Image and Vision Computing xxx (2013) xxx–xxx

Contents lists available at ScienceDirect

Image and Vision Computing

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

Face detection by structural models ☆

Junjie Yan, Xuzong Zhang, Zhen Lei *, Stan Z. Li

Center for Biometrics and Security Research & National Laboratory of Pattern Recognition, Institute of Automation, Chinese Academy of Sciences, 95 Zhongguancun Donglu, Beijing 100190, China

ARTICLE INFO

Article history:
Received 8 June 2013
Received in revised form 13 September 2013
Accepted 5 December 2013
Available online xxxx

Keywords:
Face detection
Structural model
Face-body co-occurrence

ABSTRACT

Despite the successes in the last two decades, the state-of-the-art face detectors still have problems in dealing with images in the wild due to large appearance variations. Instead of leaving appearance variations directly to statistical learning algorithms, we propose a hierarchical part based structural model to explicitly capture them. The model enables part subtype option to handle local appearance variations such as closed and open month, and part deformation to capture the global appearance variations such as pose and expression. In detection, candidate window is fitted to the structural model to infer the part location and part subtype, and detection score is then computed based on the fitted configuration. In this way, the influence of appearance variation is reduced. Besides the face model, we exploit the co-occurrence between face and body, which helps to handle large variations, such as heavy occlusions, to further boost the face detection performance. We present a phrase based representation for body detection, and propose a structural context model to jointly encode the outputs of face detector and body detector. Benefit from the rich structural face and body information, as well as the discriminative structural learning algorithm, our method achieves state-of-the-art performance on FDDB, AFW and a self-annotated dataset, under wide comparisons with commercial and academic methods.

1. Introduction

Face detection plays an important role in face based image analysis and is one of the most important problems in computer vision. The performance of various face based applications, from traditional face recognition and verification to modern face clustering, tagging and retrieval, relies on accurate and efficient face detection. Successful frontal face detectors have been built in early years, such as [1–5]. Among these detectors, the Viola and Jones (V–J) detector [5] is the most popular one due to its advantage in efficiency. The V–J detector and its subsequences have achieved great successes. However, their performance is still not satisfactory in many real world scenes (e.g., FDDB [6]), due to the large appearance variations in pose, illumination, occlusion, expression and imaging condition.

There has been a lot of works on face detection. Successful face detectors benefit from statistical learning techniques such as SVM [3], Neutral Network [4], Bayesian [7], Boosting [5], and suitable feature representations such as Haar [5], LBP [8], and SURF [9]. Although be different in representation and learning, modern face detectors tend to follow a similar paradigm: distinguishing face and background by a "fixed" classifier. Here "fixed" means that no matter what the actual face configuration is, the same classifier is exploited. The "fixed" approach, however, results in the ambiguousness in practice, where the large appearance variations exist (e.g., the face organ layout for different individuals, the expression variations, and the heavy occlusion by sunglass or scarf).

Different from "fixed" face detectors, and motivated by recent successful deformable object detection methods, such as [10–13], we propose a structural model to capture configuration variations of face flexibly and explicitly. We define a hierarchical part based structure, which captures low frequency information at the coarse resolution level by a global template and high frequency information at fine resolution level by a set of part templates. To give meticulous description of the local appearance variation (e.g., closed eye and open eye), our model allows each part to have different subtypes. Moreover, parts in our model can have deformation in order to simulate global face variations caused by expression and pose. The detection process includes a fitting step, where firstly a candidate sliding window is fitted to the structure to find the suitable part location and part subtype, and then the detection score is calculated based on the fitted configuration. In this way, configuration variations can be handled explicitly. Due to the tree structure, the inference can be conducted efficiently. For discriminative parameter learning, we cast the problem into a structural SVM framework and show how to learn it practically.

According to our statistics on 8000 people from the real world images (see Fig. 2), there has been a strong co-occurrence between face and body. One question is that could the body information be helpful for face detection? We name this information as *face-body co-occurrence* and exploit it in the following three steps: (1) training suitable body detectors, (2) estimating the face localization by body detection, and (3) combining the activations generated by face detector and body detector. Considering the difficulty in capturing arbitrary body

☆ This paper has been recommended for acceptance by Lijun Yin, Ph.D.

* Corresponding author at: 1402, Intelligent Building, Zhongguancun Donglu, Beijing, China.
E-mail addresses: jjyan@nlpr.ia.ac.cn (J. Yan), xucong.zhang1990@gmail.com (X. Zhang), zlei@nlpr.ia.ac.cn (Z. Lei), szli@nlpr.ia.ac.cn (S.Z. Li).

0262-8856/$ – see front matter © 2013 Elsevier B.V. All rights reserved.
http://dx.doi.org/10.1016/j.imavis.2013.12.004

configuration, we propose a phrase based representation, where each phrase is a deformable part model to handle a special body configuration, such as "left orientated face on the shoulder" and "frontal face with upper body". Each activated phrase provides an estimation of face position by linear regression model defined on the part locations of the phrase. Since the body detectors and the face detectors may activate the same face, a merge procedure is needed. However, the traditional Non-maximal Suppression (NMS) procedure cannot be used in this case since the scores generated by different detections do not have equal confidences. In this paper, we propose a structural context model to encode a detection and its nearby detections to a linear feature, and learn the parameters by a structural SVM to determine whether the detection should be suppressed or not.

We conduct experiments on three challenging datasets, and achieve remarkable improvements over previous state-of-the-art methods. For example, our structural model improves the baseline [13] by 3% AP (average precision), and the face-body co-occurrence further improves 2% on the annotated faces from Pascal VOC. On AFW, the proposed method outperforms the baseline [13] by 8%, and outperforms the best academic method by 5%.

This paper is a substantial extension of our conference paper [14]. Compared with [14], we present further details of our method, and conduct more extensive experiments. We examine different experimental settings, and add experiments on AFW [13]. The rest of the paper is organized as follows. In Section 2, we review the related work. The structural face model, body and context model are presented in Sections 3 and 4. The experimental comparisons are discussed in Section 5. Finally in Section 6, we conclude the paper.

2. Related work

From the pioneering knowledge based methods [1,2] to modern learning based methods [3–5], numerous works were proposed to advance face detection. The learning based methods depend on special statistical learning algorithms, such as SVM [3], Neural Network [4] and Bayesian [7]. In [5], Viola and Jones proposed a method to combine integral image based Haar-like feature, adaboost based classifier and cascade based fast inference. Due to the advantage in speed compared with other good performance methods [7,3,4] at that time, it became very popular, and a lot of methods were proposed to further enhance it. The subsequences include new features (e.g. LBP [8] and SURF [9]), new weak classifiers (e.g. asymmetric classifier [15]), new boosting algorithms (e.g. gentle adaboost [16], multiple instance boost [17], float boost [18], KLboost [19], vector boost [20]), and new cascade structures (e.g. soft cascade [21], boosting chain [22]). Some papers aimed to achieve pose-invariant face detection, and proposed various cascade structures, as in [23,24,20]. Recently, [25] added an adaptive mechanism in calculating the confidence of weak classifiers. The more detailed face detection surveys can be found in [26–29].

Part based face representation has been explored in early years [30–32]. These methods have similar accuracy with V–J based methods. However, they are not popular at that time, mainly due to the high computation cost. These early models are still fixed, and ignore the flexibility in part based representation for face detection. Our structural face model follows a different framework with V–J detector, and is motivated by recent object detection and pose estimation systems, including [10–12]. Deformable part model (DPM) [10] is the basis of our model since our face model inherits hierarchical part based structure, spatial part deformation from it. Besides introducing the deformable part model to face detection area, we have three improvements over [10]: (1) we add subtype to each part, which is more flexible, (2) the parts are defined according to the landmark annotations instead of learning from ambiguous bounding box annotation, and (3) we use supervised structural learning algorithm instead of the Latent-SVM used in [10] to encode more supervision.

