Neurocomputing 221 (2017) 138–145

Contents lists available at ScienceDirect

Neurocomputing

journal homepage: www.elsevier.com/locate/neucom

Funnel-structured cascade for multi-view face detection with alignment-awareness

Shuzhe Wu[a], Meina Kan[a,b], Zhenliang He[a], Shiguang Shan[a,b,*], Xilin Chen[a]

[a] Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China
[b] CAS Center for Excellence in Brain Science and Intelligence Technology, China

ARTICLE INFO

Communicated by Zhou Xiuzhuang

Keywords:
Face detection
Multi-view
Cascade
Multilayer perceptron

ABSTRACT

Multi-view face detection in open environment is a challenging task due to diverse variations of face appearances and shapes. Most multi-view face detectors depend on multiple models and organize them in parallel, pyramid or tree structure, which compromise between the accuracy and time-cost. Aiming at a more favorable multi-view face detector, we propose a novel funnel-structured cascade (FuSt) detection framework. In a coarse-to-fine flavor, our FuSt consists of, from top to bottom, (1) multiple view-specific fast LAB cascade for extremely quick face proposal, (2) multiple coarse MLP cascade for further candidate window verification, and (3) a unified fine MLP cascade with shape-indexed features for accurate face detection. Compared with other structures, on the one hand, the proposed one uses multiple computationally efficient distributed classifiers to propose a small number of candidate windows but with a high recall of multi-view faces. On the other hand, by using a unified MLP cascade to examine proposals of all views in a centralized style, it provides a favorable solution for multi-view face detection with high accuracy and low time–cost. Besides, the FuSt detector is alignment-aware and performs a coarse facial part prediction which is beneficial for subsequent face alignment. Extensive experiments on two challenging datasets, FDDB and AFW, demonstrate the effectiveness of our FuSt detector in both accuracy and speed.

1. Introduction

Fast and accurate detection of human faces is greatly demanded in various applications. While current detectors can easily detect frontal faces, they become less satisfactory when confronted with complex situations, e.g. to detect faces viewed from various angles, in low resolution, with occlusion, etc. Especially, the multi-view face detection is quite challenging, because faces can be captured almost from any angle – even exceeding 90° in extreme cases, leading to significant divergence in facial appearances and shapes.

Along with the steady progress of face detection, there have been mainly three categories of face detectors with different highlights. The most classic are those following the boosted cascade framework [1–3], originating in the seminal work of Viola and Jones [4]. These detectors are quite computationally efficient, benefited from the attentional cascade and fast feature extraction. Then to explicitly deal with large appearance variations, deformable part models (DPM) [5] are introduced to simultaneously model global and local face features [6–8], providing an intuitive way to cover intra-class variations and thus being more robust to deformations due to pose, facial expressions, etc. DPM has established a reputation for its promising results on challenging datasets, but detection with DPM is time-consuming, inspiring researches on speeding up techniques [7]. Recently, detectors based on neural networks, e.g. convolutional neural networks (CNN) [9–14], have attracted much attention and achieved magnificent accuracy on the challenging FDDB dataset [15], as they enjoy the natural advantage of strong capability in non-linear feature learning. The weakness of CNN-based detectors is their high computational cost due to intensive convolution and complex nonlinear operations.

Most works mentioned above focus on designing an effective detector for generic faces without considerations for specific scenarios such as multi-view face detection. In order to handle faces in different views, a straightforward solution is to use multiple face detectors in parallel [2,1,8], one for each view, as shown in Fig. 1a. The parallel structure requires each candidate window to be classified by all models, resulting in an increase of the overall computational cost and false alarm rate. To alleviate this issue, each model needs to be elaborately trained and tuned for better discrimination between face and non-face windows, ensuring faster and more accurate removal of non-face windows.

[*] Corresponding author at: Key Laboratory of Intelligent Information Processing of Chinese Academy of Sciences (CAS), Institute of Computing Technology, CAS, Beijing 100190, China.

http://dx.doi.org/10.1016/j.neucom.2016.09.072
Received 30 January 2016; Received in revised form 10 July 2016; Accepted 10 September 2016
Available online 01 October 2016

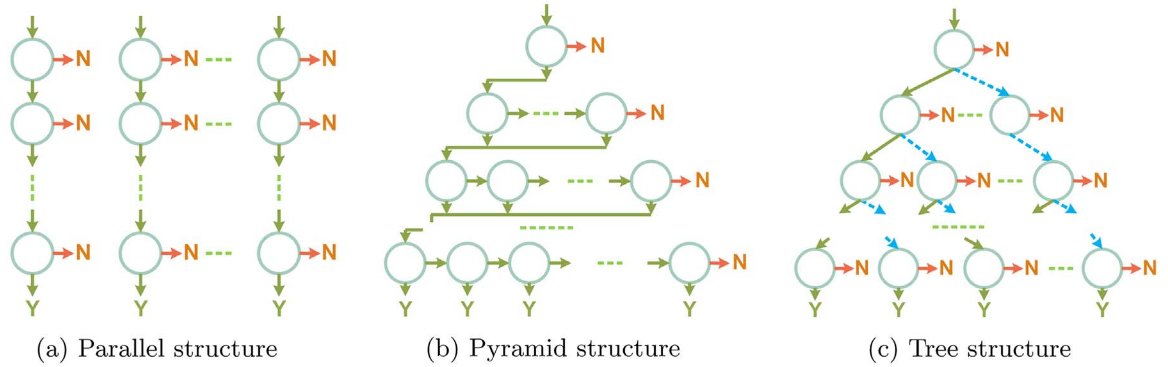


Fig. 1. Different structures for multi-view face detection.

More efficiently, the multiple models for multi-view face detection can be organized in a pyramid [16] or tree structure [17], as shown in Fig. 1b and c, forming a coarse-to-fine classification scheme. In such structures, the root classifier performs the binary classification of face vs. non-face, and then at subsequent layers, faces are divided into multiple sub-categories with respect to views in a finer granularity, each of which is handled by an independent model. The pyramid structure is actually a compressed parallel structure with shared nodes in higher layers or a stack of parallel structures with different view partitions. Therefore the pyramid-structured detectors suffer from similar problems that parallel-structured ones are faced with. The tree-structured detectors are different in that branching schemes are adopted to avoid evaluating all classifiers at each layer, but this can easily lead to missing detections with incorrect branching. To relax the dependence on accurate branching, Huang et al. [17] design a vector boosting algorithm to allow multiple branching.

