Detecting Smiles of Young Children via Deep Transfer Learning

Yu Xia, Di Huang,* and Yunhong Wang
Beijing Advanced Innovation Center for Big Data and Brain Computing
Beihang University, Beijing 100191, China
{yuxia,dhuang,yhwang}@buaa.edu.cn

Abstract

Smile detection is an interesting topic in computer vision and has received increasing attention in recent years. However, the challenge caused by age variations has not been sufficiently focused on before. In this paper, we first highlight the impact of the discrepancy between infants and adults in a quantitative way on a newly collected database. We then formulate this issue as an unsupervised domain adaptation problem and present the solution of deep transfer learning, which applies the state of the art transfer learning methods, namely Deep Adaptation Networks (DAN) and Joint Adaptation Network (JAN), to two baseline deep models, i.e. AlexNet and ResNet. Thanks to DAN and JAN, the knowledge learned by deep models from adults can be transferred to infants, where very limited labeled data are available for training. Cross-dataset experiments are conducted and the results evidently demonstrate the effectiveness of the proposed approach to smile detection across such an age gap.

1. Introduction

Smile Detection has attracted extensive interests within the community in recent years due to its considerable applications such as smiling payment, entertainment, and mental status examination. However, variations in pose, illumination, and occlusion impose great difficulties to this task, making it more challenging in the real world. Early methods with hand-crafted features have shown promising performance on some small databases [10, 25, 28]. But those low level and mid level features cannot well capture Smile-related information implied in facial images, evidenced by significant drops in accuracy when test sets are largely expended with disturbance factors included. Recently, along with the rejuvenation of deep neural networks, the baselines of computer vision tasks, particularly face analysis, are greatly increased, and the robustness to lighting and viewpoint changes is substantially improved, for example, [7, 22] on this issue.

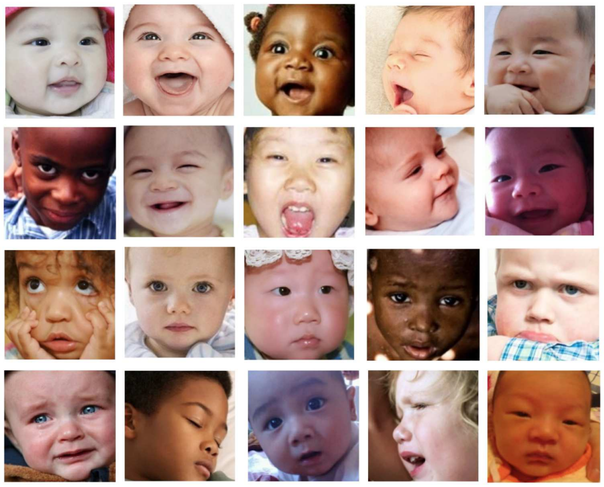

Figure 1. Examples of smile faces (top two rows) and non-smile faces (bottom two rows) on the Baby and Child Smile (BCS) dataset.

Despite of the progress achieved, there still exists a tough problem, namely age variations, which explicitly degrades the performance but has not been sufficiently studied before. To be specific, we find that the models trained on the current benchmarks do not perform well on faces of babies and young children[1] (as Figure 1 depicts). The reasons lie in twofold: (1) facial configuration of infants is different from that of adults due to immaturity of skull growth; and (2) sample distribution in public datasets is uneven, where images of kids are much less than the ones of adults. The latter has always been concealed in the previous papers by the average accuracy scores reported, since the number of such little guys for test is also quite small. This phenomenon tends to be more serious in deep learning based approaches, as these end-to-end solutions are data driven and require a big amount of diverse samples in model building or fine-tuning.

*indicates the corresponding author

[1] We use infant, kid, and child interchangeably in this study as the same concept.

While, it is indeed important to solve such a problem in smile detection, because it impedes further popularization of this technique in practical scenarios. For instance, the smile shutter in digital cameras is more useful to parents of infants for capturing decent photos, because young children are less tractable and display expressions more arbitrarily. Another example also appears in automatic monitoring systems for autism diagnosis, where smile detection contributes much and is required to apply to small kids in a non-intrusive way for auxiliary prediction.

Unfortunately, it is not easy to deal with such an age gap in smile detection. As we mention, the existent models cannot directly generalize for the structural difference between faces of infants and adults. Furthermore, compared with the labeled images of adults in the training set, the ones of babies and children are rather limited, and this disparity induces imbalanced results, especially for deep models. One may suggest to collect additional data from Internet, but images of infants are not as extensive as expected, in particular when other facial attributes are considered, *e.g.*, gender and race. Additionally, manually labeling those data greatly consumes manpower. Both make it unrealistic.

Considering the similarity between infant and adult faces, there is a strong incentive for leveraging transfer models to learn transferable features from off-the-shelf labeled data in a different but related source domain (*i.e.*, adult samples). This is a typical domain adaptation problem, which aims to establish knowledge transfer from the labeled source domain to the unlabeled target domain by exploring domain-invariant structures that bridge different domains of substantial distribution discrepancy [21], *e.g.*, the shift in datasets for an identical task as the case discussed in this study. On the other side, recent studies have illustrated that deep neural networks can learn more transferable features for domain adaptation [17, 20, 19, 33], which produce breakthrough results in image classification on standard benchmarks. It ultimately suggests the way to solve this problem.