The most similar work on deformable face representation is [13], which exploited local parts around landmarks for joint face detection, landmark localization and pose estimation with promising performance. However, there are still problems in [13]. It did not consider the local variation of part and ignored the global structure information. The defined structure model in [13] is not robust to occlusion since the location of parent node may affect the location of its child node. In particular, the detection model in [13] can be seen as a special case of our face model by removing the hierarchical structure and part subtype option. Recently, [33,34] proposed methods to speed up part based face detection by cascade classifiers.

Face detection using body information has been discussed in several early works [35,36]. However, these works are limited to constrained setting, and how to use it for images in the wild is still unclear. Recent Pascal VOC person layout competition aims to predict the location of head, hands and feet given the location of the body [37]. Nevertheless, in real applications, the body location is not always known. Our phrase based body representation is motivated by the Poselet [38], but we add part information in the Poselet and get a phrase level representation to extract useful context information for face detection.

Compared with these prior works, there are three main contributions in our paper:

- We enrich the prior work on deformable part based face detection [13] by introducing hierarchical structure and part subtype.
- We present a phrase based model for body representation, and propose a structural context model to improve face detection by exploring the face-body co-occurrence.
- We achieve state-of-the-art performance on FDDB [6], AFW [13] and a self-annotated dataset compared with a lot of academic and commercial systems.

3. Structural face model

In this part, we first present the structural face model for flexible face representation, and then discuss the corresponding inference and learning algorithms.

3.1. Model definition

The lack of alignment for real world faces results in large appearance variation, which needs to be handled by the face model. To represent faces in arbitrary configurations without information loss, the ideal way is to model every pixel and the high order relationships between them. However, it is impracticable because the joint distribution is too complex to be learned by current machine learning techniques. Instead, a simple but flexible structure is defined to approximate the complex joint distribution. To provide flexibility in representing faces with complex variations, we conduct part based representation. We sequentially enrich the face model $\mathcal{M}$, to get the final hierarchical part based structure with part subtype option and part deformation.

3.1.1. Hierarchical part based structure

Useful information for face detection exists in different resolution levels. The global information is salient at low resolution level and more detailed face texture information can be found at high resolution level. We define a global root template to represent faces at low resolution level, and a set of part templates to represent faces at high resolution level, which is two times larger than the low resolution level. The parts are defined around the landmarks, such as eye corner and mouth center, as examples in Fig. 1(a). To capture the relationships between different levels, we also define spatial relationships between root and parts. In this way, we get the hierarchical part based structure, and factorize the face model $\mathcal{M}$ into three components:

$$\mathcal{M} \rightarrow \left\{ \mathcal{M}_r, \mathcal{M}_p, \mathcal{M}_s \right\}, \tag{1}$$

Please cite this article as: J. Yan, et al., Face detection by structural models, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.12.004

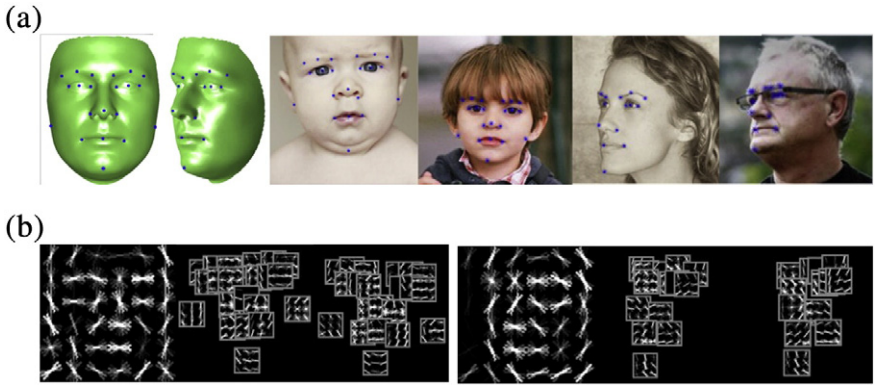


Fig. 1. Examples of structural face model.

where $\mathcal{M}_r, \mathcal{M}_p, \mathcal{M}_s$ are the global root face model at low resolution level, part set model at high resolution level and spatial relationship model, respectively. Since $\mathcal{M}_p$ consists of a series of parts, $\mathcal{M}_p$ can be further factorized to:

$$\mathcal{M}_p \rightarrow \left\{\mathcal{M}_{p_1}, \mathcal{M}_{p_2}, \cdots, \mathcal{M}_{p_N}\right\}, \tag{2}$$

where $\mathcal{M}_{p_i}$ is the i-th part model and N is the number of parts.

3.1.2. Part subtype

Although the part is local enough and tends to have less appearance variations compared with the full face, a single linear model is not enough to capture the local face appearance variations. For example, the nose type can range from "fleshy" to "celestial", the eye can range from open eye, closed eye and obscured eye (by sunglass or hair). To capture these local variations, we enable subtype option in our model, where the number of subtypes in each part is K. Denoting the j-th subtype model of the i-th part as $\mathcal{M}_{p_{i,j}}$, we get the following factorization:

$$\mathcal{M}_{p_i} \rightarrow \left\{\mathcal{M}_{p_{i,1}} | \mathcal{M}_{p_{i,2}} | \cdots | \mathcal{M}_{p_{i,K}}\right\}. \tag{3}$$

where "|" is the "OR" operation. Note that by introducing the subtype, our model can represent K^N different face configurations by KN part models, since each part can have K different choices. It equals to sharing KN parts in K^N different models.

3.1.3. Part deformation

There are two sources of global appearance variations for faces in the real world: (1) faces from different individuals have different face organs layout, such as the distance between two eyes; and (2) faces from the same individual can also have deformation, due to the pose and expression variations. To capture these appearance variations, we add a deformation model on each part. Although the deformation model of a single part can be very simple, the combination of deformations for all parts can simulate complex nonlinear face variation.

Given the definition of $\mathcal{M}_r$ and $\mathcal{M}_s$, a model is constructed to constrain the deformation of parts by modeling their spatial relationships with the root. Adding pairwise or higher order interactions between arbitrary parts can capture more structural information, but it will result in a loopy graph which is not efficient in inference. To keep the model to be tree-structured, we only define the spatial relationships between parts and root, and ignore pairwise relationships between different parts. We have N parts, each of which has K subtypes, and the corresponding factorization of $\mathcal{M}_s$ is:

$$\mathcal{M}_s \rightarrow \left\{\mathcal{M}_{s_1}, \mathcal{M}_{s_2}, \cdots, \mathcal{M}_{s_N}\right\} \tag{4}$$

$$\mathcal{M}_{s_i} \rightarrow \left\{\mathcal{M}_{s_{i,1}} | \mathcal{M}_{s_{i,2}} | \cdots | \mathcal{M}_{s_{i,K}}\right\}, \tag{5}$$

where $\mathcal{M}_{s_i}$ is the spatial model between i-th part and the root, and $\mathcal{M}_{s_{i,j}}$ is the j-th subtype of $\mathcal{M}_{s_i}$.

