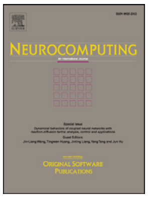
Neurocomputing 306 (2018) 246–259

Contents lists available at ScienceDirect

Neurocomputing

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



A complementary facial representation extracting method based on deep learning

Wenyun Sun[a,b], Haitao Zhao[c], Zhong Jin[a,b,*]

[a] School of Computer Science and Engineering, Nanjing University of Science and Technology, China
[b] Key Laboratory of Intelligent Perception and Systems for High-Dimensional Information of Ministry of Education, Nanjing University of Science and Technology, China
[c] School of Information Science and Engineering, East China University of Science and Technology, China

ARTICLE INFO

Article history:
Received 3 August 2017
Revised 27 April 2018
Accepted 29 April 2018
Available online 7 May 2018

Communicated by Junwei Han

Keywords:
Complementary facial representation
Facial expression
Deep learning

ABSTRACT

The identification and expression are two orthogonal properties of faces. But, few studies considered the two properties together. In this paper, the two properties are modeled in a unified framework. A pair of 18-layered Convolutional Deconvolutional Networks (Conv-Deconv) is proposed to learn a bidirectional mapping between the emotional expressions and the neutral expressions. One network extracts the complementary facial representations (i.e. identification representations and emotional representations) from emotional faces. The other network reconstructs the original faces from the extracted representations. Two networks are mutually inverse functions. Based on the framework, the networks are extended for various tasks, including face generation, face interpolation, facial expression recognition, and face verification. A new facial expression dataset called Large-scale Synthesized Facial Expression Dataset (LSFED) is presented. The dataset contains 105,000 emotional faces of 15,000 subjects synthesized by computer graphics program. Its distorted version (LSFED-D) is also presented to increase the difficulty and mimic real-world conditions. Good experiment results are obtained after evaluating our method on the synthesized clean LSFED dataset, the synthesized distorted LSFED-D dataset, and the real-world RaFD dataset.

1. Introduction

In the past several years, deep learning based face verification methods have achieved great success [1–3]. These methods try to learn good facial representations which maximize the between-class scatter and minimize the within-class scatter. The distances in the space of these representations are good facial similarity metrics for face verification. Another related research topic is the deep learning based Facial Expression Recognition (FER) [4,5]. Most FER methods are discriminative models that map the facial crops to emotional labels. The identification and expression are two orthogonal properties of faces. But, few studies considered the two properties together in a unified framework. There are two unsolved problems for such a challenging task. The one is what kinds of network should we used for modeling the multi-properties of faces? The other one is how to get a large-scale facial dataset labeled with detailed properties?

One of the trends in deep learning is using large-scale datasets. There are some large-scale facial datasets presented recently such as CASIA WebFace Database [6] and MegaFace [7]. An alternative way is to synthesize more images using Generative Adversarial Nets (GANs) [8–10]. The GANs are very promising methods for solving the data size limitation problem. Rendering images using computer graphics techniques is another solution [11,12] for building large-scale datasets. Based on the existing knowledge of computer graphics, rendered faces look real. Rendered facial expressions can strictly follow the definition of Facial Action Coding System (FACS) [13] and Emotional Facial Action Coding System (EMFACS) [14]. The synthesized datasets may be powerful tools for analyzing the real-world data.

The main contributions of this work include:

- A pair of 18-layered Convolutional Deconvolutional Networks (Conv-Deconv) is proposed to learn a bidirectional mapping between the emotional expressions and the neutral expression. From the view of facial representation, one network extracts the complementary facial representations from emotional faces. The other network reconstructs the original faces from the extracted representations. They are mutually inverse functions. The extracted complementary facial representations are used to reconstruct the original faces, generate new faces and

[*] Corresponding author.
E-mail addresses: zhongjin@njust.edu.cn, jinzhong@njust.edu.cn (H. Zhao).

https://doi.org/10.1016/j.neucom.2018.04.063
0925-2312/© 2018 Elsevier B.V. All rights reserved.

interpolate new faces. The facial representations can also be used for facial expression recognition and face verification.

- Based on computer graphics techniques, a new facial expression dataset called Large-scale Synthesized Facial Expression Dataset (LSFED) is presented. The dataset contains 105,000 emotional faces of 15,000 subjects. All the faces are synthesized using computer graphics techniques. To increase the difficulty and mimic real-world conditions, we create a distorted version of the LSFED, and name it as LSFED-D.
- Good experiment results are obtained after evaluating our method on the synthesized clean LSFED dataset, the synthesized distorted LSFED-D dataset, and the real-world RaFD dataset.

The rest of the paper is organized as follows. Section 2 reviews the related work. In Section 3, the main method is proposed. The experiments and results are presented in Section 4. Section 5 gives the conclusions.

2. Related work

2.1. Convolutional neural networks

The Convolutional Neural Networks (CNNs) is proposed by LeCun et al. [15]. Some large-scale modern CNNs [16–19] are proposed in the past several years. The recent CNNs based studies show significant improvement for varieties of computer vision tasks. For example, CNNs based bounding box regression is employed for automatically photo cropping and aesthetics assessment [20]. CNNs based representations are transferred between different tasks and data domains [21,22]. CNNs can also be used for video segmentation [23–25], remote sensing image scene classification[26], survival prediction[27], etc. In this work, CNN is used to learn a deep facial identification representation which is orthogonal to facial expressions.

