Image and Vision Computing 32 (2014) 790–799

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 15 December 2013

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:

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