3.2. Configuration score

Given a configuration H in an image I, we need a function to measure whether the configuration H matches face model $\mathcal{M}$ or not. A face configuration H includes configurations of root h_0 and parts h_i. Here $h_i = \{l_i, t_i\}$, where l_i defines a location, which consists of upper-left corner (x_i, y_i), width w_i (we assume the face region is a square) and subtype label t_i. We compute the score following the structure of the model as:

$$\begin{aligned} &S(I, H, \mathcal{M}) \\ &= S_r(I, l_0, \mathcal{M}_r) + S_p\left(I, H, \mathcal{M}_p\right) + S_S(I, H, \mathcal{M}_s), \end{aligned} \tag{6}$$

where $S_r(I,l_0,\mathcal{M}_r)$, $S_P(I,H,\mathcal{M}_p)$, and $S_S(I,H,\mathcal{M}_s)$ are the match scores of the root, part and deformation. The $S_P(I,H,\mathcal{M}_p)$ and $S_S(I,H,\mathcal{M}_s)$ are further parsed into:

$$S_P\left(I, H, \mathcal{M}_p\right) = \sum_{i=1}^{N} S_p\left(I, l_i, \mathcal{M}_{p_{i,t_i}}\right) \tag{7}$$

$$S_S(I, H, \mathcal{M}_s) = \sum_{i=1}^{N} S_s\left(I, l_0, l_i, \mathcal{M}_{s_{i,t_i}}\right), \tag{8}$$

where S_p and S_s are the appearance and deformation score of each part, l_i is the location of the i-th part, and l_0 is the location of the root. Considering the efficiency in detection and learning, we assume all are of the linear form:

$$S_r(I, l_0, \mathcal{M}_r) = w_r^T \Phi_a(I, l_0) \tag{9}$$

$$S_p\left(I, l_i, \mathcal{M}_{p_{i,t_i}}\right) = w_{p_{i,t_i}}^T \Phi_a(I, l_i) \tag{10}$$

$$S_s\left(I, l_0, l_i, \mathcal{M}_{s_{i,t_i}}\right) = w_{s_{i,t_i}}^T \Phi_s\left(I, l_0, l_{i,t_i}, l_i\right), \tag{11}$$

where w_r, $w_{p_{i,t_i}}$ and $w_{s_{i,t_i}}$ are the model parameters of root appearance, part appearance and deformation. l_{i,t_i} is the anchor point of $\mathcal{M}_{s_{i,t_i}}$ relative to l_0. We use the appearance feature and spatial feature discussed in [10]. The appearance feature $\Phi_a(I, l_i)$ is 31 dimensional HOG features on location l_i of image I, which includes 9 dimensional contrast insensitive features,18 dimensional contrast sensitive features and 4 dimensional gradient energy features. The bin size in HOG is set to be 4. The deformation feature is defined as (dx, dy, dx^2, dy^2), where dx and dy are the relative deformation of each part to its anchor.

Please cite this article as: J. Yan, et al., Face detection by structural models, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.12.004

For the linear property, the total score of configuration H in image I can be simplified as:

$$S(I,H,\mathcal{M}) = w^T\Phi(I,H), \tag{12}$$

where w is the concatenation of all the parameters including w_r, $w_{p_{i,t_i}}$ and $w_{s_{i,t_i}}$. $\Phi(I,H)$ is the concatenation of all the features with the same order. For subtypes which are not activated, the corresponding dimensions in $\Phi(I,H)$ are filled with 0.

3.3. Detection and learning

3.3.1. Detection

In detection, standard sliding window procedure is used to scan images in different locations and scales to determine whether the special window corresponds to a face or not. For each scanning window in image I, we first fit the window to the structure model $\mathcal{M}$ to get the part configuration on it, and then calculate the score of the window according to the inferred configuration by Eqs. (6)–(11).

The fitting step aims to find the configuration H^* with the highest match score according to the learned model $\mathcal{M}$ on all configurations of a sliding window. Here we set the root location l_0 exactly be the location of sliding window. Mathematically, the optimization problem is to find H^* that satisfies:

$$H^{**} = \arg\max_H \left(S(I,l_0,\mathcal{M}_r) + \sum_{i=1}^{N} \left(S_p\left(I,l_i,\mathcal{M}_{p_{i,t_i}}\right) + S_s\left(I,l_0,l_i,\mathcal{M}_{s_{i,t_i}}\right) \right) \right). \tag{13}$$

The score of each part in the model is independent once the root is specified, so that we can maximize the following problem instead:

$$h_i^{**} = \arg\max_{h_i=\{l_i,t_i\}} \left(S_p\left(I,l_i,\mathcal{M}_{p_{i,t_i}}\right) + S_s\left(I,l_0,l_i,\mathcal{M}_{s_{i,t_i}}\right) \right), \tag{14}$$

where h_i is composed of part subtype label t_i and the part location l_i. In optimization, we first fix part subtype label t_i and find the best l_i, and then travel all subtypes to find the h_i^*. The complexity of maximizing one single sliding window is high, but benefit from the generalized distance transform proposed in [39], the average complexity in simultaneously optimizing all the sliding windows in one pyramid level is of linear complexity with the size of image. Actually, the optimization problem defined in Eq. (13) is very efficient compared with the HOG feature and appearance score computation. The confidence of the sliding window is computed by $S(I,H^*,\mathcal{M})$ in Eq. (6). By adding the fitting step before calculating, the latent face configuration is inferred, and the influence of configuration variation is reduced.

3.3.2. Learning

We use supervised learning instead of Latent-SVM which adopted in [10] to discriminatively learn parameters in the structural model. The positive training set in training is denoted as $\{I_a, H_a\}$, where H_a is the annotated face configuration in image I_a, including part subtype labels and part locations. The negative samples are denoted as $\{I_b, H_b\}$. We ensure that there is no face in I_b, so that any configuration H_b on I_b is not a face. In structural SVM, the optimal w should ensure that the score of positive samples $w^T\Phi(I_a, H_a)$ is bigger than 1, and the score of negative samples $w^T\Phi(I_b, H_b)$ is smaller than -1. Thus we get the following structural SVM problem:

$$\min_{w,\xi_i\geq 0} \quad \frac{1}{2}\|w\|^2 + C\sum \xi_n \tag{15}$$

$$\begin{aligned} s.t. \quad & \forall\{I_a,H_a\} \quad w^T\Phi(I_a,H_a) \geq 1-\xi_n \\ & \forall\{I_b,H_b\}, \quad w^T\Phi(I_b,H_b) \leq -1+\xi_n, \end{aligned} \tag{16}$$

where ξ is the penalty for violation. The problem is difficult because the number of negative configuration H_b is of combinatorial explosion. We use the coordinate descent solver, which iteratively: (1) solves w in dual, (2) mines violated constraints according to current learned w, and adds them to the constraint pool, until the convergence condition is satisfied.

In practical learning, we first learn the appearance parameters of root and all the parts independently, and then take them as initial values for structural SVM learning. Since there is no annotation of subtype, we use K-means to cluster annotated landmarks to K subtypes according to the positions in face. To cover the large poses, we train three models, including face yaw angles in $(-90°, -30°)$, $(-30°, 30°)$ and $(30°, 90°)$, respectively.

4. Structural body context model

It seems that no matter how powerful an appearance based face detector is, it would not always perform well for all scenes, due to the large appearance variations in novel faces. Besides the face itself, the most possible information that can contribute to face detection is the human body. As an intuitive example, a heavy occlusion on face can defeat the state-of-the-art face detectors, but we human beings can easily localize it with the help of un-occluded body.

To validate the co-occurrence between face and body in real world images, we analyze 8000 people annotations provided by [40] on Pascal VOC dataset. These images are sampled from Flickr, and we believe that the statistic can reflect the distributions of images in real world. Some of the annotation examples and the occurrence frequency of face and human body are shown in Fig. 2.

The interesting observation in Fig. 2 is that the shoulder of above 80% annotated people is available in the images and the statistic for hip is 50%. It proves the co-occurrence between face and other body parts on real world images. Here the question is how to effectively use the body context to enhance face detection. In the part, we first show our phrase based body detection and then present the structural context model to combine the information of face and body detection.

4.1. Phrase based body detection

Body detection itself is a challenging problem since the body can have large appearance variations, such as articulation of different parts. The best body (people) detector on Pascal VOC benchmark can only achieve about 50% AP (Average Precision) [37], which is far from satisfactory. Learning a better people detector would always be helpful in improving face detection, but it's beyond the scope of the paper. Here we only focus on how to get reliable body detection to provide useful information for face detection.

