J. Vis. Commun. Image R. 56 (2018) 83–91

Contents lists available at ScienceDirect

J. Vis. Commun. Image R.

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

General-to-specific learning for facial attribute classification in the wild ☆

Yuechuan Sun, Jun Yu *

Department of Automation, University of Science and Technology of China, China

ARTICLE INFO

Article history:
Received 3 April 2018
Revised 15 August 2018
Accepted 2 September 2018
Available online 7 September 2018

Keywords:
Facial attribute
Deep convolutional network
Joint learning
Task-aware learning

ABSTRACT

Recent studies have shown that facial attributes provide useful cues for a number of applications such as face verification. However, accurate facial attribute interpretation is still a formidable challenge in real life due to large head poses, occlusion and illumination variations. In this work, we propose a general-to-specific deep convolutional network architecture for predicting multiple attributes from a single image in the wild. First, we model the interdependencies among all attributes by joint learning them all. Second, task-aware learning is adopted to explore the disparity regarding each attribute. Finally, an attribute-aware face cropping scheme is proposed to extract more discriminative features from where a certain attribute naturally shows up. The proposed learning strategy ensures both robustness and performance of our model. Extensive experiments on two challenging publicly available datasets demonstrate the effectiveness of our architecture and the superiority to state-of-the-art alternatives.

1. Introduction

Automatic facial attribute (e.g., gender, age, smile) classification, which aims at detecting the presence and absence of a certain facial attribute from a single image or video clip, is an actively researched problem in computer vision. Facial attributes are semantically meaningful to human as they give local facial representations and allow for higher-level descriptions of faces, people and activities. Facial attribute has proved useful in many real-life applications of surveillance [1], entertainment/human-computer interaction [2,3] and medical treatment [4,5]. Some methods treat attribute as prior knowledge for facial analysis tasks to facilitate final decision, such as face verification [6,7], face image retrieval/search [8,9] and face alignment/detection [10,11]. The success of these methods relies heavily on attribute detection results for giving auxiliary information. For instance, police can automatically search for persons of interest in surveillance videos on the basis of several attribute descriptions such as "Caucasian female with blond hair and eyeglasses", whereas traditionally users are required to watch the whole video streams in order to locate the target persons. Despite the wide application of facial attribute, however, analyzing facial attributes still remains challenging in real-world environments: arbitrary occlusions, non-frontal facing images, non-uniform illumination conditions and low image quality (see Fig. 1). Previous methods addressing these issues include point set matching [12], face frontalization [13], and using thermal face images [14]. However, [12] and [13] require the whole image area for reliable face modeling while an attribute is normally connected to a facial part and thermal images are usually hard to obtain. Hence, they are not successfully applied to facial attribute classification. Some attributes are particularly susceptible to the aforementioned challenges and are hard to recognize under difficult scenarios. For example, it is hardly possible to predict *lipstick* when the identity wears a mask as the lip is invisible.

Existing methods addressing the attribute classification problem can be generally categorized into global [7,15,16] and local [17–20] ones. Global methods usually extract features from the entire face and do not require localization of landmarks or object parts. They assume that different attributes are interdependent and each face part should be equally considered. All attributes are treated equivalently and no customized processing is conducted. On the contrary, local methods treat each attribute independently by first detecting face parts and applying feature descriptors to each part for training a classifier accordingly, where face alignment is critical to the final result. They may fail when accurate face localization and alignment are difficult to obtain due to occlusion under unconstrained conditions. However, local methods generally outperform the global ones when reliable

☆ This paper has been recommended for acceptance by Dr. Zicheng Liu.

* Corresponding author at: Department of Automation, University of Science and Technology of China, Room 301, Experimental Building, West Campus, Huangshan Road, Hefei, China.

E-mail addresses: ycsun@mail.ustc.edu.cn (Y. Sun), harryjun@ustc.edu.cn (J. Yu).

https://doi.org/10.1016/j.jvcir.2018.09.003
1047-3203/

Fig. 1. Exampler images from CelebA (top) and LFWA (below) with large head poses, intense occlusion, uneven illumination and low quality.

preprocessing is available as distinct spatial feature associated to each attribute is captured and less extra noise information is introduced. Considering the pros and cons of two methods mentioned before, our work integrates both methods by extracting facial representations from the holistic region of the aligned face and emphasizing on the functional parts in which a certain attribute naturally shows up using different face cropping schemes.

Over the past years, a variety of feature descriptors have been proposed for facial representation, such as local binary pattern (LBP) [21] and Gabor features [22], and they are improved and successfully applied to real-life face recognition [23–25]. Recently, Deep Convolutional Neural Networks (DCNNs) are widely used for handling most image recognition or classification tasks. This includes generic image classification [26,27], face verification [28–32], and pedestrian detection [33,34]. Thanks to the large-scale facial attribute dataset released by Liu et al. [15], the performance of facial attribute classification has also been greatly boosted by methods using DCNNs [7,15,35].

In this paper, we attempt to advance the state of the art in facial attribute classification under uncontrolled settings using DCNNs. Most prior methods treat attributes as independent from each other and each attribute is usually directly learned without intermediate stages. Nevertheless, we believe that facial attribute classification is not a standalone problem, but heavily influenced by other attributes. In addition, direct learning cannot fully exploit useful features, especially under challenging circumstance. To this end, we present a general-to-specific learning strategy composed of three major steps (see Fig. 2). In the first step, all 40 attributes are jointly learned to utilize underlying information provided by all attributes. In the second step, each attribute is individually learned to gain distinction information. In the last step, an attribute-aware cropping strategy is developed to refine attribute-related features and eliminate excessive information introduced by irrelevant face regions. By applying the general-to-specific learning strategy, we implicitly discover the correlation of all the attributes while specifically focus on the distinctions. The final model in Step 3 is obtained by step-to-step training and these three training stages are separately organized while highly correlated. Importantly, though this process is computationally expensive, it is applied only for training and merely Step 3 is required for a new testing image once the training is finished. The key contributions of this paper are:

1. Traditionally, all facial attributes are disconnected and treated equally [7,9], leading to information loss when some attribute are visually inaccessible under conditions like large occlusions and head poses. We overcome this limitation by introducing joint learning of all 40 attributes. Through joint learning, our model learns interdependencies of different attributes and thus yields higher robustness to challenging scenarios, which is unobtainable in single task learning.
2. Most DCNN-based methods directly take face images as input and produce output classification results for each attribute. To associate all attributes, we develop a general-to-specific learning framework that extracts both interconnections and disparities to improve facial attribute classification in the wild. By step-to-step learning, subtle features are thoroughly extracted without overfitting resulting from single attribute learning. Meanwhile, distinct information is captured in separate learning using task-aware face cropping and used to ensure exceptional performance.
3. We achieve the state-of-the-art average performance on two public benchmark datasets for facial attribute analysis, i.e., CelebA [15] and LFWA [36] without using external datasets. We also present the highest individual accuracies for most of the attributes on both datasets.

The rest of this paper is organized as follows: We first review related work in Section 2. The details of the proposed general-to-specific attribute prediction architecture are then elaborated in Section 3. Extensive experiments and results on two datasets are reported in Section 4. Section 5 finally concludes the paper with a brief discussion and future works.

2. Related work

2.1. Multi-task learning

Multi-task learning seeks to solve several problems at the same time by utilizing shared information when they are similar enough or are related in some sense [37,38]. Compared with independent single task learning, multi-task learning often leads to significantly improved performance in that simultaneous optimization with respect to multiple tasks enforces the algorithm to look for the