Therefore, in this paper, we propose a novel and effective approach to smile detection on facial images of small kids with limited training data. Specifically, to quantitatively analyze the impact of this age gap, we first collect a special sample set, namely Baby and Child Smile (BCS), to test current models, and in contrast to available databases where images of young children are few, BCS contains 1,245 images, all of which are for kids. More importantly, the samples are balanced smile/non-smile and ethnicity (Caucasian, Asian, and Africa-American). The evaluation confirms that the age gap between infants and adults greatly challenges smile detection systems. We then formulate such an issue as an unsupervised domain adaptation problem, where the present dataset of full but uneven age groups is as the source domain and the newly built BCS of pure children as the target one. We introduce the state of the art deep transfer learning methods, namely Deep Adaptation Networks (DAN) [17] and Joint Adaptation Network (JAN) [19], applied to two baseline deep models, *i.e.*, AlexNet and ResNet. In the DAN architecture, hidden representations of all task-specific layers are embedded in a reproducing kernel Hilbert space where the mean embeddings of different domain distributions can be explicitly matched while the shift in marginal distributions is corrected across domains under the assumption that conditional distributions remain unchanged after marginal distribution adaptation. To make it more generalized, JAN aligns the shift in joint distributions of input features and output labels in multiple domain-specific layers across domains. They are both able to transfer the knowledge learned from adults by deep models to infants. Comprehensive Experiments are carried out in the cross-dataset scenario, and thanks to the advantage of deep transfer learning models, significant performance gains are reached compared to the state of the arts when detecting smiles on faces of babies and young children. It clearly indicates the competency of the proposed approach.

The contributions of this paper are summarized as follows: (1) to the best of our knowledge, it is the first time that aging is pointed out as a challenging factor to smile detection, supported by quantitative measurements; and (2) this problem is addressed in the viewpoint of transfer learning, and DAN and JAN based deep learning solutions are presented and competitive results are delivered.

The remainder of the paper is organized as follows. Section 2 briefly reviews the related work of smile detection and transfer learning. Section 3 describes the details of the proposed method, including the baseline deep networks as well as the deep transfer models. Experimental results are shown and analyzed in Section 4. Section 5 concludes the paper with perspectives.

2. Related Work

Our study involves in smile detection and transfer learning, and this section gives a brief review of related work in the two aspects.

2.1. Smile Detection

Early smile detection methods typically work in two steps: *i.e.*, hand-crafted feature extraction from face images and binary classifier design. Shan [25] used intensity differences between pixels in gray-scale face images as features and adopted AdaBoost for classification, and this method achieved 89.70% on the GENKI-4K database (1,000 images for test). Jain and Crowley [13] exploited Multi-scale Gaussian Derivatives (MGD) to extract facial features, and after combined with PCA based dimensionality reduction, they were fed into the SVM classifier for prediction, reporting the accuracy of 92.97% on the same database. Cui *et al*. [3] employed HOG features combined with RBF kernel based

SVM, and reached an accuracy of 93.25% on GENKI-4K. Smolka and Nurzynska [28] proposed Power LBP, a variant of the traditional LBP operator and applied the SVM classifier, which displayed the accuracy of 90.75% on the All database (712 images). Gao *et al.* [10] claimed that HOG31 with the cell-based feature map ($m \times m \times 31$) is the most optimized HOG feature, which led to the improved performance at 94.06% on GENKI-4K. They further jointly used a set of features (HOG31, GSS, and Raw Pixels) and multiple classifiers (AdaBoost and Linear ELM) and obtained state of the art performance of 94.61%.

Motivated by outstanding performance of Convolutional Neural Networks (CNNs) in a variety of computer vision tasks, several deep model based smile detection approaches have been proposed in recent years. They applied end-to-end deep CNN methods, which built high-level hierarchical features from raw data, and made estimation through a soft-max classifier. Zhang *et al.* [34] presented a 6-layer deep network and achieved a 94.6% classification accuracy on GENKI-4K. Glauner [11] extracted CNN features both from the entire face and the mouth region and integrated them, which displayed the accuracy of 99.45% on the DISFA database with dynamic sequences shot in a lab-controlled environment. Current state of the art approaches of smile detection mostly depend on the multi-task deep learning architecture which jointly dealt with multiple facial attribute classification with a single deep network. Zhang *et al.* [35] leveraged the VGG based multi-task deep network and adopted general-to-specific fine-tuning to both enhance the performance in gender classification and smile detection, which achieved the accuracy of 89.34% for the latter issue on the Faces of the World (FotW) [6] validation set (3,072 images). Ranjan *et al.* [23] introduced a method which established the multi-task learning framework based on AlexNet to simultaneously perform classification on seven facial attributes, and for smile detection they reported a state of the art performance up to 90.83% on the validation set of FotW.

Even if considerable performance gain has proved the effectiveness of those methods, they principally assume that the training and test sets are i.i.d., especially for the deep models that are required to be trained on a large amount of data. But this is not true in the given topic, since the learned smile features are probably different across individual age groups (*e.g.* infants vs. adults) where upper lip thickness decreases at rest and upper incisors are more exposed on smiling as the person become aged [4].

2.2. Transfer Learning

Transfer learning techniques aim to build models to jointly learn adaptive classifiers and transferable features from labeled data in the source domain and unlabeled data in the target domain. The early attempts were made on hand-crafted features [18, 26, 27]. In [18], Maximum Mean Discrepancy (MMD) was projected into the PCA subspace to represent domain discrepancy features. Si *et al.* [27] proposed to minimize the Bregman divergence between the distribution of source and target domains in the selected PCA, LDA, and LPP subspaces.