Deformable part model (DPM) [10] and some of its subsequences [11,41] can achieve state-of-the-art performance on people detection task. However, the part used in DPM has no clear correspondence to semantic part, which makes it not suitable to estimate face localization with body location. Another promising people detector is Poselet [40], which divides complex body configuration into simple local configurations and every Poselet just describes a single one. However, the Poselet defined in [40] can only provide the coarse body location template which is not rich enough to estimate the face location.

To provide richer body information, we conduct a phrase based representation. The phrase can be “left-orientated face on the shoulder” and “frontal half face with torso”. Although the detection of body under arbitrary configuration is a hard task, the detection of body with more constrained configuration is much easier and expected to have higher precision at a given detection rate. To generate phrase and its corresponding training samples, we use the technique described in [40] to find image patches similar to a seed patch according to annotated landmarks. The patches with similar configuration are used to train a DPM detector. In training DPM based phrase detector, we

Please cite this article as: J. Yan, et al., Face detection by structural models, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.12.004

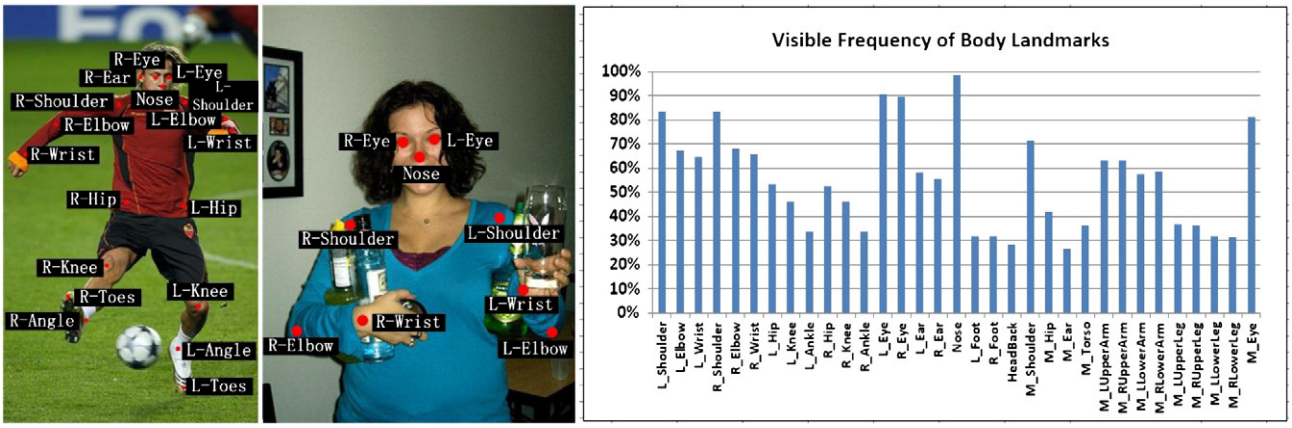


Fig. 2. Occurrence frequency of annotated human body landmarks on Pascal VOC. From this figure we can see that most of human faces are along with bodies in real world images.

modified the DPM code [42] to be fully supervised since all latent information in DPM are available given the body landmark annotation. Examples of trained phrase model are shown in Fig. 3(a) and (b).

Scanning an image using all the phrase independently is very time consuming. In this paper, we use the branch and bound implementation proposed in [43], which achieves logarithmic complexity in the image size. In our final implementation, we use 43 phrases and get a 10 times speed-up than original dynamic programming described in [10] with the same performance. Recent advances in detection such as [44] can handle 100,000 DPM detectors in 10 s on a single machine by hashing technique, which could be used to further speed up the phrase based detection.

4.2. Structural body context model

4.2.1. Predicting face location

Given the rich part locations provided by phrase based body detectors, we conduct linear model to predict the location of face. For each phrase based detector, we can get one root location and n part locations. We adopt a $2n + 3$ dimensional vector $v = (\rho_0, x_0, y_0, x_1, y_1, \cdots, x_n, y_n)^T$, where ρ_0 is the width of the root, and (x_0, y_0) is the upper-left corner coordinate of the root. (x_i, y_i) is the upper-left corner coordinate of the i-th part. The ρ_0 encodes the scale of the face, and the coordinates of root and parts encode the location information. We feed the vector v into a linear regression model to estimate the upper-left corner $\left(f_{x_l}, f_{y_u}\right)$ and lower-right corner $\left(f_{x_r}, f_{y_l}\right)$ of the corresponding face:

$$\left[f_{x_l}, f_{y_u}, f_{x_r}, f_{y_l}\right]^T = [w_1, w_2, w_3, w_4]^T v, \tag{17}$$

where w_i, $i \in \{1,2,3,4\}$ are the parameter vectors of linear regression model conducted on v independently. Since some of the parts may be not correlated with the face location, we use lasso instead of ridge regression, which can automatically select the correlated parts.

4.2.2. Structural context model

Since face detector and phrase based body detectors may activate the same face, a merge mechanism is needed to remove repetitive activations in combining face detection and body detection. Widely used merge method such as non-maxima suppression (NMS) cannot be directly used here, because the scores in different models do not have equal confidence. [45] proposed a learning based merge method for multiple classes object detection and reported advantages over NMS. However, the pairwise relationship results in a loopy graph in inference, which harms the convergence of the learning algorithm as pointed in [46].

Motivated by [45,47], we present a learning based context model to merge detection results of face detection and body detection. Given an image I, we use the face detector and phrase based body detectors to

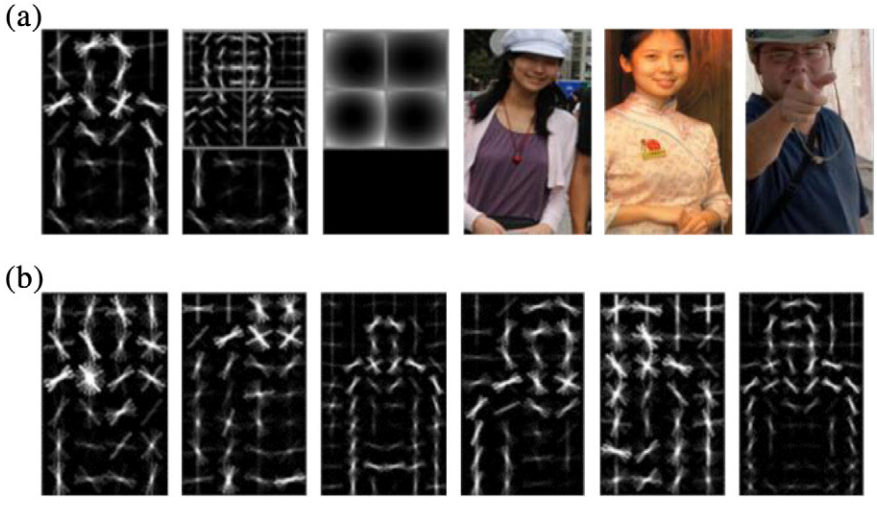


Fig. 3. Examples of phrased body detectors. Each phrase is modeled by a deformable part model to capture a special configuration of human body.

Please cite this article as: J. Yan, et al., Face detection by structural models, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.12.004

get initial face candidates $B = \{b_1, b_2,..., b_D\}$, where $b_i = \{l_i, s_i, c_i\}$ includes the location l_i, detection score s_i and the corresponding detector ID c_i. Here the detector ID identifies which detector the candidate comes from, which can be either face detector or 43 phrase based body detectors. The merge procedure is to assign a binary label set $T = \{t_1, t_2,..., t_D\}$ to B, where $t_i = 1$ means that b_i corresponds a face and should be kept. Otherwise, b_i is not a face or the face in b_i has been activated by nearby candidate, which should be suppressed.