Considering the appearance divergence of multi-view faces from the perspective of feature representation, the intra-class variations are mainly due to features extracted at positions with inconsistent semantics. For instance, in Fig. 2, three faces in different views are shown and the window at the same positions on different faces contains completely distinct semantics, resulting in features describing eye, nose and cheek respectively. Thus there does not exist a good correspondence between representations of faces in different views. Chen et al. [3] compare densely extracted features with shape-indexed features and finds the latter to be more discriminative. By using features at aligned landmarks, faces in different views can be more compactly represented and better distinguished from non-face regions.

To provide a more effective framework for multi-view face detection, we design a novel funnel-structured cascade (FuSt) multi-view face detector, which enjoys both high accuracy and fast speed. The FuSt detector, as shown in Fig. 3, features a funnel-like structure, being wider on the top and narrower at the bottom, which is evidently different from previous ones. At early stages from the top, multiple fast but coarse classifiers run in parallel to rapidly remove a large proportion of non-face windows. Each of the parallel classifiers is trained specifically for faces within a small range of views, so they are able to ensure a high recall of multi-view faces. By contrast, at subsequent stages, fewer classifiers, which are slightly more time-consuming but with higher discriminative capability, are employed to verify the remaining candidate windows. Gathering the small number of windows surviving from previous stages, at the last stages at the bottom, a unified multilayer perceptron (MLP) cascade with shape-indexed features is leveraged to output the final face detection results. From top to bottom, the number of models used decreases while the model complexity and discriminative capability increase, forming a coarse-to-fine framework for multi-view face detection (Fig. 11).

Compared with previous multi-view face detectors, the proposed FuSt detector is superior in that a more effective framework is used to organize multiple models. The contribution of our work compared to existing literature is listed as below.

- First, a unified MLP cascade is leveraged as last few stages to examine proposals provided by previous stages, which addresses the problem of increased false alarm rate resulting from using multiple models in other structures, e.g. parallel or tree structure.
- Second, the proposed FuSt detector operates in a gathering style instead of adopting any branching mechanism as in pyramid- or tree-structured detectors. Therefore it can naturally avoid missing detections caused by incorrect branching and reach a high recall.
- Third, in the final unified MLP cascade, features are extracted in semantically consistent positions by integrating shape information rather than fixed positions as in conventional face detectors, and thus multi-view faces can be better distinguished from non-face regions. Moreover, the extra shape output from our FuSt detector can provide a good initialization for subsequent alignment.
- With extensive experiments on challenging face detection datasets including FDDB [15] and AFW [6], the FuSt detector is demonstrated to have both good performance and fast speed.

The rest of the paper is organized as follows. Section 2 describes the proposed FuSt detector in detail, explaining the design of different stages from top to bottom. Section 3 presents the experimental results on two challenging face detection datasets together with analysis on the structure and shape prediction. Finally Section 4 concludes the paper and discusses the future work.

2. Funnel-structured cascade multi-view face detector

An overview of the framework of FuSt detector is presented in Fig. 3. Specifically, the FuSt detector consists of three coarse-to-fine stages in consideration of both detection accuracy and computational cost, i.e. Fast LAB Cascade classifier, Coarse MLP Cascade classifier, and Fine MLP Cascade classifier. An input image is scanned according to the sliding window paradigm, and each window goes through the detector stage by stage.

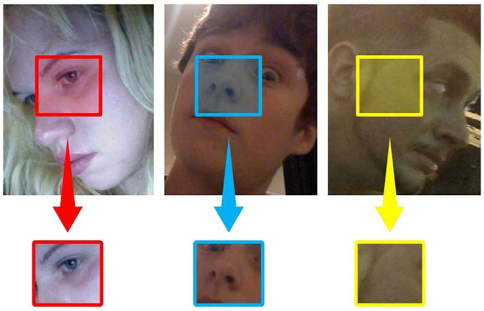


Fig. 2. The window at the same position on three faces in varied views contain totally distinct semantics.

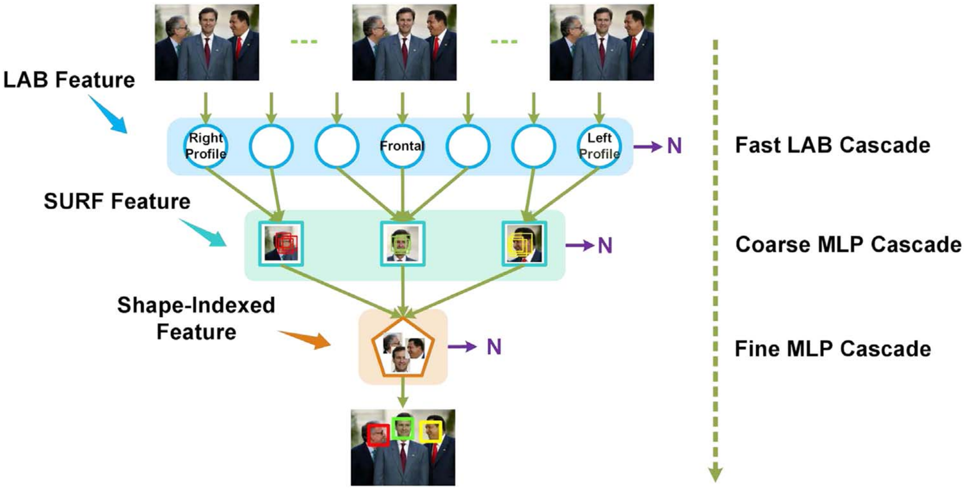


Fig. 3. An overview of our proposed funnel-structured cascade framework for multi-view face detection.

The Fast LAB Cascade classifiers aim to quickly remove most non-face windows while retaining a high call of face windows. The following Coarse MLP Cascade classifiers further roughly refine the candidate windows at a low cost. Finally the unified Fine MLP Cascade classifiers accurately determine faces with the expressive shape-indexed features. In addition, it also predicts landmark positions which are beneficial for subsequent alignment.

2.1. Fast LAB cascade

For real-time face detection, the major concern in the sliding window paradigm is the large quantity of candidate windows to be examined. For instance, to detect faces with sizes larger than 20×20 on a 640×480 image, over a million windows need to be examined. Hence it is quite necessary to propose a small number of windows that are most likely to contain faces at minimal time cost.