Regarding deep features, although they are expected to be more generalized due to its training on large-scale data, it has been point out that such hierarchical features can only reduce but do not remove dataset bias [31]. In fact, dataset shift is a major bottleneck to the transferability of deep features. Recent research [5, 9, 17, 20, 30] has extended CNNs to domain adaptation. For example, DAN [17] matched the shift in marginal distributions across domains by adding multiple adaptation layers through which the mean embeddings of distributions are matched, assuming conditional distributions remain unchanged. Ganin and Lempitsky [9] added a domain classifier connected to the feature extractor via a gradient reversal layer to align the distributions of features across the two domains. Despite significant improved performance, these methods are under the assumption that learned domain-invariant feature representations can be directly transferred from the source classifier to the target domain.

While this assumption does not hold when the source and target classifiers cannot be shared. More recently, the JAN network has been proposed to align the joint distributions of multiple domain-specific layers across domains based on a Joint Maximum Mean Discrepancy (JMMD) criterion without any assumption on the marginal distributions, which captures full interactions between different variables in the joint distributions, and can thus work in more general cases.

Transfer learning methods have been successfully applied to a number of face analysis tasks, such as face verification [1], expression recognition [2], pain intensity prediction [8], and age estimation [29]. In this paper, we investigate them in facial smile detection across the age gap between infants and adults.

3. Methods

Recall that we formulate smile detection across age as a transfer learning problem, and propose deep transfer learning solutions. In the subsequent, we introduce the framework of deep convolutional networks as well as the DAN- and JAN-based deep transfer learning methods.

3.1. Deep Convolutional Network

Given a labeled domain $\mathcal{D}_s$, we denote the set of parameters, weights $\mathbf{W}^\ell$ and bias $\mathbf{b}^\ell$ at the ℓth layer of CNN as $\Theta = \{\mathbf{W}^\ell, \mathbf{b}^\ell\}_{\ell=1}^{l}$, and the empirical risk is defined as

$$\underset{\Theta}{min} \frac{1}{n_s} \sum_{i=1}^{n_s} J(\theta(\mathbf{x}_i^s), \mathbf{y}_i^s) \tag{1}$$

where $J(\cdot,\cdot)$ is the cross-entropy loss function, and θ ($\mathbf{x}_i^s$) is the conditional probability that the CNN assigns label $\mathbf{y}_i^s$ to $\mathbf{x}_i^s$.

We select two CNN models, a basic one (AlexNet) [15] and an advanced one (ResNet) [12] as the baseline methods, since they are widely used in related work. To keep the integrity, we will first introduce the two deep models.

AlexNet consists of five convolutional layers ($conv1$-$conv5$), and three fully-connected layers ($fc6$-$fc8$). The output of the last fc layer is linked to a softmax layer which produces a distribution over the number of classes. Response-normalization layers follow the first and second convolutional layers. Max-pooling layers follow both response-normalization layers as well as the fifth convolutional layer. The ReLU non-linearity is applied to the output of every convolutional and fully-connected layer. This is the first model which makes breakthrough on the ImageNet dataset, by applying the techniques such as data augmentation and dropout. AlexNet illustrates the benefits of CNNs, it is still widely used for its clear and simple architecture, especially in multi-task learning [23] and transfer learning [17, 19, 20].

ResNet is a "ultra-deep" layer network architecture that sets latest records in classification, detection, and localization. Residual learning is formulated as input $\mathbf{x}$ plus its going through a residual function $\mathcal{F}(\mathbf{x})$ to form the result $\mathcal{H}(\mathbf{x}) = \mathbf{x} + \mathcal{F}(\mathbf{x})$, which benefits back propagation. Different from traditional CNN models, where the higher layer gradient must pass through the weight layer to reach the lower layer, during back propagation in ResNet, the gradient of higher layer can directly pass to the lower one which reduces the risk of vanishing gradient or exploding gradient. In our study, a 50-layer ResNet is considered as it is efficient in training. After 50 convolutional layers, the last layer is a fully-connected layer.

3.2. Deep Transfer Learning

Recent studies on transferring features in deep convolutional networks reveal that the transferability decreases and eventually changes from general to specific by the last layer of the network while convolutional layers can learn generic features that are transferable across domains [33], and when the cross-domain discrepancy increases, the transferability of features and classifiers both degrades. Meanwhile, the quantification study on deep transfer learning also suggests that the features can reduce the cross-domain distribution discrepancy, but cannot eliminate it [17, 19, 20, 33]. While the deep features at higher layers $\mathcal{L}$ are task-specific, the discrepancy between the training and test domain lingers in the activations $\mathbf{Z}^1$,..., $\mathbf{Z}^{|\mathcal{L}|}$ ($|\mathcal{L}|$ is the number of task-specific layers). For example, in the case of AlexNet, the activations in the higher fully-connected layers $\mathcal{L} = \{fc6, fc7, fc8\}$ are not safely transferable in a new domain. Usually, fine-tuning is required by the trained deep models for domain adaptation. However, target monitoring with limited data tends to make the fine-tuning work fail, and even if sufficient data are available, manual annotation on them is still annoying. An alternative is to apply unsupervised learning to unlabeled target data, making it fit for the paradigm of transfer learning. Therefore, it is necessary to develop transfer learning methods for deep learning models to solve this problem.