2.2. Deconvolutional networks

The deconvolutional network is proposed by Zeiler et al. [28,29] for visualization of CNNs. In fact, the deconvolution network can be used to generate images [10,12]. The Conv-Deconv network consists of a pair of convolutional network and deconvolutional network. It can be used for some image to image tasks, such as saliency map generation [30–34], semantic segmentation map generation [29,35], optical-flow map calculation [36], and pose heat map estimation [37]. In this work, a pair of Conv-Deconv networks is designed to learn a bidirectional mapping between the emotional expressions and the neutral expression. To our knowledge, this is the first time to use Conv-Deconv networks to extract the complementary facial representations from emotional faces and to reconstruct the original faces from the extracted representations.

2.3. Extracting properties from images and generating images from properties

Kulkarni et al. [11] proposed the Deep Convolutional Inverse Graphics Network (DC-IGN) to learn an interpretable representation of faces. The representation is disentangled with respect to three-dimensional scene structure and viewing transformation including rotations and lighting variations. Dosovitskiy et al. [12] proposed a generative convolutional neural network to generate images from semantic labels, viewpoints, and colors. The two pieces of work use deconvolutional networks to generate objects from different viewpoints. In practice, collecting training data from different viewpoints is a hard work. An alternative way is rendering the training images using computer graphics techniques. In fact, using rendered datasets is a key idea in such kinds of work. Inspired by this idea, we use computer graphics techniques to synthesize a large-scale facial expression dataset. Based on this dataset, training our networks will be possible.

3. The proposed method

3.1. Extracting complementary facial representations

First of all, we are going to propose a pair of networks, namely NET-1 and NET-2. As illustrated in Figs. 1 and 2, both two networks are Convolutional Deconvolutional Networks (Conv-Deconv) [28,29]. The NET-1 is designed for splitting identifications and emotions. It accepts emotional faces as its input and calculates the corresponding neutral faces and the emotional predictions. The NET-2 is designed for merging identifications and emotions in the middle representation. It accepts the neutral faces and the emotional predictions as its input and generates the corresponding emotional faces. The NET-2 is the inverse function of the NET-1, and vise versa.

As illustrated in Fig. 3, the combination of the NET-1 and the NET-2 is referred as NET-3. The NET-3 extracts identification representations / emotional representations and then reconstructs the original faces. We use the neutral faces as the invariant anchors of all facial images belonging to a same identity, and use the emotional labels as the invariant anchors of all facial images belonging to a same emotion. The whole network disentangles complementary facial properties in deep representations.

Formally, denoting the convolutional/deconvolutional procedure of the NET-1 as f_1 / g_1, denoting the convolutional/deconvolutional procedure of the NET-2 as f_2 / g_2, and denoting the original emotional face and its corresponding normalized emotional label, unnormalized emotional label, neutral face as $\boldsymbol{X}$, $\boldsymbol{y} = (\boldsymbol{y}_1, \boldsymbol{y}_2, \ldots, \boldsymbol{y}_7)$, $\boldsymbol{y}_{logit} = (\boldsymbol{y}_{logit\,1}, \boldsymbol{y}_{logit\,2}, \ldots, \boldsymbol{y}_{logit\,7})$, $\boldsymbol{Z}$ respectively, the NET-1, the NET2 and the NET-3 can be defined as

$$(\boldsymbol{h}_1, \boldsymbol{y}_{logit}) = \boldsymbol{f}_1(\boldsymbol{X}, \boldsymbol{w}_{f_1}), \tag{1}$$

$$\boldsymbol{Z} = \boldsymbol{g}_1(\boldsymbol{h}_1, \boldsymbol{w}_{g_1}), \tag{2}$$

$$\boldsymbol{h}_2 = \boldsymbol{f}_2(\boldsymbol{Z}, \boldsymbol{w}_{f_2}), \tag{3}$$

$$\boldsymbol{X} = \boldsymbol{g}_2(\boldsymbol{h}_2, \boldsymbol{y}, \boldsymbol{w}_{g_2}), \tag{4}$$

$$\boldsymbol{y}_i = \frac{\boldsymbol{exp}(\boldsymbol{y}_{logit\,i})}{\sum_{j=1}^{7} \boldsymbol{exp}(\boldsymbol{y}_{logit\,j})}, i = 1, 2, \ldots, 7, \tag{5}$$

where $\boldsymbol{h}_1$ and $\boldsymbol{h}_2$ denote the middle activations of the two Conv-Deconv networks. $\boldsymbol{w}_{g_1}$, $\boldsymbol{w}_{f_2}$, $\boldsymbol{w}_{f_2}$, and $\boldsymbol{w}_{g_2}$ denote the parameter sets of the two Conv-Deconv networks.

($\boldsymbol{Z}$, $\boldsymbol{y}$) and ($\boldsymbol{h}_2$, $\boldsymbol{y}$) are two alternative pairs of complementary representations. Both $\boldsymbol{Z}$ and $\boldsymbol{h}_2$ are emotion-invariant and identification complete representations. In most case, $\boldsymbol{h}_2$ is preferred for the reason that $\boldsymbol{h}_2$ is a compressed representation. More empirical comparisons will be shown in the experiment section.

The framework above needs two Conv-Deconv networks (i.e. NET-1 and NET-2) to learn the bidirectional mapping between $\boldsymbol{X}$ and $\boldsymbol{Z}$. Two practical configurations of the Conv-Deconv networks are proposed. The configuration details including the layer types and the numbers of activations/parameters are listed in Tables 1 and 2. Each convolutional layer is followed by a Batch Normaliza-