Structural Models for Face Detection

Junjie Yan Xucong Zhang Zhen Lei Dong Yi Stan Z. Li
Center for Biometrics and Security Research & National Laboratory of Pattern Recognition
Institute of Automation, Chinese Academy of Sciences, China
{jjyan,xczhang,zlei,dyi,szli}@nlpr.ia.ac.cn

***Abstract*— Despite the success in the last two decades, the state-of-the-art face detectors still have problems in dealing with images in the wild for the large appearance variations. Instead of taking appearance variations as black boxes and leaving them to statistical learning algorithms, we propose a structural face model to explicitly represent them. Our hierarchical part based structural face model enables part subtype option to describe appearance variations of the local part, and part deformation to capture the deformable variations between different poses and expressions. In the process of detection, the input candidate is first fitted by the structural model to infer the part location and part subtype, and the confidence score is then computed based on the fitted configuration to reduce the influence of structure variation. Besides the face model, we utilize the co-occurrence of face and body to further boost the face detection performance. We present a method for training phrase based body detectors, and propose a structural context model to jointly use the results of face detector and various body detectors. Experiments on the challenging FDDB show that our method has state-of-the-art performance, compared with other commercial and academic systems.**

I. 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 to modern face clustering, tagging and retrieval, are relying on accurate face detection. Although the frontal face detection in controlled scene is reasonably solved, face detection in the wild is still a challenging problem, due to large appearance variations caused by pose, illumination, occlusion and expression.

There has been a lot of work on face detection. Successful face detectors often benefit from statistical learning techniques (e.g. Neutral Network [22], Bayesian [24], Boosting [26]). Although be different in feature representation, learning algorithm and classifier structure, they follow a basic framework, which feeds a "fixed" feature representation to a pre-trained classifier. Here the "fixed" means that no matter what the actual face configuration is, the fixed feature is extracted at fixed position. However, it would have problems in practice. For example, the relative positions of two eyes may vary for different individuals, which gives the ambiguousness. Moreover, even for faces from the same individual there often have various poses and expressions. The "fixed" detectors do not explicitly model these variations and leave them as black boxes.

Stan Z. Li is the corresponding author.

Different from these "fixed" face detectors, we propose a structural face model to describe face configuration variations flexibly and explicitly. We define a hierarchical part based structure, which captures low frequency information at the coarse resolution level by a global filter and high frequency information at fine resolution level by a set of part filters. To give meticulous description of the appearance variations (e.g., closed eye and open eye), our model allows the parts to have different subtypes. Moreover, parts in our model can be deformed to capture the face variations caused by expression and pose. Our detection procedure includes a fitting step, where the candidate region is first 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, the proposed model can handle configuration variations explicitly. Benefit from the tree structure, the model can be inferred efficiently. For discriminative parameters estimation, we cast the problem into a structural SVM learning framework and present the practical learning method.

It's observed that there has strong co-occurrence between the face and body for humans, especially for images in the real world. One question is can the body information be helpful for face detection? We name this information as body context and use it by training suitable body detectors and learning the relationship between face detector and body detectors. Considering the difficulty in body detection at arbitrary body configuration, we propose a phrase based representation, where every phrase is a deformable part model to describe a special configuration of body, such as "left orientated face on shoulder", "frontal face with upper body" and so on. Each activated phrase can give an estimation of face location. In order to merge repetitive activations for the same face by face detector and various body detectors, we further propose a learning based structural context model to encode the nearby activations and determine whether the activation should be suppressed or not.

A. Related Work

There are numerous papers on face detection or more general object detection. We mainly review the most related due to the space limitation. Among these early detectors (e.g. [20][22][24][26]), Viola-Jones(V-J) detector [26] is the most dominant one in the last decade for its efficiency, and there have been a lot of incremental improvements on it (e.g. [27][18][10]).

Our structural face model follows a different framework with V-J detector, and it is motivated by recent object detection and pose estimation systems, including [6][9][28]. Deformable part model (DPM) [6] is the basis of our model since our face model inherits the hierarchical part based structure, spatial part deformation from it. Besides introducing the deformable part model to face detection field, we have three improvements over [6]. Firstly, we add subtype to each part, which gives more flexibility. Secondly, the parts are defined according to the landmark annotations instead of learning from ambiguous weakly labeled data. Finally, in model learning phase, we use fully supervised structural learning algorithm instead of the Latent-SVM used in [6] to encode more supervision.

