutional networks. arXiv preprint arXiv:1604.02878 (2016)
15. Ranjan, R., Patel, V.M., Chellappa, R.: Hyperface: a deep multi-task learning framework for face detection, landmark localization, pose estimation, and gender recognition. arXiv preprint arXiv:1603.01249 (2016)
16. Viola, P., Jones, M.: Robust real-time object detection. Int. J. Comput. Vis. **4**, 34–47 (2001)
17. Yang, B., Yan, J., Lei, Z., Li, S.Z.: Aggregate channel features for multi-view face detection. In: 2014 IEEE International Joint Conference on Biometrics (IJCB), pp. 1–8. IEEE (2014)
18. Zhu, X., Ramanan, D.: Face detection, pose estimation, and landmark localization in the wild. In: 2012 IEEE Conference on Computer Vision and Pattern Recognition (CVPR), pp. 2879–2886. IEEE (2012)
19. Yan, J., Lei, Z., Wen, L., Li, S.: The fastest deformable part model for object detection. In: Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, pp. 2497–2504 (2014)
20. Mathias, M., Benenson, R., Pedersoli, M., Gool, L.: Face detection without bells and whistles. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8692, pp. 720–735. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10593-2_47
21. Ranjan, R., Patel, V.M., Chellappa, R.: A deep pyramid deformable part model for face detection. In: 2015 IEEE 7th International Conference on Biometrics Theory, Applications and Systems (BTAS), pp. 1–8. IEEE (2015)
22. Chen, D., Ren, S., Wei, Y., Cao, X., Sun, J.: Joint cascade face detection and alignment. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8694, pp. 109–122. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10599-4_8
23. Liao, S., Jain, A., Li, S.: A fast and accurate unconstrained face detector. IEEE Trans. Pattern Anal. Mach. Intell. **38**, 211–223 (2016)
24. Goldmann, L., Mönich, U.J., Sikora, T.: Components and their topology for robust face detection in the presence of partial occlusions. IEEE Trans. Inf. Forensics Secur. **2**, 559–569 (2007)
25. Azizpour, H., Laptev, I.: Object detection using strongly-supervised deformable part models. In: Fitzgibbon, A., Lazebnik, S., Perona, P., Sato, Y., Schmid, C. (eds.) ECCV 2012. LNCS, vol. 7572, pp. 836–849. Springer, Heidelberg (2012). doi:10.1007/978-3-642-33718-5_60

26. Bourdev, L., Maji, S., Brox, T., Malik, J.: Detecting people using mutually consistent poselet activations. In: Daniilidis, K., Maragos, P., Paragios, N. (eds.) ECCV 2010. LNCS, vol. 6316, pp. 168–181. Springer, Heidelberg (2010). doi:10.1007/978-3-642-15567-3_13
27. Zhang, N., Paluri, M., Ranzato, M.A.: Panda: Pose aligned networks for deep attribute modeling. In: Computer Vision and Pattern Recognition, pp. 1637–1644. IEEE, Springer, Berlin, Heidelberg (2014)
28. Zhang, N., Donahue, J., Girshick, R., Darrell, T.: Part-based R-CNNs for fine-grained category detection. In: Fleet, D., Pajdla, T., Schiele, B., Tuytelaars, T. (eds.) ECCV 2014. LNCS, vol. 8689, pp. 834–849. Springer, Heidelberg (2014). doi:10.1007/978-3-319-10590-1_54
29. Zhang, H., Xu, T., Elhoseiny, M., Huang, X., Zhang, S., Elgammal, A., Metaxas, D.: SPDA-CNN: Unifying semantic part detection and abstraction for fine-grained recognition. In: The IEEE Conference on Computer Vision and Pattern Recognition (CVPR) (2016)
30. Köstinger, M., Wohlhart, P., Roth, P.M., Bischof, H.: Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization. In: 2011 IEEE International Conference on Computer Vision Workshops (ICCV Workshops), pp. 2144–2151. IEEE (2011)
31. Bourdev, L., Brandt, J.: Robust object detection via soft cascade. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 2, pp. 236–243. IEEE (2005)
32. Liao, S., Jain, A.K., Li, S.Z.: Unconstrained face detection. Technical report, MSU-CSE-12-15, Department of Computer Science, Michigan State University (2012)
33. Dalal, N., Triggs, B.: Histograms of oriented gradients for human detection. In: IEEE Computer Society Conference on Computer Vision and Pattern Recognition, CVPR 2005, vol. 1, pp. 886–893. IEEE (2005)