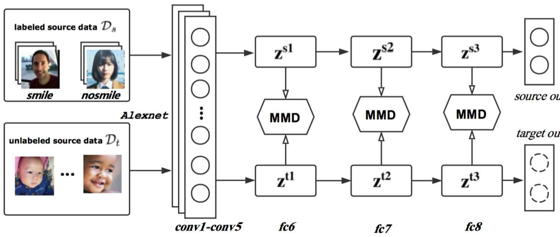

Figure 2. The DAN architecture for learning transferable features on AlexNet. Deep features eventually transit from general to specific along the network, and in AlexNet the domain-specific layers are $\mathcal{L} = \{fc6, fc7, fc8\}$ and $|\mathcal{L}| = 3$, which are not transferable and should be adapted with MMD.

In unsupervised domain adaptation, given a source domain $\mathcal{D}_s = \{(\mathbf{x}_i^s, \mathbf{y}_i^s)\}_{i=1}^{n_s}$ with n_s labeled samples and a target domain $\mathcal{D}_t = \{\mathbf{x}_j^t\}_{j=1}^{n_t}$ with n_t unlabeled samples, we apply transfer learning methods to bridge the two domains and generate labels in the target domain. The source domain and target domain are sampled from different probability distributions P and Q respectively, where their joint distributions are $P(\mathbf{X}^s, \mathbf{Y}^s)$ and $Q(\mathbf{X}^t, \mathbf{Y}^t)$ and $P \neq Q$. In our case, $\mathcal{D}_s$ is a labeled training set of current public databases with an uneven image distribution in age (sample of infants are very limited), and $\mathcal{D}_t$ is an unlabeled test set (*i.e.* BCS) only with images of kids. During domain adaptation, unsupervised deep neural network $\mathbf{y} = \theta(\mathbf{x})$ is designed to reduce the bias between different age groups and learn transferable features and classifiers, where the target risk $R_t(\theta) = \mathbb{E}_{(\mathbf{x},\mathbf{y})\sim Q}[\theta(\mathbf{x}) \neq \mathbf{y}]$ by jointly optimizing the source risk and domain discrepancy.

The main idea of domain adaptation is to decrease the domain discrepancy by finding an abstract feature representation through which the source and target domains are similar. In the case of deep transfer learning, it is central to minimize the CNN error defined in (1) on the source labeled data and the distinguishable feature representation of higher domain-specific layers $\mathcal{L}$ in the source and target domains jointly.

In this study, we consider two representative deep transfer learning techniques, *i.e.*, Deep Adaptation Networks (DAN) and Joint Adaptation Network (JAN), where the latter is the extension of the former.

In DAN [17], Maximum Mean Discrepancy (MMD) is used to measure the difference between the source and target domain. MMD is designed based on the idea that all samples in the same generating distribution are identical, *i.e.*, $P(\mathbf{X}^s) = Q(\mathbf{X}^t)$. The definition of MMD is the distance between embeddings of distributions in Reproducing Kernel Hilbert Spaces (RKHS) $\mathcal{H}$, and that between probability distributions P and Q respectively corresponding to source and target domain is:

$$d_{\mathcal{H}}(P,Q) \triangleq \sup_{\theta \in \mathcal{H}} \left(\mathbb{E}_{\mathbf{X}^s}[\theta(\mathbf{X}^s)] - \mathbb{E}_{\mathbf{X}^t}[\theta(\mathbf{X}^t)] \right). \tag{2}$$

The most important property is that $P = Q$ if and only if $d_{\mathcal{H}}(P,Q) = 0$ [24]. The characteristic kernel associated with the feature map ϕ, $k(\mathbf{x}^s, \mathbf{x}^t) = \langle \phi(\mathbf{x}^s), \phi(\mathbf{x}^t) \rangle$. So the square distance between the empirical kernel mean embeddings $d^2_{\mathcal{H}}(P,Q)$ as metrics to compare the discrepancy in each domain-specific layer,

$$\begin{aligned} d^2_{\mathcal{H}}(P,Q) = \ & \frac{1}{n_s^2} \sum_{i=1}^{n_s} \sum_{j=1}^{n_s} k(\mathbf{x}_i^s, \mathbf{x}_j^s) \\ & + \frac{1}{n_t^2} \sum_{i=1}^{n_t} \sum_{j=1}^{n_t} k(\mathbf{x}_i^t, \mathbf{x}_j^t) \\ & - \frac{2}{n_s n_t} \sum_{i=1}^{n_s} \sum_{j=1}^{n_t} k(\mathbf{x}_i^s, \mathbf{x}_j^t). \end{aligned} \tag{3}$$

Therefore, we can compute the MMD-based multi-layer adaptation regularizer by integrating the MMD estimator (3) to the CNN error (1) as follows:

$$\min_{\Theta} \frac{1}{n_s} \sum_{i=1}^{n_s} J(\theta(\mathbf{x}_i^s), \mathbf{y}_i^s)) + \lambda \sum_{\ell \in \mathcal{L}} d^2_{\mathcal{H}}(\mathcal{D}_s^\ell, \mathcal{D}_t^\ell), \tag{4}$$

where λ is a positive penalty parameter. $\mathcal{D}_*^\ell$ is the ℓth layer hidden representation for the source or target samples, and $d^2_{\mathcal{H}}(\mathcal{D}_s^\ell, \mathcal{D}_t^\ell)$ is the MMD between the source and target domain evaluated on the representation at the ℓth layer ($\ell \in \mathcal{L}$, and $\mathcal{L}$ is the task-specific layers which cannot be safely transferred). We illustrate the DAN optimization framework based on AlexNet in Figure 2, where $\mathcal{L} = \{fc6, fc7, fc8\}$. DAN learns transferable features from the source domain to the related target domain by measuring MMD (3) at each layer ℓ.