The most related work is [30], which used local parts around landmarks to represent face, and present a tree structure deformable model for joint face detection, landmark location and pose estimation with promising performance. However, it still has problems in the model. It did not consider the part variation problem and ignored the global structure information. The defined structure model in [30] was not robust to occlusion since the location of parent node may affect its children node. In particular, the model in [30] can be seen as a special case of our face model by removing the hierarchical structure and part subtype option.

Face detection using body information has been discussed in some early works [16][19], but how to use these information in unconstraint setting is still unclear. Recently Pascal VOC person layout competition aims to predict the location of head, hands and feet given the location of the body [5]. However, in real applications, the body location is always unknown. Our phrase based body representation is motivated by the Poselet [1], and we add part information in the Poselet to get a phrase level representation, in order to extract more useful context information for face detection.

The rest of the paper is organized as follows: Section 2 and Section 3 present the structural face model and structural body context model, respectively. The experiments on two challenging face databases are shown in Section 4, and finally in Section 5, we conclude the paper.

II. Structural Face Model

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

A. Model Definition

To represent faces in arbitrary configuration without information loss, the ideal way is to model all the pixels and high order relationship between them. However, it is impossible since that the joint distribution is too complex to be learned by current machine learning techniques. Instead, we define a structure on it, to simplify the complex joint distribution. In order to give more flexibility in representing faces with complex variations, we use part based representation, where the parts are defined around facial landmarks. Given an initial face model $\mathcal{M}$, we sequentially enrich it, to get the final hierarchical part based structure with part subtype option and part deformation.

Fig. 1. Example of training samples and annotations for frontal and profile views, respectively.

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. Here we use two resolution levels: a global root template for representing the faces at low resolution level and a set of parts to represent faces at the high resolution level, which is two times larger than the low resolution level. Similar to [30], the parts are defined around the landmarks, (e.g., eye corner, mouth center, as shown in Fig. 1). To build the relationship between the two levels, we also define a spatial constraint between location of root and part models. In this way, we get the hierarchical part based structure, and factorize the face model $\mathcal{M}$ into the following three components:

$$\mathcal{M} \longrightarrow \{\mathcal{M}_r, \mathcal{M}_p, \mathcal{M}_s\} \tag{1}$$

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

$$\mathcal{M}_p \longrightarrow \{\mathcal{M}_{p_1}, \mathcal{M}_{p_2} \cdots, \mathcal{M}_{p_N}\} \tag{2}$$

where $\mathcal{M}_{p_i}$ is the ith part model and N is the number of part.

2) Part Subtype: Although the part is local enough and tends to have less appearance variations compared with the full face, a single model is not enough to model the appearance variations of the part for the possible appearance variation, especially when only linear classifier is used. For example, the nose type can range from "fleshy" to "celestial". In order to capture these large variations, we enable the subtype option in our model. We set the number of possible subtypes in each part to be K. Denoting the jth subtype model of the ith part as $\mathcal{M}_{p_{i,j}}$, we get the following factorization:

$$\mathcal{M}_{p_i} \longrightarrow \{\mathcal{M}_{p_{i,1}} | \mathcal{M}_{p_{i,2}} | \cdots | \mathcal{M}_{p_{i,K}}\} \tag{3}$$

In this factorization, one and only one subtype can be activated at one time. Note that by introducing the subtype, our model can represent K^N different face configurations by KN part models, since every part can have K different choices. It equals to sharing KN parts in K^N different models. The part subtype mechanism gives more flexibility than global classifier subtype option widely used in pose-invariant face detectors.

3) Part Deformation: There are two sources of deformations in face images. The faces from different individuals have different face organs layout, (e.g. the width between two eyes). Faces from the same individual can also have deformation, due to the pose and expression variations. To capture the variation, we add a deformation model on every part. Note that although the deformation model of every part is very simple, the combination of deformations in all the parts could simulate complex nonlinear face variations. Given the definition of $\mathcal{M}_r$ and $\mathcal{M}_p$, a model is constructed to constrain the deformation of part by modeling the spatial relationship between the relative location in the two resolution levels. Adding pairwise or higher order interactions between different parts can capture more structural information, however, it will result in a loopy graph without efficient inference algorithm. To keep the model to be tree structured, we only define the spatial relationship between parts and root, and ignore pairwise relationship between different parts. Suppose we have N parts and every part has K subtypes, the factorization of $\mathcal{M}_s$ is:

$$\mathcal{M}_s \longrightarrow \{\mathcal{M}_{s_1}, \mathcal{M}_{s_2}, \cdots, \mathcal{M}_{s_N}\} \quad (4)$$

$$\mathcal{M}_{s_i} \longrightarrow \{\mathcal{M}_{s_{i,1}} | \mathcal{M}_{s_{i,2}} | \cdots | \mathcal{M}_{s_{i,K}}\} \quad (5)$$

where $\mathcal{M}_{s_i}$ is the spatial model of the i-th part, based on the relative location of $\mathcal{M}_{p_i}$ and $\mathcal{M}_r$, $\mathcal{M}_{s_{i,j}}$ is the j-th subtype of $\mathcal{M}_{s_i}$.

B. Configuration Score

Given a configuration H, we need a measurement function to calculate the score of its match to the learned face model $\mathcal{M}$. $H = \{h_0, h_1, h_2, \cdots, h_N\}$ consists of the configurations of root h_0 and parts h_i. Here $h_i = \{l_i, t_i\}$, where l_i is a location, which consists of upper-left corner (x_i, y_i), height h_i and width w_i; t_i is the subtype label. Following the structure of the defined model, the match score is defined as:

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

where $S_r(I, l_0, \mathcal{M}_r)$, $S_p(I, H, \mathcal{M}_p)$, $S_s(I, H, \mathcal{M}_s)$ are the match scores of the root level, part level, and spatial consistency, respectively. The $S_p(I, H, \mathcal{M}_p)$ and $S_s(I, H, \mathcal{M}_s)$ are further parsed into the appearance score and spatial score of each part:

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

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

For efficient detection and learning, we assume all the sub-scores are of the linear form:

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

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

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

where w_r^T, $w_{p_{i,t_i}}^T$ and $w_{s_{i,t_i}}^T$ are the model parameters. l_{i,t_i} is the anchor point of $\mathcal{M}_{s_{i,t_i}}$ relatively to l_0. We use the appearance feature and spatial feature discussed in [6]. The appearance feature $\Phi_a(I, l_i)$ are the modified 31 dimension HOG feature on location l_i of image I, and the cell size in HOG is set to be 4. The spatial feature is defined as a (dx, dy, dx^2, dy^2), where dx and dy are the relative deformation of l_i to its anchor l_{i,t_i}. Here $dx = x_i - x_0 - x_{i,t_i}$ and $dy = y_i - y_0 - y_{i,t_i}$.

Since all components in the model are of the linear form, the total score of configuration H in image I can be simplified as:

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

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

C. Detection and Learning

1) Detection: In the detection, we use the standard scanning window strategy to scan images in possible locations and scales to determine whether the special location corresponding to a face or not. For the scanning window located at L in image I, We fix the root model at L, and allow the deformation of every part. We first fit the candidate to the structure model $\mathcal{M}$ to get the best part configuration, and then calculate the score of the window according to the inferred configuration by Eq. [6-11].

The fitting is conducted by finding the configuration H^* with the highest match score according to the learned model $\mathcal{M}$ on all possible configurations:

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

Since the match score of every part is independent once the root filter is fixed, we can maximize the following problem instead:

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

There are two components in h_i, the part subtype label t_i and the part location l_i. In the optimization, we first fix part subtype label t_i and find the best l_i. By traversing all subtypes, we can find the h_i^* which maximizes the match score. The complexity of maximizing one single sliding window is high, but benefiting from the generalized distance transform proposed in [7], simultaneous optimization on all the sliding windows in the same pyramid level can be very efficient. The algorithm is of linear complexity with the size of image. For the detail of the algorithm, please refer to [7] for the lack of space here.

The confidence of the sliding window L is computed by $S(I, H^*, \mathcal{M})$ in Eq. 6. By adding the fitting step in advance, we can infer the latent face configuration and reduce the influence of configuration variation in face detection.