A good option for fast face proposal is to use boosted cascade classifiers, which are very efficient for face detection task as shown by Viola and Jones [4]. Yan et al. [18] propose an efficient LAB (Locally Assembled Binary) feature, which only considers the relative relations between Haar features, and can be accelerated with a look-up table. Extracting an LAB feature in a window requires only one memory access, resulting in constant time complexity of $O(1)$. Therefore we employ the more preferable LAB feature with boosted cascade classifiers, leading to the extremely fast LAB cascade classifiers, which are able to rapidly reject a large proportion of non-face windows at the very beginning.

Although the LAB feature is quite computationally efficient, it is less expressive and has difficulty modeling the complicated variations of multi-view faces for a high recall of face windows. Therefore, we adopt a divide-and-conquer strategy by dividing the difficult multi-view face detection problem into multiple easier single-view face detection problems. Specifically, multiple LAB cascade classifiers, one for each view, are leveraged in parallel and the final candidate face windows are the union of surviving windows from all of them.

Formally, denote the whole training set containing multi-view faces as S, and it is partitioned into v subsets according to view angles, denoted as S_i, $i = 1, 2, \ldots, v$. With each S_i, an LAB cascade classifier c_i is trained, which attempts to detect faces in the i-th view angle. For a window x within an input image, whether it is possible to be a face is determined with all LAB cascade classifiers as follows:

$$y = c_1(x) \vee c_2(x) \vee \cdots \vee c_v(x), \tag{1}$$

where $y \in \{0, 1\}$ and $c_i(x) \in \{0, 1\}$ indicate whether x is determined to be a face or not. As can be seen from Eq. (1), a window will be rejected if and only if it is classified as negative by all LAB cascade classifiers. Using multiple models will cost more time, but all models can share the same LAB feature map for feature extraction. Therefore more models add only minor cost and the overall speed is still very fast as a high recall is reached.

Besides the high recall, the parallel structure also allows more flexibility in view partitions. Since it does not suffer from missing detections caused by incorrect branching as in tree structure, a rough rather than an accurate view partition is enough. In other words, degenerated partitions with incorrect view labeling of faces has minor influences on the overall recall of all LAB cascade classifiers. It is even applicable for automatic view partition from clustering or that based on other factors.

2.2. Coarse MLP cascade

After the stages of LAB cascade, most of the non-face windows have been discarded, and the remaining ones are too hard for the simple LAB feature to handle. Therefore, on subsequent stages, the candidate windows are further verified by more sophisticated classifiers, i.e. MLP with SURF (Speeded-Up Robust Feature) [19]. To avoid imposing too much computational cost, small networks are exploited to perform a better but still coarse examination.

SURF features are more expressive than LAB features, but are still computationally efficient benefited from the integral image trick. Therefore face windows can be better differentiated from non-face windows with low time cost. Furthermore, MLP is used with SURF feature for window classification, which can better model the non-linear variations of multi-view faces and diverse non-face patterns with the equipped nonlinear activation functions.

MLP is a type of neural network consisting of an input layer, an output layer, and one or more hidden layers in between. An n-layer MLP $F(\cdot)$ can be formulated as

$$F(x) = f_{n-1}(f_{n-2}(\cdots f_1(x))), \tag{2}$$

$$f_i(z) = \sigma(W_i z + b_i). \tag{3}$$

where x is the input, i.e. the SURF features of a candidate window; W_i and b_i are the weights and biases of connections from layer i to $i + 1$ respectively. The activation function $\sigma(\cdot)$ is commonly designed as a nonlinear function such as a sigmoid function $\sigma(x) = \frac{1}{1+e^{-x}}$. As can be seen in Eqs. (2) and (3), units in hidden layers and output layer are both equipped with nonlinear functions, so the MLP is endowed with strong capability to model highly nonlinear transformations. The training of MLPs aims to minimize the mean squared error between the predictions and the true labels as below

$$\min_{F} \sum_{i=1}^{n} \| F(x_i) - y_i \|^2, \tag{4}$$

where x_i is the feature vector of the i-th training sample and y_i the corresponding label as either 1 or 0, representing whether the sample is a face or not. The problem in Eq. (4) can be easily solved by using gradient descent under the back propagation framework [20].

We employ multiple coarse MLPs to construct an attentional cascade, in which the number of features used and the size of the network gradually increase stage by stage. The SURF features used at each stage are selected by using group sparse [21]. Since the MLP cascade classifiers have stronger ability to model face and non-face variations, windows passing through multiple LAB cascade classifiers can be handled together by one model, i.e. one MLP cascade can connect to multiple LAB cascade classifiers.

2.3. Fine MLP cascade with shape-indexed feature

Surviving from the previous stages, the small number of windows have been quite challenging, among which face and non-face windows are more difficult to be distinguished. Considering that multiple models running in parallel tend to introduce more false alarms, it is desirable to process the remaining windows in a unified way. Hence we leverage one single MLP cascade following the previous Coarse MLP Cascade classifiers.

Prominent divergence exists in appearances of multi-view faces, which is mainly due to the unaligned features, i.e. features are extracted at positions that are not semantically consistent. For example, the central region of a frontal face covers the nose, while that of a profile face is part of the cheek, as shown in Fig. 2. To address this issue, we adopt shape-indexed features extracted at semantically consistent positions as the input of the Fine MLP Cascade classifier. As shown in Fig. 5, four semantic positions are selected, corresponding to the facial landmarks of left and right eye center, nose tip and mouth center. For profile faces, the invisible eye is assumed to be at the same position as the other eye. The SIFT (Scale-Invariant Feature Transform) [22] feature is computed at each semantic position on candidate windows, and they are robust to large face variations such as pose, translation, etc.

