Neurocomputing 147 (2015) 380–386

Contents lists available at ScienceDirect

Neurocomputing

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

Double layer multiple task learning for age estimation with insufficient training samples

Yao Ma [a,b], Jianyi Liu [a,*], Xi Yang [a], Yuehu Liu [a], Nanning Zheng [a]

[a] Institute of Artificial Intelligence and Robotics, Xi'an Jiaotong University, China
[b] Department of Computer Science, Tokyo Institute of Technology, Japan

ARTICLE INFO

Article history:
Received 11 December 2013
Received in revised form
25 April 2014
Accepted 17 June 2014
Communicated by Y. Fu
Available online 8 July 2014

Keywords:
Age estimation
Facial images
Multi-task learning
Insufficient sample problem
Support vector machine
Multiple kernel learning

ABSTRACT

One of the main difficulty of facial age estimation is the lack of training sample problem. In this paper, we point out that when age estimation is treated as a multiple task learning (MTL) problem, the impact of training sample problem can be relieved. By this idea, we re-formulate the age estimation task using the multi-class score function and develop a double layer multiple task learning (DLMTL) approach. In the subject layer, the personalized age estimation models as well as the global model are used to share knowledge of common aging pattern among different subjects; in the age label layer, the sub-tasks of score function estimation on any specific age label are further modeled to fully exploit the sequential information along the age axis. The proposed DLMTL model can be formulated into a very concise inner product representation, and it is finally solved using the multiple kernel learning (MKL) tool. The experimental results upon the FG-NET and MORPH aging databases verified that our method outperforms many other popular age estimation algorithms especially for the extremely training sample insufficient applications.

1. Introduction

Facial age estimation is a sub-problem in face recognition area [1,2]. Compared to all other facial attributes like illumination, pose, and expression, the main difficulty of age estimation is the insufficient training sample problem: one can easily collect as many as wanted number of facial photos with illumination/pose variations in laboratory, while it is a headache to collect enough number of age variation images for any specific subject. This makes age estimation the most challenging problem among various facial attribute recognition tasks.

The age estimation algorithms can be divided into two categories: classification based approaches [3–7] and regression based approaches [8–12]. The earliest report of age estimation algorithm that studies the training sample problem is the AGES proposed by Geng et al. [4]. They defined the sequence of personal face images sorted in time order as *aging pattern*. Due to lacking of training samples, many missing entries inevitably exist on each aging pattern. AGES uses a PCA reconstruction based algorithm to fullfill these highly incomplete aging patterns for training the classifier. In their another classification based approach [6], a novel label description degree is defined to represent the *label distribution* based on the fact that faces at close ages look similar, so that the samples of neighboring ages can be repeatedly exploited in the classifier to relieve the lacking sample problem. A similar idea named *label-sensitive* learning was proposed by Chao et al. [13] to alleviate the potential data imbalance problem. In their method, the label similarity was studied during the phase of manifold training to better exploit the ordinal relationship among age labels.

Another attempt coping with the insufficient sample problem is to use web image mining instead of the traditional sample collection way. Ni et al. [14] proposed a regression based age estimator under the multiple instance learning (MIL) framework. Their work also shows that the age label uncertainty problem has to be carefully addressed when using the huge web image resource. Liu et al. [11] proposed the hybrid constraint support vector regression (HC-SVR) algorithm that took the web collected samples to compensate the distribution bias of the aging database. The *fuzzy label* and *deterministic label* were defined for different data collection ways, and a regression based learning method was formulated to jointly exploit these different labels.

More recently, Chen and Hsu [15] found that *age ranks* play an important role in human cognition of age, and proposed a semi-supervised age regression method via learning the age ranks of unlabeled data, so that more available data can be exploited to

* Corresponding author.
E-mail address: jyliu@xjtu.edu.cn (J. Liu).

http://dx.doi.org/10.1016/j.neucom.2014.06.047
0925-2312/

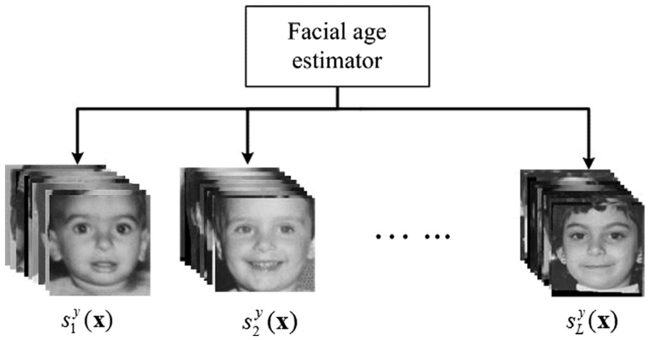

Fig. 1. Facial age estimation can be treated as a multiple task learning (MTL) problem. The estimations of score functions $s^y(\mathbf{x})$ for L individuals are different but related tasks.

improve the feature discriminability compared to traditional supervised learning methods. Weng et al. [16] also constructed an age ranking model named multi-feature ordinal ranking (MFOR). Different from most existing classification/regression based methods, they formulated age estimation as a group of *ordinal ranking* subproblems to better extract complementary information from different facial features.

In the pioneer work of age estimation, Lanitis et al. [8] already pointed out that the aging process of human face contains not only individual discrepancy but also some common trends such as skin relaxing and wrinkle increasing. The distributions of “appearance-age” among different subjects are similar to each other. In addition, the similarity information between adjacent/close age labels from the same subject is also useful [17,16,13]. Therefore, the tasks of estimating these distributions can be regarded as different but related tasks, as illustrated in Fig. 1. Due to the insufficient training sample problem, it is hard to estimate any individual distribution only by its own training data. However, it is possible to do it by jointly exploiting training data from all subjects and on all ages. This is a typical multi-task learning (MTL) problem statement [18].