2) Learning: We use full supervised learning instead of Latent-SVM [6] to discriminatively learn the model parameters. The positive training samples are defined as $\{I_a, H_a\}$, where H_a is the annotated face configuration of image I_a, including part subtype labels and part locations. The negative samples is defined as $\{I_b\}$. We ensure that these is no face in I_b, so that any configuration H_b on I_b is negative. We use the concatenation representation in Eq. 12. The optimal w should ensure that the score of positive samples above 1, and the score of negative samples below -1. Thus we get the following structural SVM problem:

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

$$s.t. \quad \forall\{I_a, H_a\} \quad w^T\Phi(I_a, H_a) \geq 1 - \xi_n$$

$$\forall\{I_b\}, \forall H_b \quad w^T\Phi(I_b, H_b) \leq -1 + \xi_n \tag{16}$$

where ξ is the penalty for violation. The problem is difficult for the number of negative configuration H_b is combinatorial explosion. We use the dual coordinate descent solver from [28], which iteratively (1) solves w in dual; (2) mines violated constraints according to current learned w, and adds them to the constraint pool, until converged.

In practical learning, we first learn the appearance parameters of root model and all the part models independently, and then use them as the initialization for structural SVM training. We set the number of subtypes K to be 4 according to cross validation. Since there is no annotation of subtype, we use a K-means to cluster annotated landmarks to K subtypes according to the relative location in face. To capture the large pose variation, we train three models, including faces yaw angles in $(-90°, -30°)$, $(-30°, 30°)$, $(30°, 90°)$, respectively. For frontal faces, we use 21 landmarks and for profile faces, we use 14 landmarks.

III. 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, for the possible large appearance variation 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 make the state-of-the-art face detectors to fail, but we human beings can easily locate it with the help of un-occluded body.

A. Body Detection Model

Body detection itself is a more challenging problem since it often has large appearance variations, the best people detectors on Pascal VOC benchmark can only achieve about 50% AP (Average Precision) [5], which is far from satisfactory. Learning a better people detector would always be helpful in improving face detection, but it is beyond the scope of the paper. Here we focus on training suitable body model that can provide useful context information for face detection.

Mixture deformable part model (DPM) [6] and some of its improvements [9][29] can get state-of-the-art performance on people detection task, however, we find that the original DPM is not suitable for helping face detection for the ambiguity in latent variable inference. The part used in DPM has no clear corresponding to semantic part. Another promising people detector is Poselet [1], which divides complex people body configuration into simple local configurations and every Poselet just describes a single one. However, the Poselet defined in [1] is not part based, the only available global template is not accurate enough in estimate the face location.

To provide a better body context, we train a phrase model for body representation. The phrase can be "left-orientated face on the shoulder", "frontal half face with torso" and so on. Although detecting body under arbitrary configuration is a hard task, detecting body with more constrained configuration is simpler and is expected to have better precision. In order to generate phrase and corresponding training samples we use the technique described in [1] to find image patches similar to a seed patch according to annotated landmarks. We use patches with similar configuration to train a part based body detector. In training each phrase model, we modified the DPM code [8] to an fully supervised version since all latent information in DPM is available given the landmark annotation.

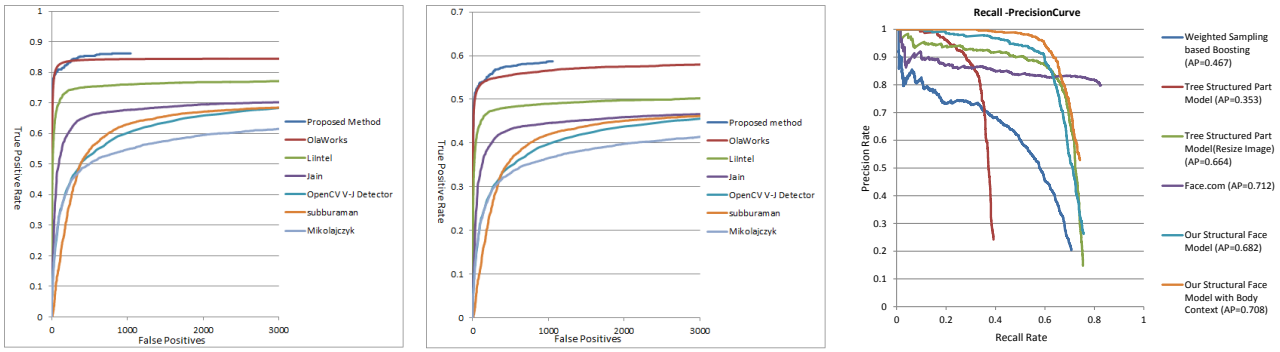
Each phrase model is a DPM model, and scanning an image using all DPMs is time consuming. In the paper, we use the branch and bound implementation as described in [15], which achieves logarithmic complexity in the image size. In our experiment, we use 43 phrase models, and get nearly 10 times speed up than original DPM inference code [6] with the same performance. We refer to [15] for the detail of the inference algorithm.