With the more expressive shape-indexed features, larger MLPs with higher capacity of nonlinearity are used to perform finer discrimination between face and non-face windows. Moreover, different from previous ones, the larger MLPs predict both class label, indicating whether a candidate window is a face, and shape simultaneously. An extra term of shape prediction errors is added to the objective function in Eq. (4). The new optimization problem is the following:

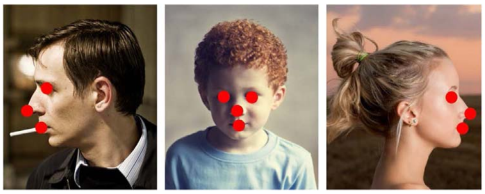


Fig. 5. The four semantic positions (landmarks) used to extract shape-indexed feature: left and right eye center, nose tip and mouth center.

$$\min_{F} \sum_{i=1}^{n} \left\| F_c(\phi(x_i, \hat{s}_i)) - y_i \right\|^2 + \lambda \sum_{i=1}^{n} \left\| F_s(\phi(x_i, \hat{s}_i)) - s_i \right\|_2^2, \tag{5}$$

where F_c corresponds to the face classification output, and F_s the shape prediction output; $\phi(x_i, \hat{s}_i)$ indicates the shape-indexed feature (i.e. SIFT) extracted from the i-th training sample x_i according to a mean shape or predicted shape $\hat{s}_i$; s_i is the groundtruth shape for the sample; λ is the weighting factor to maintain the balance between the two types of errors, which is set to $\frac{1}{d}$ with d as the dimension of shape. As can be seen from Eq. (5), a more accurate shape $F_s(\phi(x_i, \hat{s}_i))$ than the input $\hat{s}_i$ can be obtained with the MLP. Hence a subsequent model can exploit more compact shape-indexed features extracted according to the refined shape $F_s(\phi(x_i, \hat{s}_i))$. As so, in multiple cascaded MLPs, the shapes used for feature extraction become more and more accurate stage by stage, leading to more and more distinctive shape-indexed features and further making multi-view faces more distinguishable from non-face regions. The process is shown in Fig. 4.

Additionally, predicting shapes has made the detector alignment-aware in the sense that an alignment model can be initialized with landmark coordinates directly instead of bounding boxes of detected faces.

3. Experiments

To evaluate the proposed FuSt detector for multi-view face detection, as well as to analyse the detector in various aspects, extensive experiments are performed on two challenging face datasets.

3.1. Experimental settings

The most popular dataset for evaluating face detectors is the FDDB [15]. It contains 5,171 labeled faces from 2845 news images. FDDB is challenging in the sense that the labeled faces appear with great variations in view, skin color, facial expression, illumination, occlusion, resolution, etc.

Another widely used face detection dataset is the AFW [6]. This set

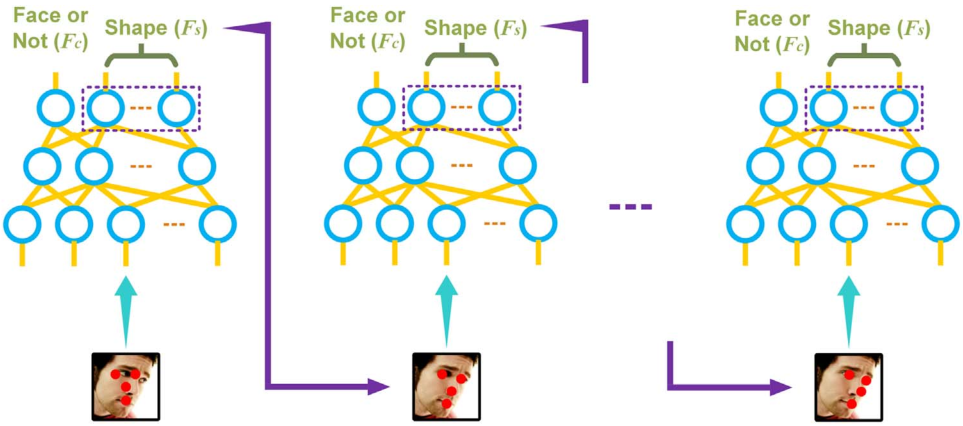


Fig. 4. The Fine MLP Cascade with shape-indexed feature. The input of each stage of MLP is the shape-indexed feature extracted according to the shape predicted by the previous stage (or mean shape for the first stage). The output includes the class label indicating whether the window is a face or not as well as a more accurate shape, which is used to extract more distinctive shape-indexed features for the next stage.

contains 205 images from Flickr with 468 faces. It is a small set, yet is challenging, since faces appears in cluttered backgrounds and with large variations in viewpoints.

For evaluation of the detection accuracy, we apply the officially provided tool to our detection results on FDDB to obtain the ROCs, and draw precision–recall curve for the results on AFW, following most existing works.

For the training data of the FuSt detector, we use faces from MSRA-CFW [23], PubFig [24], and AFLW [25] as positive samples, and randomly crop patches from 40,000 collected images not containing faces as negative samples. To augment the training set with more variations, we add random distortions to the face samples. Besides, all samples are resized to 40×40 for training.

We use 1 stage with a total of 150 LAB features for the Fast LAB Cascade, and 3 stages for the Coarse MLP Cascade, which exploit 2, 4 and 6 SURF features respectively. SURF features are extracted based on local patches, which will cover redundant information if there is considerable overlap between them. Therefore a large step of 16 are chosen for adjacent SURF patches, resulting in a pool of 56 SURF features on a 40×40 sample image. The three stages of MLP all have only one hidden layer, and there are 15 hidden units in the first-stage MLP and 20 hidden units in the second- and third-stage MLP. The final Fine MLP Cascade contains 2 stages of single-hidden-layer MLP with 80 hidden units with SIFT features extracted around the four semantic positions as mentioned in Section 2.3.

3.2. Analysis of the funnel-structured cascade

We first conduct a detailed analysis of the proposed FuSt detector to evaluate its performance from various perspectives. Specifically, we compare different view partitions, verify the effectiveness of shape-indexed features, assess the accuracy of shape predictions, and compare the final MLP cascade with two widely used CNN models.

Different view partitions: At the beginning, we adopt a divide-and-conquer strategy to treat faces in different views with separate LAB cascade classifiers. This makes it possible for such simple classifiers to reject a large proportion of non-faces windows, while retaining a high overall recall of faces. To explore the impact of different view partitions, we compare two typical partition schemes: (1) five-view partition, i.e. *left full profile, left half profile, near frontal, right half profile, and right full profile*; (2) two-view partition, i.e. *near frontal, profile*. Note that in the second two-view partition scheme, left and right profile faces are mixed together, and half profile faces are mixed with frontal ones. To supplement the training set with more half profile face images, we also use some images from CelebA dataset [26]. The recall of faces with the two schemes are presented in Table 1. Here we manually partition the FDDB into two subsets of profile and frontal faces to evaluate on them separately. The former contains 527 profile faces from 428 images, and the latter, i.e. the frontal face subset, contains the rest faces including both near frontal and some half profile faces.