Training a deep CNN model requires a large amount of labeled data, but (3) generates a complexity of $O(n^2)$, which makes its application impossible. Moreover, mini-batch Stochastic Gradient Descent (SGD) is difficult to realize due to the summation over pairwise similarities between data points, which is important to the training effectiveness. To solve this problem, DAN [17] exploits the unbiased estimate of MMD which is computed with linear complexity according to $\hat{d}^2_{\mathcal{H}}(p,q) = \frac{2}{n_s} \sum_{i=1}^{n_s/2} g_{\mathcal{H}}(\mathbf{z}_i)$, where $g_{\mathcal{H}}(\mathbf{z}_i) \triangleq k(\mathbf{x}_{2i-1}^s, \mathbf{x}_{2i}^s) + k(\mathbf{x}_{2i-1}^t, \mathbf{x}_{2i}^t) - k(\mathbf{x}_{2i-1}^s, \mathbf{x}_{2i}^t) - k(\mathbf{x}_{2i}^s, \mathbf{x}_{2i-1}^t)$ as in [17]. It allows the computation of an expectation of independent variables as in (3) with cost $O(n)$, making MMD implementation in CNNs possible.

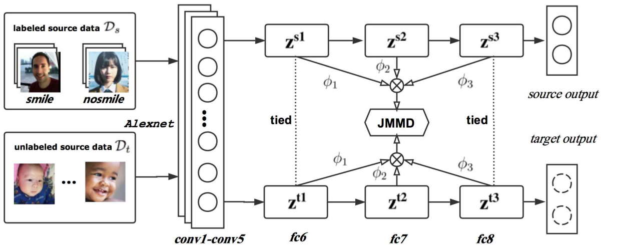

Figure 3. The JAN architecture for learning transferable features on AlexNet. Deep features eventually transit from general to specific along the network, and activations in multiple domain-specific layers $\mathcal{L}$ are not safely transferable. The joint distributions of the activations $P(\mathbf{Z}^{s1}, ..., \mathbf{Z}^{s|\mathcal{L}|})$ and $Q(\mathbf{Z}^{t1}, ..., \mathbf{Z}^{t|\mathcal{L}|})$ in these layers should be adapted by JMMD minimization. Here, in AlexNet, $\mathcal{L} = \{fc6, fc7, fc8\}$ and $|\mathcal{L}| = 3$.

It should be noted that DAN is under a hypothesis that feature layers and classifier layer are independent representations, and features can thus be directly transferred from the source classifier to the target one. However, this assumption may not hold. In this case, JAN is proposed to align the joint distributions of multiple domain-specific layers across domains using JMMD.

Different from MMD used in DAN which applies only uniform weights and does not take the influence of other variables in other layers into consideration, JMMD adopts non-uniform weight activations $\mathbf{Z}^{|\ell|}$ at each layer $\ell \in \mathcal{L}$. It thus emphasizes full interactions between different variables in joint distributions $P(\mathbf{Z}^{s1}, ..., \mathbf{Z}^{s|\mathcal{L}|})$ and $Q(\mathbf{Z}^{t1}, ..., \mathbf{Z}^{t|\mathcal{L}|})$.

Similar to MMD, JMMD is to measure the difference of Hilbert-Schmidt norm between kernel mean embedding of empirical joint distributions $P(\mathbf{Z}^{s1}, \ldots, \mathbf{Z}^{s|\mathcal{L}|})$ and $Q(\mathbf{Z}^{t1}, \ldots, \mathbf{Z}^{t|\mathcal{L}|})$ in the Hilbert space

$$d^2_{\mathcal{L}}(P,Q) \triangleq \| \mathcal{C}_{\mathbf{Z}^{s,1:|\mathcal{L}|}}(P) - \mathcal{C}_{\mathbf{Z}^{t,1:|\mathcal{L}|}}(Q) \|^2_{\bigotimes_{\ell=1}^{|\mathcal{L}|} \mathcal{H}^\ell}, \tag{5}$$

where $\mathcal{C}_{\mathbf{Z}^{s,1:|\mathcal{L}|}}(P) \triangleq \mathbb{E}_{\mathbf{Z}^{1:|\mathcal{L}|}}[\otimes_{\ell=1}^{|\mathcal{L}|} \theta(\mathbf{Z}^{s\ell})]$ is the joint distribution of P of variables $\mathbf{Z}^{s1}, \ldots, \mathbf{Z}^{s|\mathcal{L}|}$ embedded in space $\bigotimes_{\ell=1}^{|\mathcal{L}|} \mathcal{H}^\ell$. In unsupervised deep learning, we have activations at the layers $\mathcal{L}$ as $\{(\mathbf{z}_i^{s1}, \ldots, \mathbf{z}_i^{s|\mathcal{L}|})\}_{i=1}^{n_s}$ from n_s labeled samples in source domain $\mathcal{D}_s$ and $\{(\mathbf{z}_j^{t1}, \ldots, \mathbf{z}_j^{t|\mathcal{L}|})\}_{j=1}^{n_t}$ from n_t unlabeled samples in target domain $\mathcal{D}_t$. The square distance of JMMD between the empirical kernel mean embeddings is formulated as

$$\begin{aligned} d^2_{\mathcal{L}}(P,Q) = & \frac{1}{n_s^2}\sum_{i=1}^{n_s}\sum_{j=1}^{n_s}\prod_{\ell\in\mathcal{L}} k^\ell(\mathbf{z}_i^{s\ell},\mathbf{z}_j^{s\ell}) \\ & + \frac{1}{n_t^2}\sum_{i=1}^{n_t}\sum_{j=1}^{n_t}\prod_{\ell\in\mathcal{L}} k^\ell(\mathbf{z}_i^{t\ell},\mathbf{z}_j^{t\ell}) \qquad (6) \\ & - \frac{2}{n_s n_t}\sum_{i=1}^{n_s}\sum_{j=1}^{n_t}\prod_{\ell\in\mathcal{L}} k^\ell(\mathbf{z}_i^{s\ell},\mathbf{z}_j^{t\ell}). \end{aligned}$$

By explicitly minimizing JMMD, the discrepancy between the source and target domain can be decreased to enable domain adaptation.