B. Structural Body Context Model

1) Predicting Face Location: Benefit from the rich part information in the phrase based body detectors, we can use simple models to predict the location of face. In this paper, we use a simple linear regression model. For each phrase model we have one root and n parts, and we construct a 2n+3 dimensional vector with the width of the root filter and the location of upper-left corner of each filter in image $v = (w_0, x_0, y_0, x_1, y_1, \cdots, x_n, y_n)$. We feed the vector v into a linear regression model to estimate upper-left corner (f_{x_l}, f_{y_u}) and lower-right corner (f_{x_l}, f_{y_r}) of face. In order to remove the effect of parts with little correlation, we use lasso instead of the ridge regression.

2) Structural Context Model: Since face detector and phrase based body detectors may activate the same face, we need a merge mechanism to remove the repetitive activations. Widely used merge method like non-maxima suppression (NMS) can not be directly used here, because the scores in different models do not have equal confidence. [4] gives a learning based merge method that outperforms NMS, where pairwise relationship are learned. However, it yields a loopy graph and only approximate inference algorithms are available.

Motivated by [4][23], 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



(a) Discrete score ROC (b) Continuous score ROC (c) ROC on Pascal Face

Fig. 2. ROC comparison of different algorithms on FDDB and Pascal Face.

phrase based body detectors to scan the image, and get the initial D face candidates $B = \{b_1, b_2, \cdots, b_D\}$, where $b_i = \{l_i, s_i, c_i\}$ includes the location l_i, detection score s_i and the corresponding model ID c_i. The merge procedure is to assign a binary label set $T = \{t_1, t_2, \cdots, t_D\}$ to B, where $t_i = 1$ means the b_i should be kept otherwise should be suppressed.

In order 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 number of models. The first two dimensions of x_i are the score s_i and a bias term 1. The other dimensions of x_i are defined as follows. For every $b_j, j \neq i$, if there is an overlap between b_j with b_i, we set the value in $3c_j$,$3c_j+1$ and $3c_j+2$ dimensions to be the overlap ratio, the scale ratio between them, and the confidence score respectively. Other unset dimensions are set to be 0. There may exist more than one overlapped activations with the same model ID, in this condition, we just keep the one with the largest overlap ratio.

We assume a linear model parameterized by w_{c_i} on the feature x_i to determine whether b_i should be kept or not. The label t_i of b_i is set to be 1 if $w_{c_i}^T x_i > 0$ otherwise 0. The procedure equals $\arg\max_{t_i=0,1} w_{c_i}^T \Phi(b_i, t_i)$, where $\Phi(b_i, t_i) = x_i$ if $t_i = 1$, and $\Phi(b_i, t_i) = -x_i$ if $t_i = 0$. For simplicity, we assign labels to b_i independently, then label set T is inferred by $\arg\max_T w_c^T \Phi(B, T) = \sum \arg\max_{t_i} w_{c_i}^T \Phi(b_i, t_i)$. We concatenate w_{c_i} to a long vector w_c. The optimal w should ensure that the score of true hypothesis H corresponding to a higher score $w_c^T \Phi(B, H)$ than any other hypothesis T with a margin $L(T, H)$ on the image. w_c can be discriminatively learned from labeled training image set $\{I_n, B_n, H_n\}$ by the following Structural SVM problem:

$$\arg \min_{w_c, \xi_x \geq 0} \quad \frac{1}{2}\|w_c\|^2 + C\sum_n \xi_n \tag{17}$$

$$s.t. \quad \forall H, n \quad w_c^T \Phi(B_n, H_n) - w_c^T \Phi(B_n, T_n) \geq L(T_n, H_n) - \xi_n \tag{18}$$

where T_n is arbitrary hypothesis of B_n in I_n, $L(T_n, H_n)$ is the Hamming loss to measure the difference between two detection hypotheses.