The basic idea of the context model is to use the candidates nearby to help decide whether the candidate should be suppressed or not. To use information of nearby activations, we encode the nearby overlapped activations into feature representation of b_i, denoted as x_i. x_i is a $(3M + 2)$ dimensional vector, where M is the number of detectors (including both face and body detectors). The first two dimensions of x_i are the score s_i of the candidate b_i and a bias term 1. The following $3M$ dimensions of x_i are defined as follows: for each $b_j, j \neq i$, if there is an overlap between b_j with b_i, we set the $3c_j$, $3c_j + 1$ and $3c_j + 2$ dimensions to be the overlap ratio, the scale ratio between b_i and b_j, and the detection score s_j of b_j, respectively. If some dimensions are not activated by any b_j, they are filled with 0. For overlapped activations with the same detector ID, we just keep the one with the largest overlap ratio. In this way, the context information of nearby candidates are encoded into a feature vector which is ready for a learning algorithm, instead of the heuristic NMS. The demonstration of the feature component is shown in Fig. 4.

We build a linear model parameterized by w_{c_i} on feature x_i to infer the label t_i of b_i, where t_i is set to be 1 if $w_{c_i}^T x_i > 0$ otherwise 0. It equals to $max_{t_i=0,1} w_{c_i}^T \Phi(b_i, t_i)$, where $\Phi(b_i, t_i) = x_i$ if $t_i = 1$ and $\Phi(b_i, t_i) = -x_i$ if $t_i = 0$. Given an image, we assign labels to each activations b_i independently, and then whole label set T of the image I can be inferred by $max_T w_c^T \Phi(B, T) = \sum max_{t_i} w_{c_i}^T \Phi(b_i, t_i)$. Here to simplify the notation, we concatenate w_{c_i} to be a long vector w_c, and concatenate $\Phi(b_i, t_i)$ to be a long vector $\Phi(B, H)$ with the same order. The w_c combines the parameters for all the detectors, and the $\Phi(B, H)$ encodes the contextual information for the whole image. The ideal w_c should ensure the true hypothesis H of the image has a bigger score $w_c^T \Phi(B,H)$ than any wrong hypothesis T with a margin $L(T, H)$. w_c can be discriminatively learned from labeled training image set $\{I_n\}$ by the following structural SVM problem:

$$\min_{w_c, \xi_n \geq 0} \quad \frac{1}{2}\|w_c\|^2 + C\sum_n \xi_n \tag{18}$$

$$s.t. \quad \forall n, T_n \qquad w_c^T \Phi(B_n, H_n) - w_c^T \Phi(B_n, T_n) \geq L(T_n, H_n) - \xi_n, \tag{19}$$

where T_n is arbitrary hypothesis of B_n in I_n. The difference between T_n and H_n is measured by the $L(T_n, H_n)$, which is defined as a Hamming loss.

Since the number of hypothesis T_n is of exponential order, the constraints in above optimization problem cannot be stored into memory once in practice. We further use the coordinate descent procedure in our optimization. At each loop, we use the current model to scan images and find the hardest wrong hypotheses, and then use hard wrong hypotheses as the negative samples to update the model.

5. Experiment

In this part, we first describe the training data used in learning structural face model and body context model. After that, we examine the affect of different parameters on face detection performance on a self-annotated dataset. Finally, we compare our method with various commercial and academic methods on FDDB [6] and AFW [13].

5.1. Training datasets

The positive samples used in training structural face model come from AFLW database [48], which is a large scale face dataset collected from Flickr, with maximum 21 landmark annotations. Some of the faces in the dataset are not fully annotated, and we just ignore these faces for simplicity. In our experiments, we use 3065 faces for training frontal face model with 21 landmark annotations and 1552 faces for training profile face model with 14 landmark annotations. Our phrase based body detection models are trained on the training and validation set of Pascal VOC 2009 detection [37] and H3D dataset [38]. We use 4087 images with 8566 people in Pascal detection dataset, and we use the 33 landmark annotations from [49]. In H3D dataset, there are 2000 people annotations on 520 images. For both the face and body model learning, the negative samples are all from non-people images in Pascal VOC dataset [37]. To train the context model between face and body, we select 1000 images from Pascal detection test set and annotate faces on them by ourselves.

5.2. Parameter selection

To investigate the influence of different parameter settings on real world images, we annotate a face detection benchmark collected from the test set of Pascal person layout dataset, which is a subset from Pascal VOC [37]. It contains 1335 faces from 851 images with large appearance variations. A detection is taken as correct if the area of overlap ratio (the intersection of two regions against the union of the two regions) is greater than 50%, and the performance is reported by Recall–Precision curve used in [37]. To give a summary of the Recall–Precision curve, we use the average precision (AP) metric, which is defined as the area under the Recall–Precision curve. Here we explore each parameter sequentially while keep others constant.

5.2.1. Structure

The proposed structure face model has some components, such as the root, parts and deformations. To evaluate the contribution of each aspect of the structure model, we test the following structure settings: (1) root, which has only the root and can be taken as a standard multi-view HOG + SVM detector; (2) parts, where only fixed parts are used (similar to the structure used in [13]) and the parts are structured to be a tree, (3) root + parts, which is the tree structure defined in this paper that combines both the root and parts; and (4) root + part + deformation, which further adds the deformation

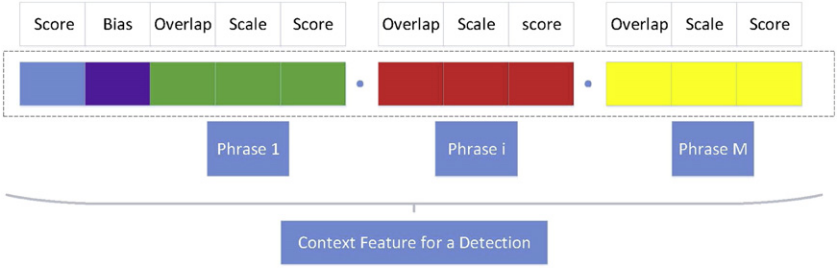


Fig. 4. Demonstration of context feature for a candidate. For each candidate, we encode the information of its nearby candidates to be a feature. The first two dimensions are the detection score and a bias term 1. The following dimensions encode the nearby candidates according to detector ID.

Please cite this article as: J. Yan, et al., Face detection by structural models, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.12.004