As can be seen, the recall of faces with the five-view partition, especially the recall of profile faces, are higher than that with the two-view partition when both scheme remove over 99% of candidate windows. As expected, the finer partition allows classifiers to cover more variations within each view of faces, and is beneficial for obtaining higher recall. This demonstrates the effectiveness of using a reasonably wide top in the proposed funnel structure.

Funnel structure vs parallel structure: To demonstrate the effectiveness of the proposed funnel structure employing a unified model to handle candidate windows coming from different classifiers, we compare the parallel and the funnel structure on frontal and half profile faces in the coarse MLP cascade stage. Specifically, for the parallel structure, we train three MLPs, one for each of the three views, which follows the corresponding fast LAB cascade. For the funnel structure, only one MLP is trained for frontal, left half profile and right half profile faces. The parallel structure obtains a recall of 94.41% with 297.06 windows per image, while the funnel structure reaches a higher recall of 94.43% with only 268.10 windows per image. This demonstrates that a unified model can effectively control the false positives with less sacrifice of recall.

Shape-indexed feature: To verify the effectiveness of the shape-indexed feature, we train two types of two-stage Fine MLP Cascade classifiers with mean shape and refined shape respectively, which are used to extract shape-indexed feature. Namely, one MLP cascade uses SIFT extracted according to mean shape as input at both stages, while the other uses SIFT extracted with refined and thus more accurate shapes as input at the second stage.

Fixing previous stages, we compare the two types of Fine MLP Cascades on FDDB. The performance curves are presented in Fig. 6. As expected, using more accurate shapes brings performance gain, demonstrating the effectiveness of shape-indexed features for multi-view faces. Shape-indexed features from two faces have good semantic consistence, thus reducing intra-class variations and increasing inter-class distinctions. This makes it easier to distinguish face from non-face windows.

We also evaluate the coarse shape predictions on AFW. Fig. 7 compares the predicted shape with the mean shape. With only two stages of refinement, the predicted shapes achieve significant improvement over the mean shape, leading to more semantically consistent shape-indexed features. When followed by an alignment model, the predicted shape from our FuSt detector can be directly used as a good initialization, which is more preferable than only bounding boxes of detected faces. Fig. 8 gives several examples of predicted shapes on faces in different views.

MLP vs CNN: The powerful CNN models have achieved good results in face detection task [9–11], so we also compare MLP with CNN under the proposed funnel-structured cascade framework. Two commonly used CNN models are considered in the comparison, i.e. LeNet [27] and AlexNet [28], and they serve as replacements for the final Fine MLP Cascade. The input sizes of LeNet and AlexNet are $40 \times 40 \times 3$ and $256 \times 256 \times 3$ respectively, and the output layers are adjusted for two-class classification of face or non-face. Both CNN models are fine-tuned using the same data as that used in training the MLP cascade. The performance curves on FDDB are given in Fig. 9. As is shown, the MLP cascade outperforms LeNet by a large margin and also performs better than the 8-layer AlexNet. This is most likely because the semantically consistent shape-indexed features are more effective than the learned convolutional features. Considering the result that the MLP with hand-crafted features has the ability to defeat deep CNN models, it implies that a well-designed model with considerations for the problem can be better than an off-the-shelf CNN.

Detection speed: Our FuSt detector enjoys a good advantage of detection speed with the coarse-to-fine framework design and is faster than complex CNN-based detectors. When detecting faces no smaller than 80×80 on a VGA image of size 640×480, our detector takes 50 ms with step-1 sliding window using a single thread on an i7 CPU. The Fast LAB Cascade and Coarse MLP Cascade cost only 30 ms, and the final Fine MLP Cascade 20 ms. By contrast, Cascade CNN takes 110 ms over an image pyramid with scaling factor of 1.414 on CPU [10]. Moreover, further speed-up of FuSt detector can be easily obtained with GPU since a large amount of data parallelism exists in our framework, e.g. feature extraction for each window, the inner product

Table 1
Recall of faces with different view partitions with over 99% windows removed.

View partition	Recall of faces (%)		
	Frontal	Profile	Overall
5 Views	96.27	95.83	96.15
2 Views	95.07	92.60	94.82

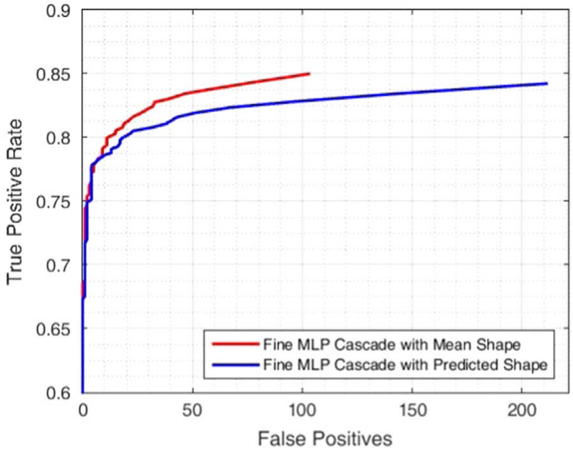


Fig. 6. Comparison between shape-indexed features extracted with mean shape and refined shape.

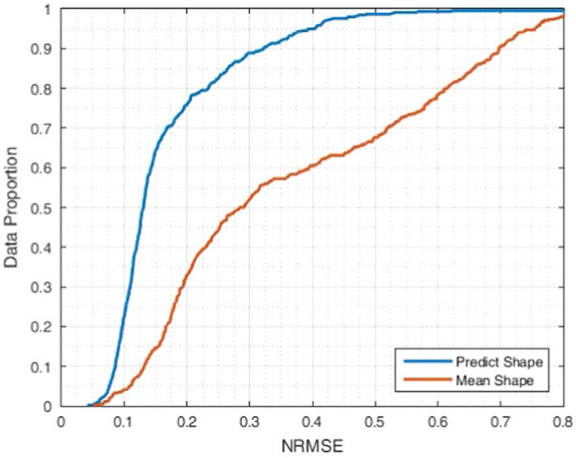


Fig. 7. Comparison between predicted shape and mean shape on AFW.