Since that the number of hypothesis T_n is of exponential order, the constraints in Eq. 17 can not be fit into memory once in practice. Instead we use cutting problem algorithm in our optimization. At every loop of the solver, we use the current model to scan images and find the most hard wrong hypothesis, and then use these hard wrong hypothesis as the negative samples to update the model.

IV. Experiment

In this part, we give the training data used in training structural face model and body context model. After that, we compare the learned model with various commercial and academic face detectors on challenging database.

A. Training Databases

The positive samples used in training structural face model come from AFLW database [14], which is a large scale face database that consists of 25,993 faces collected from Flickr, with maximum 21 landmark annotations. Some of the faces in this database are not fully annotated, currently we just ignore these faces for simplicity. In our experiments, we use 3065 faces for training frontal face model and 1552 faces for training profile face model. Our phrase models are trained on the train and validation set of the latest Pascal VOC 2012 detection database [5] and H3D database [3]. In Pascal detection database, there are 4087 images with 8566 people, we use the 33 landmark annotations from [2]. In H3D database, there are 2000 people annotation on 520 images. We combine these two databases to train our body model. For the two tasks, the negative samples are from non-people images in Pascal database [5]. In order to train the context model between face and body, we select the 1000 images from Pascal detection test set and annotate faces on them. Note that there is overlap between Pascal detection set and Pascal layout set, which will be used as test set in our experiment, so that we remove these overlapped images in training body detectors and body context model.

B. Experiment on FDDB

The FDDB database [11] is a challenging face detection benchmark designed for comparing unconstrained face detection. It contains 5171 faces in 2845 images taken from news

photographs. In the database, we compare our model with the following systems: (1) Olaworks face detector[1], which is a commercial system, and it achieved the best performance according to the FDDB result page[2]; (2) Li-intel face detector [17], which is based on SURF cascade; (3) Jain's face detector [12], which uses the image level context to improve boosting based algorithm; (4) OpenCV V-J face detector [3], which is one of the most popular open source face detector; (5) Subburaman's face detector [25], which improves V-J detector in sliding window phase; (6) Mikolajczyk's face detector [19], which used body information. ROC results under discrete score metric and continuous score metric are shown in Fig. 2(a) and Fig. 2(b), respectively.

We can find that our detector and the commercial Olaworks detector are among the leading methods on FDDB that outperform other detectors with large margin. Our true positive rate are slight higher than Olaworks when the number of false positive samples is above 258 on discrete score ROC and 43 on continuous score ROC. We also measure the Average Precision (AP) of different systems by cumulating the areas under the Recall-Precision curve, our detector achieves 0.837 AP, better than the 0.820 of Olaworks.

C. Experiment on Pascal Face

Besides FDDB, we also annotate a more challenging face detection database from the Pascal person layout database [5], which is collected from Flicker. There are 1335 faces from 851 images with arbitrary appearance variation. The state-of-the-art systems are compared, including the Face.com, Tree structure part model [30], V-J, and weighted sampling based boosting [13].

We use the Recall-Precision curves and corresponding AP (Average Precision) to measure the performance. Our system achieves comparable performance with the state-of-the-art commercial system Face.com, and outperforms other academic systems including [30], even only using the structural face model. Out structural face model outperforms [30] with 1.8% AP, which can prove the advantages of part subtype and hierarchical structure, since [30] can be seen as a special case of structural face model by removing the two information. On the experiment we can also find the value of body context model, which contributes 2.5% AP, by the comparison between the experimental setting (4) and (5) in Fig. 2(c).

V. Conclusion

In this paper, we propose structural models for face detection. The structural face model allows a fitting step in detection, thus has the flexibility to capture configuration variation for faces in the wild. Further, we present a method to capture the co-occurrence between face and body, and prove its advantage in face detection.

[1]http://eng.olaworks.com/olaworks/main/
[2]http://vis-www.cs.umass.edu/fddb/results.html
[3]http://sourceforge.net/projects/opencvlibrary/

VI. Acknowledgement

This work was supported by the Chinese National Natural Science Foundation Project #61070146, #61105023, #61103156, #61105037, #61203267, National IoT R&D Project #2150510, National Science and Technology Support Program Project #2013BAK02B01, Chinese Academy of Sciences Project No. KGZD-EW-102-2, European Union FP7 Project #257289 (TABULA RASA), and AuthenMetric R&D Funds.