In JAN, our main objective is to reduce the discrepancy in the joint distributions of the activations of higher domain-specific layers $\mathcal{L}$, *i.e.*, $P(\mathbf{Z}^{s1},...,\mathbf{Z}^{s|\mathcal{L}|})$ and $Q(\mathbf{Z}^{t1},...,\mathbf{Z}^{t|\mathcal{L}|})$ where the features are not safely transferable. The joint distributions is matched with the deep network model by adding the JMMD regularizer (6) into the risk of the deep model (1),

$$\min_{\Theta} \frac{1}{n_s}\sum_{i=1}^{n_s} J(\theta(\mathbf{x}_i^s),\mathbf{y}_i^s) + \lambda d^2_{\mathcal{L}}(P,Q), \qquad (7)$$

where λ is a positive tradeoff parameter of the JMMD penalty. As in DAN, for the JAN model we set domain specific $\mathcal{L}$ in AlexNet at the last three layers $\{fc6, fc7, fc8\}$, as shown in Figure 3, which are not safely transferable and should be jointly adapted by minimizing the CNN error and JMMD. However, limited by its quadratic complexity, JMMD (6) is impossible to be used in the mini-batch SGD algorithm. Just like the unbiased estimate of MMD [24], we apply a similar linear-time estimate of JMMD in practice for a linear measurement by sampling the same number of source and target points to eliminate the bias caused by the domain size. For ResNet, we do it in the same way as on AlexNet, but apply DAN and JAN to the last two layers, *i.e.*, the 5th pooling layer and the fully-connected layer.

4. Experiments

To validate the proposed approach, we conduct experiments on the popular public benchmarks as well as our newly collected BCS dataset. The data, protocols, and results are presented subsequently.

4.1. Databases

Two public databases, *i.e.* CelebFaces Attributes [16] (CelebA) and FotW, are adopted in our experiments. CelebA is a large-scale face database, which contains 202,599 images of 10,177 identities. Each image in CelebA is annotated with 40 facial attributes, and one of them is Smile/Non-Smile. Most of images in CelebA are clear and frontal faces of celebrated people. Due to its large amount of data, it is usually used to pre-train deep models for face analysis tasks. FotW is a recent standard benchmark for evaluating smile detection systems, which comes from the 2016 ChaLearn Looking at People and Faces of the World Challenge and Workshop. There are in total 17,517 images, where 6,171 are for training, 3,086 for validation, and 8,260 for test. The samples are of various ages, races, poses, backgrounds, and image qualities, which are closer to the ones captured in the real world case. Although FotW is more balanced than the other datasets, it still suffers from the uneven sample distribution in age. Specifically, according to our statistics, the ratio of images of infants and children who are younger than 5 years old is only around 3.8%.

Race	Smile	Non-Smile	All
Caucasian	222	227	449
African-American	152	160	312
Asian	247	227	484
All	621	624	1245

Table 1. Data distribution in the newly collected BCS dataset.

To demonstrate the problem that smile detection performance is largely affected by the age gap between infants and adults. We build a new database, namely Baby and Child Smile (BCS). It has 1,245 images, all of which belong to the children under 5 years. They are first roughly collected from Internet and then carefully screened out to achieve the balance in ethnicity and ratio of positive/negative (see Table 1 for more details). Meanwhile, the images also include variability in illumination, pose, accessory, *etc.* We manually annotate the images with either "Smile" or "Non-Smile". For pre-processing, we use Seetaface [32] to localize and crop faces. Figure 1 shows some examples.

4.2. Setup

We carry out three experiments as: Exp.1 to analyze the impact of age gap; Exp.2 to tune configuration in DAN; and Exp.3 to highlight the effectiveness of deep transfer learning methods. All the experiments are conducted in a cross-dataset manner, where the models are trained using the public benchmarks and tested on the entire BCS dataset. Meanwhile, we display the results on the validation partition of FotW in Exp.1 for contrast. In addition, besides average accuracy, we also report the score of recall on smile samples, for more comprehensive analysis.

In experiments, we use the similar fine-tuning pipeline as in [35], where we adopt CNNs (AlexNet and ResNet) pretrained by the large scale ImageNet for image classification. Then we fine-tune the models for the specific task smile of detection on the CelebA training set and the FotW training set. All the CNN models are implemented using the Caffe deep learning toolbox [14]. The momentum is set at 0.9, the weight decay is set at 0.0005, and the base learning rate is between 10^2 and 10^5 with a multiplicative step-size equal to $10^{1/2}$.

In DAN and JAN, we set lr_multi at 0.1 on all the convolutional layers while we set it at 0.2 on $fc6$ (AlexNet) and 0.5 on the last fully connected layers. We set the batch size at 64 and the total iterations at 200,000 in all the steps.