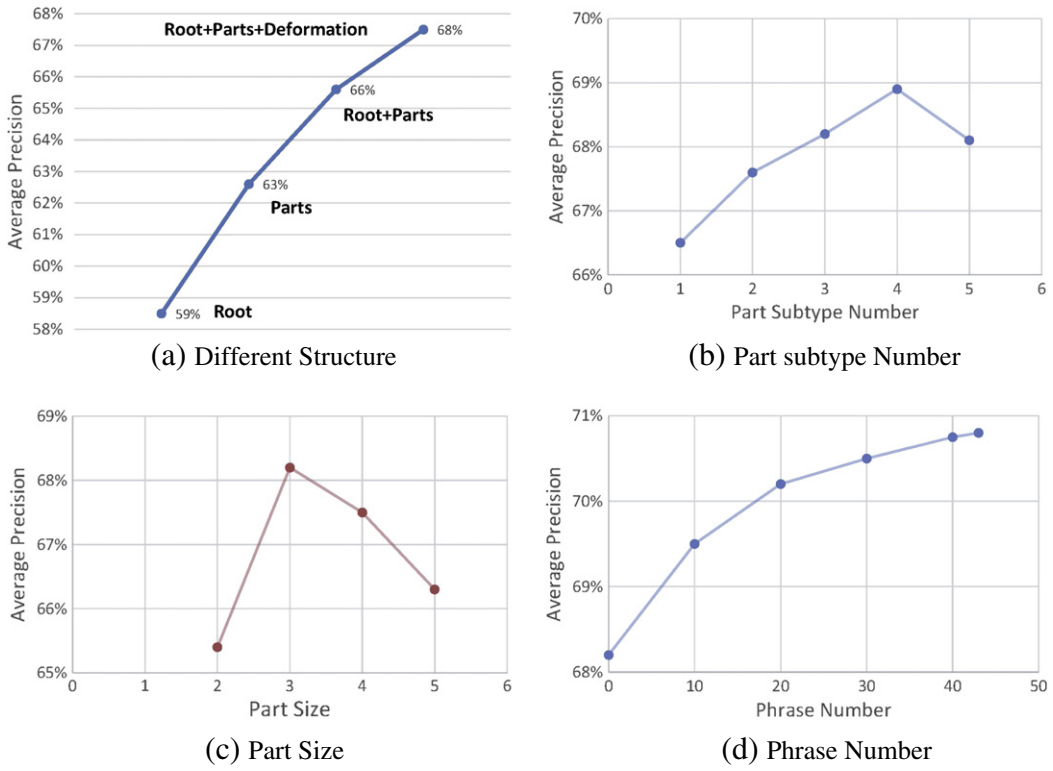


Fig. 5. Results of various parameter settings on face detection. These include (a) different structures for face detection, (b) the influence of part subtype number, (c) the influence of part size, and (d) the contribution of face-body co-occurrence.

of parts, but without part subtype option. Different structures are compared by AP and reported in Fig. 5(a). The root + parts outperforms root model by 7.1% and outperforms parts model by 3.0%. The deformation further improves 1.9%.

5.2.2. Part subtype number

Generally, more part subtypes are able to capture more local appearance variations but easier in overfitting and with higher computational cost. We test the face detection performance when the part subtype number K varies from 1 to 5, and the corresponding AP is demonstrated in Fig. 5(b). It can be seen that the best detection result is achieved when the subtype number K is set to 4, which improves the result of $K = 1$ as adopted in [10,13] by 2.4%.

5.2.3. Part size

Each part in the proposed structural model is set to be of equal size, which contains $n_d \times n_d$ HOG cells. Here we take n_d as the size of the part. Larger part could capture more information, but may lose the advantage in robustness of local part. We measure the AP when the part size varies from 2 to 5, and the best performance is achieved when the part size is set to be 3.

5.3. Phrase number

In addition, we examine the contribution of the face-body co-occurrence and the influence of phrase number, as shown in Fig. 5(d). Adding more phrases can have more body activations, thus can improve the performance. However, more phrases would take more inference and learning cost. In the following experiments, we fix the phrase number to be 43, which contributes to 2.6% AP as a trade-off between detection accuracy and computational efficiency.

5.4. Overall performance

In the following experiments (including experiments on FDDB and AFW), we use the root + parts + deformation structure, where the part subtype is set to be 4, the part size is set to be 3, and the phrase number is set to be 43. On the dataset, the proposed method is compared with other state-of-the-art methods including: (1) OpenCV implementation of 2-view Viola–Jones, (2) Kala's weighted sampling based boosting [50], including frontal and profile face detectors provided by the author, (3) TSM (tree structure model) [13], (4) the proposed structural face model, and (5) the proposed structural face model with body context AP curves are shown in Fig.6. The proposed method

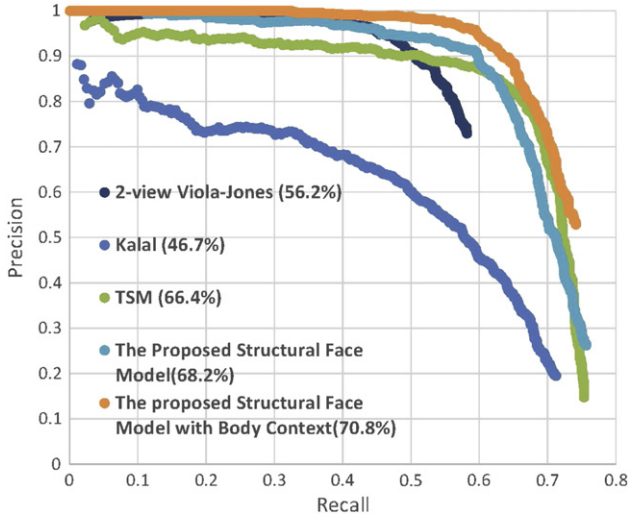
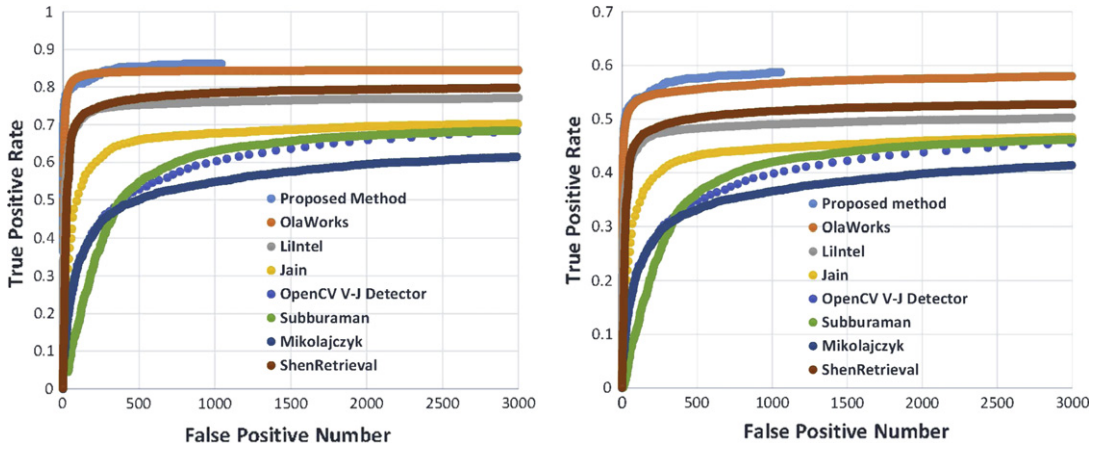


Fig. 6. Precision–Recall Curve and AP (average precision) on annotated faces from Pascal VOC.

Please cite this article as: J. Yan, et al., Face detection by structural models, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.12.004



(a) Discrete score ROC (b) Continuous score ROC

Fig. 7. Discrete and continuous ROC comparisons on FDDB. Our method and commercial Olaworks largely outperform other methods on this dataset.

outperforms TSM in [13] by 4.4% and outperforms the V–J detector in OpenCV by 14.6%. Some of the detection results are shown in Fig. 9.