References

[1] L. Bourdev, S. Maji, T. Brox, and J. Malik. Detecting people using mutually consistent poselet activations. *ECCV*, 2010.
[2] L. Bourdev, S. Maji, and J. Malik. Detection, attribute classification and action recognition of people using poselets (in submission). In *TPAMI*.
[3] L. Bourdev and J. Malik. Poselets: Body part detectors trained using 3d human pose annotations. In *ICCV*, 2009.
[4] C. Desai, D. Ramanan, and C. Fowlkes. Discriminative models for multi-class object layout. *IJCV*, 2011.
[5] M. Everingham, L. Van Gool, C. K. I. Williams, J. Winn, and A. Zisserman. The pascal voc 2012 results.
[6] P. Felzenszwalb, R. Girshick, D. McAllester, and D. Ramanan. Object detection with discriminatively trained part-based models. *TPAMI*, 2010.
[7] P. Felzenszwalb and D. Huttenlocher. Pictorial structures for object recognition. *IJCV*, 2005.
[8] P. F. Felzenszwalb, R. B. Girshick, and D. McAllester. Discriminatively trained deformable part models, release 4.
[9] R. B. Girshick, P. Felzenszwalb, and D. Mcallester. Object detection with grammar models. In *NIPS*. 2011.
[10] C. Huang, H. Ai, Y. Li, and S. Lao. High-performance rotation invariant multiview face detection. *TPAMI*, 2007.
[11] V. Jain and E. Learned-Miller. Fddb: A benchmark for face detection in unconstrained settings. Technical report, Technical Report UM-CS-2010-009, University of Massachusetts, Amherst, 2010.
[12] V. Jain and E. Learned-Miller. Online domain adaptation of a pre-trained cascade of classifiers. In *CVPR*. IEEE, 2011.
[13] Z. Kalal, J. Matas, and K. Mikolajczyk. Weighted sampling for large-scale boosting. *BMVC*, 2008.
[14] M. Koestinger, P. Wohlhart, P. M. Roth, and H. Bischof. Annotated facial landmarks in the wild: A large-scale, real-world database for facial landmark localization. In *Workshop on Benchmarking Facial Image Analysis Technologies*, 2011.
[15] I. Kokkinos. Rapid deformable object detection using dual-tree branch-and-bound. *NIPS*, 2011.
[16] H. Kruppa and B. Schiele. Using local context to improve face detection. *BMVC*, 2003.
[17] J. Li, T. Wang, and Y. Zhang. Face detection using surf cascade. In *ICCV Workshops*. IEEE, 2011.
[18] S. Li, L. Zhu, Z. Zhang, A. Blake, H. Zhang, and H. Shum. Statistical learning of multi-view face detection. *ECCV*, 2006.
[19] K. Mikolajczyk, C. Schmid, and A. Zisserman. Human detection based on a probabilistic assembly of robust part detectors. *ECCV*, 2004.
[20] E. Osuna, R. Freund, and F. Girosit. Training support vector machines: an application to face detection. In *CVPR*. IEEE, 1997.
[21] D. Ramanan. Dual coordinate descent solvers for large structured prediction problems. to appear.
[22] H. Rowley, S. Baluja, and T. Kanade. Neural network-based face detection. *TPAMI*, 1998.
[23] M. Sadeghi and A. Farhadi. Recognition using visual phrases. In *CVPR*, 2011.
[24] H. Schneiderman and T. Kanade. A statistical method for 3d object detection applied to faces and cars. In *CVPR*. IEEE, 2000.
[25] V. Subburaman and S. Marcel. Fast bounding box estimation based face detection. In *ECCV Workshop*, 2010.
[26] P. Viola and M. Jones. Robust real-time face detection. *IJCV*, 2004.
[27] J. Wu, S. Brubaker, M. Mullin, and J. Rehg. Fast asymmetric learning for cascade face detection. *TPAMI*, 2008.
[28] Y. Yang and D. Ramanan. Articulated pose estimation with flexible mixtures-of-parts. In *CVPR*. IEEE, 2011.
[29] J. Zhang, K. Huang, Y. Yu, and T. Tan. Boosted local structured hog-lbp for object localization. In *CVPR*.
[30] X. Zhu and D. Ramanan. Face detection, pose estimation, and landmark localization in the wild. In *CVPR*. IEEE, 2012.