4.3. Results

Exp.1: Discrepancy Analysis. We evaluate the impact of the discrepancy between infants and adults on performance. The results are summarized in Table 2. It can be seen in this table, when fine-tuned using the data in the FotW training set, both AlexNet and ResNet achieve more than 85% accuracy and around 80% recall, and the scores of ResNet are slightly superior to the ones of AlexNet. Such accuracies are very close to the state of the art accuracy of 90.83% reported in [23] through a multi-task deep model. It indicates the effectiveness of the two baseline models.

However, when directly applying the baseline models to BCS which only contains samples of infants and children, the accuracy scores of AlexNet and ResNet are reduced to 59.12% and 69.00% respectively, both of which suffer a drop of more about 20% to 30%. Regarding recall, the case is even worse, with a dramatical fall of 30 to 50 points. It illustrates that the age difference between the two test sets is challenging to current deep model based methods. On the other hand, the very deep ResNet performs better than the basic AlexNet, which validates that deeper models not only boost the performance of specific tasks but also learn more transferable representations for domain adaptation. It confirms that the deep model can only reduce, but cannot remove the discrepancy of different datasets [33].

Method	Accuracy (%)	Recall (%)
on FotW Validation		
AlexNet	86.51	77.08
ResNet	87.03	79.32
on BCS		
AlexNet	59.12	25.60
ResNet	69.00	45.89

Table 2. Result comparison in terms of accuracy and recall on the FotW and BCS datasets.

Exp.2: Configuration Tuning in DAN. [17] states that DAN based deep domain adaptation with multi-layers is superior to that with only one hidden layer $fc6$, since different layers in deep networks extract features at different levels. We test it on AlexNet, and the results are displayed in Table 3. We can see that the adaptation with $fc7$ and $fc8$ outperforms that with either single layer, $fc7$ or $fc8$, in terms of accuracy and recall, which confirms this claim. Therefore, in Exp.3, we use DAN with multi-layers, *i.e.* $fc6$, $fc7$ and $fc8$ on AlexNet as well as $pool5$ and fc on ResNet.

Method	Accuracy	Recall
DAN_AlexNet($fc6$+$fc7$)	73.57%	63.12%
DAN_AlexNet($fc6$+$fc8$)	72.20%	61.03%
DAN_AlexNet($fc6$+$fc7$+$fc8$)	**77.03%**	**63.77%**

Table 3. Comparison of different configurations in DAN based on AlexNet.

Method	Accuracy (%)	Recall (%)
AlexNet [15]	59.12	25.60
DAN_AlexNet	77.03	63.77
JAN_AlexNet	83.85	74.23
ResNet [12]	69.00	45.89
DAN_ResNet	80.16	69.89
JAN_ResNet	**85.06**	**78.10**

Table 4. Comparison of results with and without deep transfer learning on the BCS database.

Exp.3: Contribution Assessment of Deep Transfer Learning. The results with and without transfer learning are demonstrated in Table 4. We can see that both the accuracies of AlexNet and ResNet are largely improved by transfer learning methods. Specifically, for AlexNet, DAN and JAN increase the baseline score (59.12%) to 77.03% and 83.85%, and for ResNet, they increase the baseline result (69.00%) to 80.16% and 85.06%. More importantly, the recall scores of smile samples are sharply promoted either by DAN or JAN. Figure 4 depicts some examples whose labels are corrected by transfer learning.

When we compare the results of AlexNet and ResNet, similar conclusions can be made as those in Table 2. Regarding the comparison between DAN and JAN, JAN always reaches better results than DAN does. Although they both adapt multiple domain-specific layers, DAN is based on an assumption that the feature and classifier layers are independent. DAN uses MMD as penalty, and the shift in the marginal distribution at each layer is reduced independently, while JAN adopts the JMMD penalty to reduce the shift in the joint distributions of multiple task-specific layers, reflecting the relationship between input features and output labels.

The results prove the effectiveness of deep transfer learning in solving the problem of discrepancy between infants and adults in smile detection. The difference between labeled source data and unlabeled target data is bridged, where very

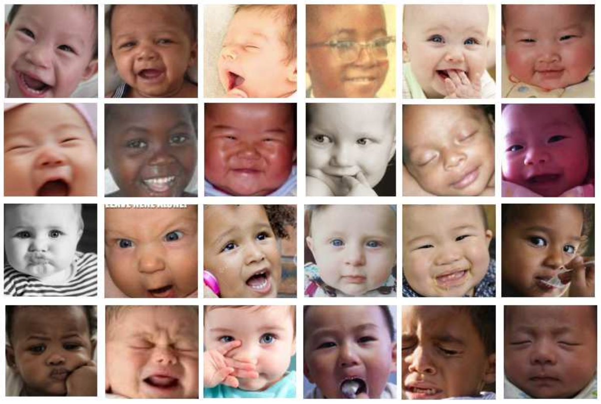

Figure 4. Some samples whose labels are corrected by transfer learning on the BCS database, where the first two rows show smile samples and the last two rows show non-smile ones.

limited labeled data are available for training. In practice, deep transfer learning methods can augment the expansibility of a trained model in a new dataset without the burden of heavy manual annotation work.

5. Conclusion

This paper discusses the issue of smile detection across the difference between infants and adults. We address it in the viewpoint of domain adaptation and propose a novel approach which integrates state of the art transfer learning techniques, *i.e.* DAN and JAN, and popular deep models. The proposed approach learns the knowledge from adults and successfully adapts it to infants, with very limited labeled data for training. Experimental results achieved in a cross-dataset scenario illustrate its effectiveness in smile detection in the presence of such an age gap.