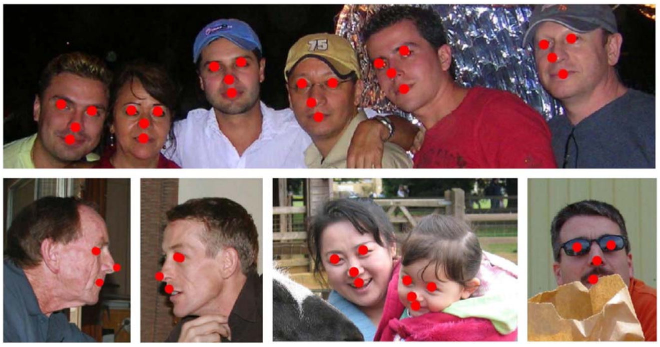


Fig. 8. Examples of predicted shapes on AFW.

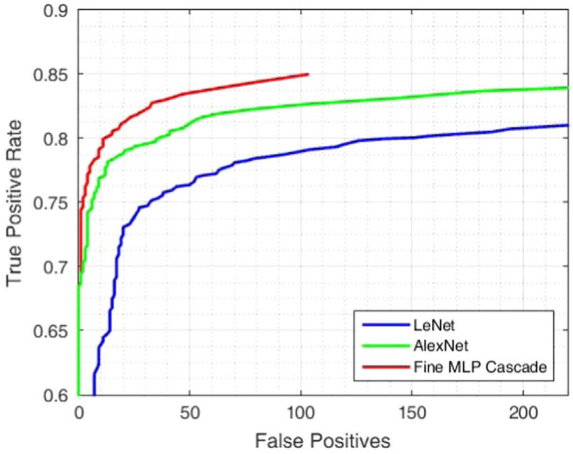


Fig. 9. Comparison of MLP cascade, LeNet and AlexNet.

Table 2
Comparison with Cascade CNN [10] in different aspects. The DR@100FPs is computed on FDDB, and the speed is compared with minimum face size set as 80×80 and image size 640×480.

Methods	DR@100FPs	Speed	Landmark prediction
Cascade CNN [10]	85%	110 ms	No
Our FuSt	85%	50 ms	Yes

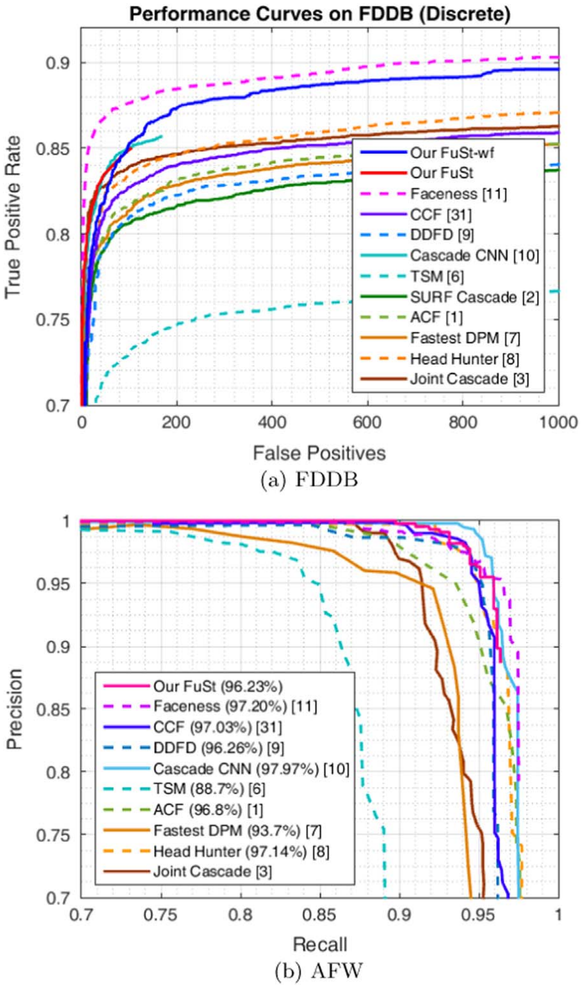


Fig. 10. Comparison with the state-of-the-art on two face detection datasets: (a) FDDB and (b) AFW.

operations in MLP, etc. (Table 2)

Discussion: Compared with CNN based methods, the proposed funnel structure is a general framework of organizing multiple models, adopting a divide-and-conquer strategy to handle multi-view faces. The MLPs used with the framework can also be replaced by CNNs. One other aspect that makes our FuSt detector different is that hand-crafted shape-indexed feature is adopted based on explicit consideration for semantically consistent feature representation. By contrast, CNN learns the feature representation merely from data without considering the semantic consistency.