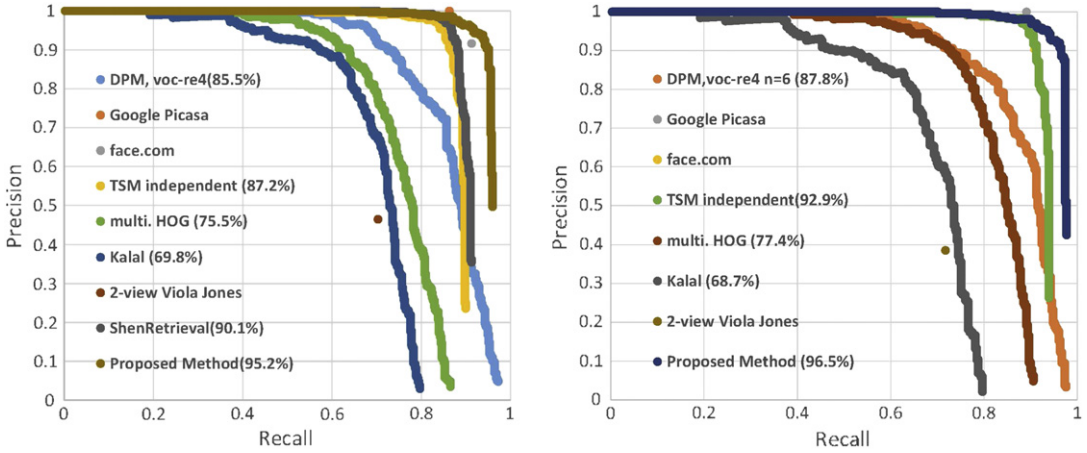
5.5. Experiment on FDDB

In this part, we use the FDDB [6] as a standard benchmark to compare the proposed method with other state-of-the-art methods. FDDB is a challenging face detection benchmark designed to study unconstrained face detection. It contains 5171 faces in 2845 images collected from news photographs. In the dataset, we compare the proposed method with the following methods: (1) Olaworks face detector, which is a commercial system that achieved the best performance on FDDB, (2) Li-intel SURF cascade detector [9], (3) Jain's face detector [51], (4) OpenCV implementation of 2-view Viola–Jones, (5) Subburaman's face detector [52], (6) Mikolajczyk's face detector [36], which used body information, and (7) Shen's face detector [53]. Following the protocol defined in [6], we report ROC curve under both the discrete and continuous metric, as shown in Fig. 7(a) and (b), respectively. The experiments are conducted on 10 subfolders, and finally the average ROC is reported. In continuous ROC, the overlap ratio is taken as the weight to measure matching quality. Some of the qualitative results on FDDB are shown in Fig. 9.

From Fig. 7, we can find that the proposed method and the commercial Olaworks are among the leading methods, which largely outperform others. The true positive rate of the proposed method is slight higher than Olaworks when the number of false positive is above 258 on discrete score ROC and 43 on continuous score ROC. The best academic method is [53], and the proposed method achieves 5% improvement on both discrete and continuous ROC over it. We also measure the average precision of different methods by cumulating the areas under the Recall–Precision curve. Our detector achieves 83.7% average precision, better than the 82.0% of Olaworks.

5.6. Experiments on AFW

We also examine our method on AFW [13]. It contains 205 images with 468 real world faces. Following the protocol in [13], we report performance on two Recall–Precision curves, the first is for all faces and the second is for large faces only (above 150 pixels in height), as in Fig. 8. The results of the following methods are also reported: (1) Open CV implementation of 2-view Viola–Jones, (2) Kala's weighted sampling based boosting [50], (3) Multi.HOG, which is a three view HOG + SVM implementation, (4) DPM [10], here the version 4 is used, (5) TSM [13], we use the best results reported in [13] to guarantee the best



(a) Precision-Recall Curve for all faces (b) Precision-Recall Curve for large faces only

Fig. 8. Precision–Recall curves for face detection on AFW test set: (a) all faces; and (b) faces larger than 150 × 150 pixels. Our method outperforms other academic method, and even outperforms the commercial method face.com on this dataset.

Please cite this article as: J. Yan, et al., Face detection by structural models, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.12.004

(a) Qualitative results on Pascal VOC.

(b) Qualitative results on FDDB.

(c) Qualitative results on AFW.

Fig. 9. Qualitative results on the three datasets. Images in the three datasets have cluttered backgrounds with large appearance variations caused by pose, glasses, expression, gender, aging and makeup.

performance, (6) face.com, and (7) Google Picasa. In the first Precision–Recall curve, we also report the result from a recent work [53].

The proposed method performs surprisingly well on this dataset. For "all faces" setting, the proposed method outperforms the baseline TSM by 8%, and outperforms the best academic method [53] by 5%. It even outperforms the commercial system face.com. Another commercial system Google Picasa has a higher precision at the same recall rate over the proposed method, but our method can recall more faces when slight more false positives are allowed. For "large faces" setting, our method achieves 96.5% average precision. On this setting, the proposed method outperforms baseline TSM by 3.6%. Some of the qualitative results on AFW are shown in Fig. 8. An interesting observation on the two curves is that the performance of traditional boosting based methods such as Viola–Jones and the Kalal's boosting does not increase with the resolution, while the HOG based methods DPM, TSM and our method can have remarkable improvements. A more detailed analysis of the phenomenon can be found in [54]. The qualitative results on Pascal, FDDB and AFW are shown in Fig. 9.

6. Conclusion

In this paper, we build structural models for face detection. The hierarchical part based representation provides flexibility in capturing appearance variations for faces in the wild. Furthermore, we propose a structural model to explore the face-body co-occurrence. We list the comparisons between proposed structural detector and the dominant V–J detector for face detection in the following three aspects:

- **Overall Accuracy** Benefit from the rich structural information captured in our method, we achieve state-of-the-art performance on challenging FDDB and AFW. The structural model can encode a lot of priors, such as the landmarks of face and body, which are not fully explored in traditional face detection models. Our work proved the advantages of complex model over the simple model (e.g. V–J) in face detection.
- **Occlusion Handling** Occlusion can be naturally handled in our method. Firstly, the hierarchical part based representation sums the activation of the whole face, thus be robust to local occlusion. Secondly, the phrase based body detector is not affected by occlusion in face regions.
- **Training Samples** Traditional V–J based methods need millions faces or more to get a desired performance, which makes data collection a hard problem. With proper modeling on the structure, the proposed method can achieve better performance by only thousands of training images.

Acknowledgement

We thank the reviewers and editors for helpful feedbacks. This work is supported by the Chinese National Natural Science Foundation Projects #61070146, #61105023, #61103156, #61105037, #61203267, and #61375037, the National IoT R&D Project #2150510, the National Science and Technology Support Program Project #2013BAK02B01, the Chinese Academy of Sciences Project No. KGZD-EW-102-2, the European Union FP7 Project #257289 (TABULA RASA), and Authen-Metric R&D Funds.

References

[1] T. Kanade, Picture Processing System by Computer Complex and Recognition of Human Faces, Department of Science, Kyoto University, 1973.
[2] C. Kotropoulos, I. Pitas, Rule-based face detection in frontal views, Proceedings of IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP-97), vol. 4, 1997, pp. 2537–2540, (IEEE).
[3] E. Osuna, R. Freund, F. Girosit, Training support vector machines: an application to face detection, Proceedings of the IEEE Computer Society Conference on Computer Vision and Pattern Recognition, 1997, pp. 130–136.
[4] H.A. Rowley, S. Baluja, T. Kanade, Neural network-based face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 20 (1) (1998) 23–38.
[5] P. Viola, M.J. Jones, Robust real-time face detection, International journal of computer vision 57 (2) (2004) 137–154.
[6] V. Jain, E. Learned-Miller, Fddb: A benchmark for face detection in unconstrained settings, University of Massachusetts, Amherst, Tech. Rep. UM-CS-2010-009, 2010.