In this paper, we will address the insufficient training sample problem of age estimation within a MTL framework. The knowledge transferring among tasks are accomplished in two layers: in the subject layer, the personalized age estimation models [19] as well as the global model are used to share knowledge of common aging pattern among different subjects; in the age label layer, the sub-tasks of estimating score functions on all age labels are further modeled to fully exploit the sequential information along the age axis. The proposed double layer MTL model can be formulated into a very concise inner product representation, and it is finally solved using the multiple kernel learning (MKL) tool [20].

The rest of this paper is organized as follows: we introduce our double layer multiple task learning model in Section 2, and formulate its solution method in Section 3, then describe the experimental results in Section 4 and conclude in Section 5.

2. Double layer MTL for age estimation

2.1. Problem statement

The human age estimation can be modeled as a multi-class classification problem, in which each class label is a specific age (in years old). A binary classifier $\mathrm{sgn}(S^y(\mathbf{x}))$ on each age y will tell whether or not the sample $\mathbf{x}$ belongs to the class y. Therefore, we need to find a set of score functions: $S^y(\mathbf{x}), y=1\ldots N$, where $\mathbf{x}\in\mathcal{R}^q$ is the q-dimension feature representation of image, $y\in\mathcal{R}$ is the class label (age label), and N is the total number of classes. The score function with the highest value will indicate the highest intensity of possible class belonging. Using the similar way as in [21], the age estimation problem can be stated as the following optimization problem:

$$\hat{y}=\max_{y\in\Omega} S^y(\mathbf{x}) \tag{1}$$

where $\Omega=\{1,2,\ldots,N\}$ is the whole set of class labels. However, the implementation of age estimation is always bothered by the lack of training sample problem. Let $\mathcal{T}_k=\{(\mathbf{x}_t^k,y_t^k)\}_t$ denote the labeled training data set coming from subject (person) $k, k=1\ldots L$, where $\mathbf{x}_t^k\in\mathcal{R}^q$ is the feature vector, $y_t^k\in\mathcal{R}$ is its label, L is the total number of subjects, and t is the index of each sample in the whole training set from all subjects. Due to the difficulty of age sample collection, the number of elements in each $\mathcal{T}_k$ is usually very limited, see Fig. 2 for instance. It is therefore impossible to estimate a satisfied score function for any of the k-th subject independently relying only the training data of $\mathcal{T}_k$.

As we all know, the distributions of “appearance-age” among various subjects share some common trends, although the individual differences also exist. This is the reason that human ourselves can easily distinguish age even from strangers. Therefore, the estimation tasks for all the L subjects can be regarded as related tasks, and we could estimate each individual distribution for $\mathcal{T}_k$ by jointly exploiting all labeled training data $\mathcal{T}=\mathcal{T}_1\cup\mathcal{T}_2\cup\cdots\cup\mathcal{T}_L=\{(\mathbf{x}_t,y_t)\}_t$. Therefore, age estimation can be modeled within a multi-task learning (MTL) framework.

2.2. The subject layer

Inspired by the idea above, we will estimate the score function $s_k^y(\mathbf{x})$ for each subject k by utilizing all the training data $\mathcal{T}$. The universal score function $S^y(\mathbf{x})$ is then represented as the linear combination of these $s_k^y(\mathbf{x})$, with the weight coefficients $\beta_k, k=1,\ldots,L$. Different tasks (subjects) will be able to share knowledge by means of the parameters β_k.

The score function $s_k^y(\mathbf{x}), y=1,\ldots,N$ can be regarded as the personalized model for the k-th subject. To reflect the common variation trend of aging pattern, we also use an additional score function $s_0^y(\cdot)$ here as the global model, which assumes that all samples yield the same distribution. The final score function can be defined as the linear combination of both the global model and the personalized models:

$$S^y(\mathbf{x})=\beta_0 s_0^y(\mathbf{x})+\sum_{k=1}^{L}\beta_k s_k^y(\mathbf{x}) \tag{2}$$

where $\beta_k, k=0,\ldots,L$ controls the weights of all models as in [17]. Different source tasks $s_k^y(\cdot)$ have different contributions to the target task $S^y(\cdot)$. The weights β_k represent the relatedness of all $L+1$ source tasks toward the target task. The higher related source task will have more influence to determine $S^y(\cdot)$. This representation has the advantage that both the common trend and the personalized differences of aging pattern are considered.

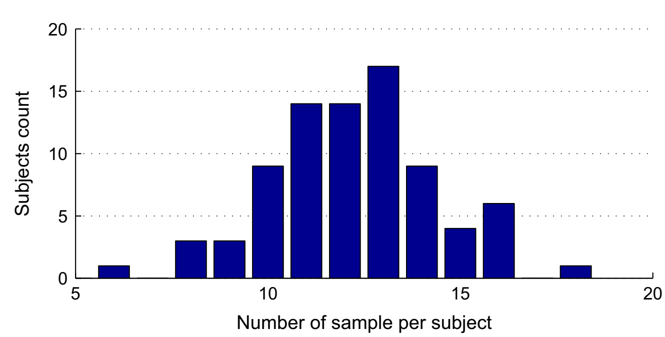

Fig. 2. Distribution of samples per subject in the FG-NET facial aging dataset.