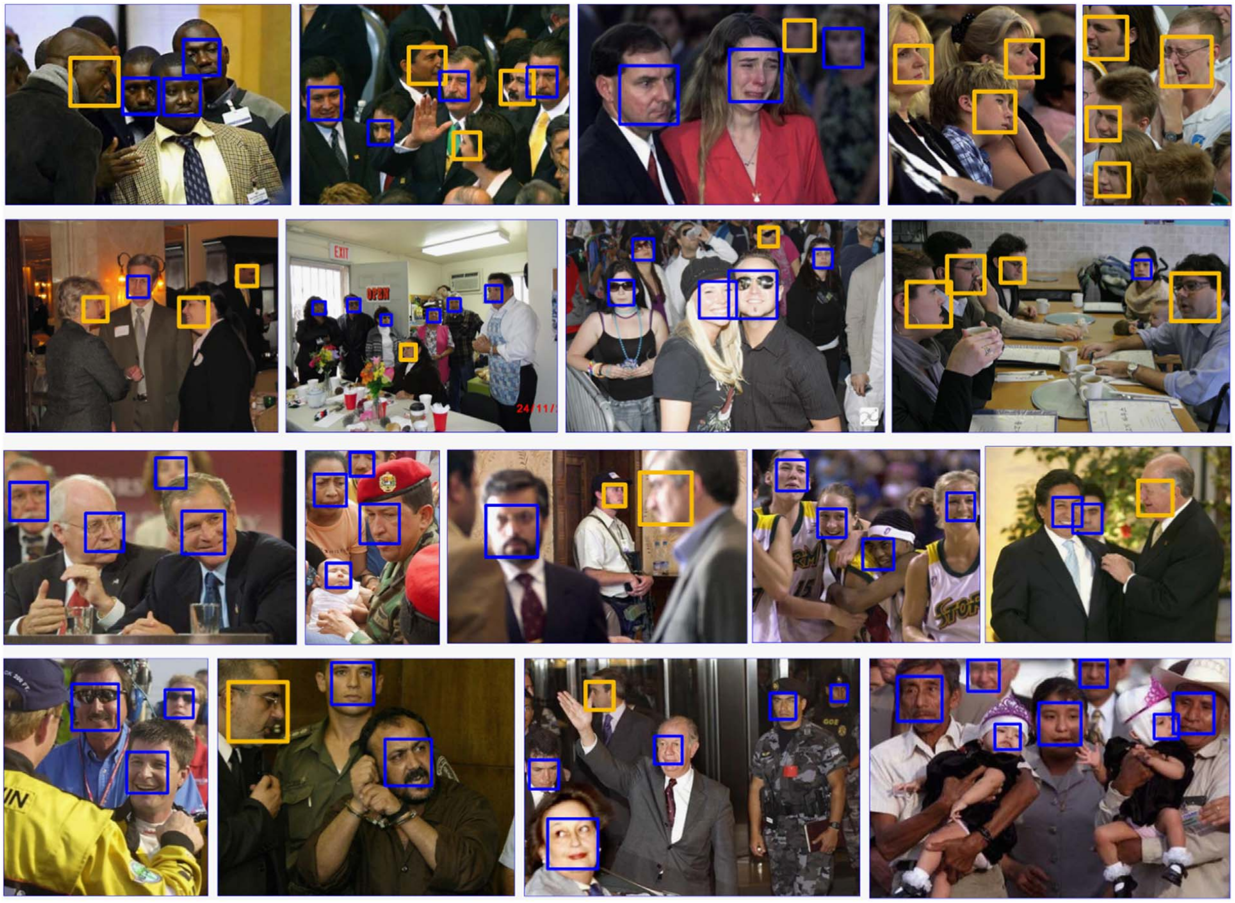


Fig. 11. Examples of detections on FDDB and AFW (blue: near frontal faces, orange: profile faces). (For interpretation of the references to color in this figure legend, the reader is referred to the web version of this article.)

3.3. Comparison with the state-of-the-art

To further evaluate the performance of the FuSt detector on multi-view face detection, we compare it with the state-of-the-art methods on FDDB and AFW as shown in Fig. 10. Methods being compared include cascade-structured detectors such as Joint Cascade [3], ACF [1], SURF Cascade [2], and Head Hunter [8], DPM-based detectors such as Fastest DPM [7], and TSM [6], and deep-network-based detectors such as DDFD [9], Cascade CNN [10], CCF [29], and FacenessNet [11].

Compared with multi-view face detectors like SURF Cascade, ACF, and Head Hunter, which all employ a parallel structure, our FuSt detector performs better on FDDB, indicating the superiority of our funnel structure. With as few as 100 false positives, the FuSt detector achieves a high recall of 85%, which is quite favorable in practical applications. Compared with the impressive deep-network-based methods, we achieve comparable performance with that of Cascade CNN. However, as stated in Section 3.2, our FuSt detector enjoys a more favorable speed, taking only 50 ms to detect a VGA image with a single thread on CPU. By contrast, Cascade CNN costs 110 ms on CPU. On AFW dataset, our PR curve is comparable to or better than most methods, further demonstrating that our FuSt detector is favorable for multi-view face detection.

To further investigate the potential of our FuSt detector on FDDB, we trained a new detector FuSt-wf with a more diverse dataset WIDER FACE [30]. WIDER FACE dataset covers much more face variations, which is beneficial for obtaining higher performance. Since WIDER FACE does not provide landmark annotations for faces, we only trained one stage for the unified MLP cascade with mean shape. As shown in Fig. 10, FuSt-wf achieves obvious performance boost, further demonstrating the effectiveness of the funnel-structure design. With higher quality and more data, the FuSt detector can continue to improve.

4. Conclusions and future works

In this paper, we have proposed a novel multi-view face detection framework, i.e. the funnel-structured cascade (FuSt), which has a coarse-to-fine flavor and is alignment-aware. The proposed FuSt detector operates in a gathering style, with the early stages of multiple parallel models reaching a high recall of faces at low cost and the final unified MLP cascade well reducing false alarms. As evaluated on two challenging datasets, the FuSt detector has shown good performance, and the speed of the detector is also quite favorable. In addition, the alignment-awareness nature of our FuSt detector can be leveraged to achieve a good initial shape for subsequent alignment models with minor cost.

For the future work, the funnel structure framework can be further enhanced with specifically designed CNN models which have good capability of learning feature representations automatically from data. It is also worth trying different hand-crafted shape-indexed features, e.g. the multi-scale pixel difference features used in [3], and comparing them with CNN-learned features. Considering the alignment-awareness nature of the FuSt detector, it is also a promising direction to design a joint face detection and alignment framework.

Acknowledgements

This work was partially supported by 973 Program under Contract no. 2015CB351802, Natural Science Foundation of China under Contract nos. 61173065, 61222211, 61402443 and 61390511, and Strategic Priority Research Program of the CAS under Contract no.

XDB02070004.

References

[1] B. Yang, J. Yan, Z. Lei, S.Z. Li, Aggregate channel features for multi-view face detection, in: IEEE International Joint Conference on Biometrics (IJCB), 2014, pp. 1–8.

[2] J. Li, Y. Zhang, Learning SURF cascade for fast and accurate object detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2013, pp. 3468–3475.

[3] D. Chen, S. Ren, Y. Wei, X. Cao, J. Sun, Joint cascade face detection and alignment, in: European Conference on Compute Vision (ECCV), 2014, pp. 109–122.

[4] P. Viola, M.J. Jones, Robust real-time face detection, Int. J. Comput. Vis. (IJCV) 57 (2) (2004) 137–154.

[5] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Trans. Pattern Anal. Mach. Intell. 32 (9) (2010) 1627–1645.

[6] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2879–2886.

[7] J. Yan, Z. Lei, L. Wen, S.Z. Li, The fastest deformable part model for object detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2014, pp. 2497–2504.

[8] M. Mathias, R. Benenson, M. Pedersoli, L. Van Gool, Face detection without bells and whistles, in: European Conference on Compute Vision (ECCV), 2014, pp. 720–735.

[9] S.S. Farfade, M. Saberian, L.-J. Li, Multi-view face detection using deep convolutional neural networks, in: International Conference on Multimedia Retrieval (ICMR), 2015.