[7] H. Schneiderman, T. Kanade, A statistical method for 3D object detection applied to faces and cars, Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition, vol. 1, 2000, pp. 746–751.
[8] L. Zhang, R. Chu, S. Xiang, S. Liao, S.Z. Li, Face detection based on multi-block LBP representation, Advances in BiometricsSpringer, 2007, pp. 11–18.
[9] J. Li, T. Wang, Y. Zhang, Face detection using surf cascade, IEEE International Conference on Computer Vision (ICCV Workshops), 2011, pp. 2183–2190.
[10] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, D. Ramanan, Object detection with discriminatively trained part-based models, IEEE Transactions on Pattern Analysis and Machine Intelligence 32 (9) (2010) 1627–1645.
[11] R.B. Girshick, P.F. Felzenszwalb, D.A. Mcallester, Object detection with grammar models, Advances in Neural Information Processing Systems, 2011, pp. 442–450.
[12] Y. Yang, D. Ramanan, Articulated pose estimation with flexible mixtures-of-parts, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2011, pp. 1385–1392.
[13] X. Zhu, D. Ramanan, Face detection, pose estimation, and landmark localization in the wild, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), 2012, pp. 2879–2886.
[14] J. Yan, X. Zhang, Z. Lei, D. Yi, S. Li, Structural models for face detection, IEEE International Conference on Automatic Face & Gesture Recognition (FG 2013), 2013.
[15] J. Wu, S.C. Brubaker, M.D. Mullin, J.M. Rehg, Fast asymmetric learning for cascade face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 30 (3) (2008) 369–382.
[16] R. Lienhart, A. Kuranov, V. Pisarevsky, Empirical analysis of detection cascades of boosted classifiers for rapid object detection, Pattern Recognition (2003) 297–304.
[17] P. Viola, J. Platt, C. Zhang, Multiple instance boosting for object detection, Advances in neural information processing systems 18 (2006) 1417.
[18] S. Li, Z. Zhang, Floatboost learning and statistical face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 26 (9) (2004) 1112–1123.
[19] C. Liu, H.-Y. Shum, Kullback–Leibler boosting, Proceedings of the EEE Computer Society Conference on Computer Vision and Pattern Recognition, vol. 1, 2003, pp. 1–587.
[20] C. Huang, H. Ai, Y. Li, S. Lao, High-performance rotation invariant multiview face detection, IEEE Transactions on Pattern Analysis and Machine Intelligence 29 (4) (2007) 671–686.
[21] L. Bourdev, J. Brandt, Robust object detection via soft cascade, IEEE Conference on Computer Vision and Pattern Recognition (CVPR), vol. 2, 2005, pp. 236–243.
[22] R. Xiao, L. Zhu, H. Zhang, Boosting chain learning for object detection, @Computer Vision, 2003. Proceedings. Ninth IEEE International Conference on, IEEE, 2003, pp. 709–715.
[23] S.Z. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, H. Shum, Statistical learning of multi-view face detection, Proceedings of the 7th European Conference on Computer Vision-Part IV, Springer-Verlag, 2002, pp. 67–81.
[24] M. Jones, P. Viola, Fast multi-view face detection, Mitsubishi Electric Research Lab TR-2003–96, 2003.
[25] K. Ali, F. Fleuret, D. Hasler, P. Fua, A real-time deformable detector, IEEE Transactions on Pattern Analysis and Machine Intelligence 34 (2) (2012) 225–239.
[26] M.-H. Yang, D.J. Kriegman, N. Ahuja, Detecting faces in images: a survey, IEEE Transactions on Pattern Analysis and Machine Intelligence 24 (1) (2002) 34–58.
[27] W. Zhao, R. Chellappa, P.J. Phillips, A. Rosenfeld, Face recognition: a literature survey, Acm Computing Surveys (CSUR) 35 (4) (2003) 399–458.
[28] C. Zhang, Z. Zhang, A survey of recent advances in face detection, Microsoft Research, June, 2010.
[29] S. Li, A. Jain, Handbook of Face Recognition, Springer, 2011.
[30] B. Moghaddam, A. Pentland, Probabilistic visual learning for object representation, IEEE Transactions on Pattern Analysis and Machine Intelligence 19 (7) (1997) 696–710.
[31] B. Heiselet, T. Serre, M. Pontil, T. Poggio, Component-based face detection, IEEE Computer Society Conference on Computer Vision and Pattern Recognition (CVPR 2001), vol. 1, 2001, pp. 1–657.
[32] H. Schneiderman, T. Kanade, Object detection using the statistics of parts, International Journal of Computer Vision 56 (3) (2004) 151–177.
[33] H. Cevikalp, B. Triggs, V. Franc, Face and landmark detection by using cascade of classifiers, Automatic Face & Gesture Recognition (FG 2013), 2013 IEEE International Conference on, IEEE, 2013.
[34] J. Yan, X. Zhang, Z. Lei, S.Z. Li, Real-time high performance deformable model for face detection in the wild, International Conference on Biometric, IEEE, 2013.
[35] H. Kruppa, B. Schiele, Using local context to improve face detection, Proc. of the BMVC, Norwich, England, 2003, pp. 3–12.
[36] K. Mikolajczyk, C. Schmid, A. Zisserman, Human detection based on a probabilistic assembly of robust part detectors, Computer Vision-ECCV 2004, Springer, 2004, pp. 69–82.
[37] M. Everingham, L. Van Gool, C.K.I. Williams, J. Winn, A. Zisserman, The PASCAL Visual Object Classes Challenge (VOC2012), Results, 2012, http://www.pascal-network.org/challenges/VOC/voc2012/workshop/index.html.
[38] L. Bourdev, J. Malik, Poselets: body part detectors trained using 3d human pose annotations, Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 1365–1372.
[39] P.F. Felzenszwalb, D.P. Huttenlocher, Pictorial structures for object recognition, International Journal of Computer Vision 61 (1) (2005) 55–79.
[40] L. Bourdev, S. Maji, T. Brox, J. Malik, Detecting people using mutually consistent Poselet activations, ECCV, 2010.
[41] J. Zhang, K. Huang, Y. Yu, T. Tan, Boosted local structured HOG-LBP for object localization, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1393–1400.
[42] P.F. Felzenszwalb, R.B. Girshick, D. McAllester, Discriminatively trained deformable part models, release 4, 2010.
[43] I. Kokkinos, Rapid deformable object detection using dual-tree branch-and-bound, Advances in Neural Information Processing Systems, 2011, pp. 2681–2689.
[44] T. Dean, M.A. Ruzon, M. Segal, J. Shlens, S. Vijayanarasimhan, J. Yagnik, Fast, accurate detection of 100,000 object classes on a single machine, IEEE Conference on Computer Vision and Pattern Recognition, 2013.
[45] C. Desai, D. Ramanan, C. Fowlkes, Discriminative models for multi-class object layout, Computer Vision, 2009 IEEE 12th International Conference on, IEEE, 2009, pp. 229–236.
[46] T. Finley, T. Joachims, Training structural SVMs when exact inference is intractable, Proceedings of the 25th International Conference on Machine Learning, ACM, 2008, pp. 304–311.
[47] M.A. Sadeghi, A. Farhadi, Recognition using visual phrases, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 1745–1752.
[48] M. Koestinger, P. Wohlhart, P.M. Roth, H. Bischof, Annotated facial landmarks in the wild: a large-scale, real-world database for facial landmark localization, Workshop on Benchmarking Facial Image Analysis Technologies, 2011.
[49] L. Bourdev, S. Maji, J. Malik, Detection, attribute classification and action recognition of people using Poselets (in submission), in: TPAMI, 2013.
[50] Z. Kalal, J. Matas, K. Mikolajczyk, Weighted sampling for large-scale boosting, Proceedings of the British Machine Vision Conference 2008, Leeds, 2008, pp. 1–10, (September 2008).
[51] V. Jain, E. Learned-Miller, Online domain adaptation of a pre-trained cascade of classifiers, Computer Vision and Pattern Recognition (CVPR), 2011 IEEE Conference on, IEEE, 2011, pp. 577–584.
[52] V.B. Subburaman, S. Marcel, Fast bounding box estimation based face detection, ECCV, Workshop on Face Detection: Where We Are, and What Next, vol. 7, 2010.
[53] X. Shen, Z. Lin, J. Brandt, W. Ying, Detecting and aligning faces by image retrieval, Computer Vision and Pattern Recognition (CVPR), 2013 IEEE Conference on, IEEE, 2013, pp. 4321–4328.
[54] J. Yan, X. Zhang, Z. Lei, S. Liao, S.Z. Li, Robust multi-resolution pedestrian detection in traffic scenes, IEEE Conference on Computer Vision and Pattern Recognition, IEEE, 2013.

Please cite this article as: J. Yan, et al., Face detection by structural models, Image Vis. Comput. (2013), http://dx.doi.org/10.1016/j.imavis.2013.12.004