[10] H. Li, Z. Lin, X. Shen, J. Brandt, G. Hua, A convolutional neural network cascade for face detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2015.

[11] S. Yang, P. Luo, C.C. Loy, X. Tang, From facial parts responses to face detection: A deep learning approach, in: IEEE International Conference on Computer Vision (ICCV), 2015.

[12] S. Zhan, Q.-Q. Tao, X.-H. Li, Face detection using representation learning, Neurocomputing 187 (2016) 19–26.

[13] Q.-Q. Tao, S. Zhan, X.-H. Li, T. Kurihara, Robust face detection using local CNN and SVM based on kernel combination, Neurocomputing 211 (2016) 98–105.

[14] X. Jiang, Y. Pang, X. Li, J. Pan, Speed up deep neural network based pedestrian detection by sharing features across multi-scale models, Neurocomputing 185 (2016) 163–170.

[15] V. Jain, E. Learned-Miller, FDDB: A Benchmark for Face Detection in Unconstrained Settings, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.

[16] S.Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, H. Shum, Statistical learning of multi-view face detection, in: European Conference on Compute Vision (ECCV), 2002, pp. 67–81.

[17] C. Huang, H. Ai, Y. Li, S. Lao, High-performance rotation invariant multiview face detection, IEEE Trans. Pattern Anal. Mach. Intell. 29 (4) (2007) 671–686.

[18] S. Yan, S. Shan, X. Chen, W. Gao, Locally assembled binary (LAB) feature with feature-centric cascade for fast and accurate face detection, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2008, pp. 1–7.

[19] J. Li, T. Wang, Y. Zhang, Face detection using SURF cascade, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2011, pp. 2183–2190.

[20] M. Schmidt, minFunc: Unconstrained Differentiable Multivariate Optimization in Matlab, ⟨http://www.cs.ubc.ca/~schmidtm/Software/minFunc.html, 2005.

[21] Y.C. Eldar, P. Kuppinger, H. Bolcskei, Block-sparse signals: uncertainty relations and efficient recovery, IEEE Trans. Signal Process. 58 (6) (2010) 3042–3054.

[22] D.G. Lowe, Distinctive image features from scale-invariant keypoints, Int. J. Comput. Vis. 60 (2) (2004) 91–110.

[23] X. Zhang, L. Zhang, X.-J. Wang, H.-Y. Shum, Finding celebrities in billions of web images, IEEE Trans. Multimed. 14 (4) (2012) 995–1007.

[24] N. Kumar, A.C. Berg, P.N. Belhumeur, S.K. Nayar, Attribute and simile classifiers for face verification, in: IEEE International Conference on Computer Vision (ICCV), 2009, pp. 365–372.

[25] M. Kostinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, in: IEEE International Conference on Computer Vision Workshops (ICCVW), 2011, pp. 2144–2151.

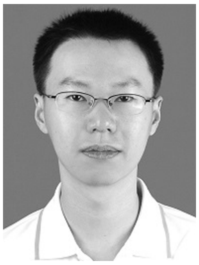
[26] X.W. Ziwei Liu, Ping Luo, X. Tang, Deep learning face attributes in the wild, in: IEEE International Conference on Computer Vision (ICCV), 2015.

[27] Y. Lecun, L. Bottou, Y. Bengio, P. Haffner, Gradient-based learning applied to document recognition, Proc. IEEE 86 (11) (1998) 2278–2324.

[28] A. Krizhevsky, I. Sutskever, G.E. Hinton, ImageNet classification with deep convolutional neural networks, in: Advances in Neural Information Processing Systems 25, 2012, pp. 1097–1105.

[29] B. Yang, J. Yan, Z. Lei, S.Z. Li, Convolutional channel features, in: IEEE International Conference on Computer Vision (ICCV), 2015.

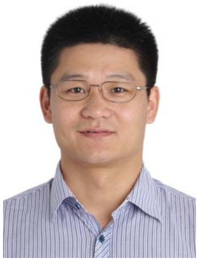
[30] S. Yang, P. Luo, C.C. Loy, X. Tang, WIDER FACE: a face detection benchmark, in: IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2016.



Shuzhe Wu is now a Ph.D. candidate at the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS). He received B.E. degree in computer science and technology from the Beijing University of Posts and Telecommunications, Beijing, China, in 2013. His research mainly focuses on face detection, object detection and deep learning.

Meina Kan is now an Associate Professor with the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), where she received the Ph.D. degree in computer science in 2013. Her research mainly focuses on face detection, face recognition, transfer learning and deep learning.

Zhenliang He received the B.E. degree from Beijing University of Posts and Telecommunications and is pursing the Ph.D. degree from Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China. His research interests include pattern recognition, machine learning and computer vision.



Shiguang Shan received M.S. degree in computer science from the Harbin Institute of Technology, Harbin, China, in 1999, and Ph.D. degree in computer science from the Institute of Computing Technology (ICT), Chinese Academy of Sciences (CAS), Beijing, China, in 2004. He joined ICT, CAS, in 2002 and has been a Professor since 2010. He is now the Deputy Director of the Key Lab of Intelligent Information Processing of CAS. His research interests cover computer vision, pattern recognition, and machine learning. He especially focuses on face recognition related research topics. He has published more than 200 papers in refereed journals and proceedings in the areas of Computer Vision and Pattern Recognition. He has served as Area Chair for many international conferences including ICCV'11, ICPR'12, ACCV'12, FG'13, ICPR'14, ICASSP'14, and ACCV'16. He is an Associate Editor of IEEE Trans. on Image Processing, Neurocomputing, and Pattern Recognition Letters.

Xilin Chen received the B.S., M.S., and Ph.D. degrees in computer science from the Harbin Institute of Technology, Harbin, China, in 1988, 1991, and 1994, respectively. He was a Professor with the Harbin Institute of Technology from 1999 to 2005. He was a Visiting Scholar with Carnegie Mellon University, Pittsburgh, PA, from 2001 to 2004. He has been a professor with the Institute of Computing Technology, Chinese Academy of Sciences (CAS), Beijing, since August 2004. He is now the vice-dean of this institute. He has published one book and over 150 papers in refereed journals and proceedings in the areas of computer vision, pattern recognition, image processing, and multimodal interfaces. Dr. Chen has served as a program committee member for more than 30 international and national conferences. He has received several awards, including China's State Scientific and Technological Progress Award in 2000, 2003, and 2005 for his research work. He is a senior member of the IEEE.