Acknowledgment

This work is partly supported by the National Natural Science Foundation of China (No. 61673033); the Research Program of State Key Laboratory of Software Development Environment (SKLSDE-2017ZX-07); and the Microsoft Research Asia Collaborative Program (FY17-RES-THEME-033).

References

[1] X. Cao, D. Wipf, F. Wen, G. Duan, and J. Sun. A practical transfer learning algorithm for face verification. In *IEEE International Conference on Computer Vision*, pages 3208–3215, 2013.

[2] J. Chen, X. Liu, P. Tu, and A. Aragones. Person-specific expression recognition with transfer learning. In *IEEE International Conference on Image Processing*, pages 2621–2624. IEEE, 2012.

[3] Z. Cui, S. Zhang, J. Hu, and W. Deng. Evaluation of smile detection methods with images in real-world scenarios. In *Asian Conference on Computer Vision*, pages 166–179. Springer, 2014.

[4] S. Desai, M. Upadhyay, and R. Nanda. Dynamic smile analysis: changes with age. *American Journal of Orthodontics and Dentofacial Orthopedics*, 136(3):310–e1, 2009.

[5] Z. Ding, M. Shao, and Y. Fu. Deep low-rank coding for transfer learning. In *International Joint Conference on Artificial Intelligence*, 2015.

[6] S. Escalera, M. Torres Torres, B. Martinez, X. Baró, H. Jair Escalante, I. Guyon, G. Tzimiropoulos, C. Corneou, M. Oliu, M. Ali Bagheri, et al. Chalearn looking at people and faces of the world: Face analysis workshop and challenge 2016. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 1–8, 2016.

[7] S. S. Farfade, M. J. Saberian, and L.-J. Li. Multi-view face detection using deep convolutional neural networks. In *ACM International Conference on Multimedia Retrieval*, pages 643–650. ACM, 2015.

[8] C. Florea, L. Florea, and C. Vertan. Learning pain from emotion: Transferred hot data representation for pain intensity estimation. In *European Conference on Computer Vision Workshops*, pages 778–790, 2014.

[9] Y. Ganin and V. Lempitsky. Unsupervised domain adaptation by backpropagation. In *International Conference on Machine Learning*, pages 1180–1189, 2015.

[10] Y. Gao, H. Liu, P. Wu, and C. Wang. A new descriptor of gradients self-similarity for smile detection in unconstrained scenarios. *Neurocomputing*, 174:1077–1086, 2016.

[11] P. O. Glauner. Deep convolutional neural networks for smile recognition. *arXiv preprint arXiv:1508.06535*, 2015.

[12] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *IEEE Conference on Computer Vision and Pattern Recognition*, pages 770–778, 2016.

[13] V. Jain and J. L. Crowley. Smile detection using multi-scale gaussian derivatives. In *WSEAS International Conference on Signal Processing, Robotics and Automation*, 2013.

[14] Y. Jia, E. Shelhamer, J. Donahue, S. Karayev, J. Long, R. Girshick, S. Guadarrama, and T. Darrell. Caffe: Convolutional architecture for fast feature embedding. In *ACM International Conference on Multimedia*, pages 675–678. ACM, 2014.

[15] A. Krizhevsky, I. Sutskever, and G. E. Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in Neural Information Processing System*, pages 1097–1105, 2012.

[16] Z. Liu, P. Luo, X. Wang, and X. Tang. Deep learning face attributes in the wild. In *IEEE International Conference on Computer Vision*, pages 3730–3738, 2015.

[17] M. Long, Y. Cao, J. Wang, and M. Jordan. Learning transferable features with deep adaptation networks. In *International Conference on Machine Learning*, pages 97–105, 2015.

[18] M. Long, J. Wang, G. Ding, J. Sun, and P. S. Yu. Transfer feature learning with joint distribution adaptation. In *IEEE International Conference on Computer Vision*, pages 2200–2207, 2013.

[19] M. Long, J. Wang, and M. I. Jordan. Deep transfer learning with joint adaptation networks. *arXiv preprint arXiv:1605.06636*, 2016.

[20] M. Long, H. Zhu, J. Wang, and M. I. Jordan. Unsupervised domain adaptation with residual transfer networks. In *Advances in Neural Information Processing Systems*, pages 136–144, 2016.

[21] S. J. Pan and Q. Yang. A survey on transfer learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(10):1345–1359, 2010.

[22] N. P. Ramaiah, E. P. Ijjina, and C. K. Mohan. Illumination invariant face recognition using convolutional neural networks. In *IEEE International Conference on Signal Processing, Informatics, Communication and Energy Systems*, pages 1–4. IEEE, 2015.

[23] R. Ranjan, S. Sankaranarayanan, C. D. Castillo, and R. Chellappa. An all-in-one convolutional neural network for face analysis. In *IEEE International Conference on Automatic Face and Gesture Recognition*, pages 17–24. IEEE, 2017.

[24] D. Sejdinovic, B. Sriperumbudur, A. Gretton, and K. Fukumizu. Equivalence of distance-based and rkhs-based statistics in hypothesis testing. *The Annals of Statistics*, pages 2263–2291, 2013.

[25] C. Shan. Smile detection by boosting pixel differences. *IEEE Transactions on Image Processing*, 21(1):431–436, 2012.

[26] M. Shao, C. Castillo, Z. Gu, and Y. Fu. Low-rank transfer subspace learning. In *IEEE International Conference on Data Mining*, pages 1104–1109, 2012.

[27] S. Si, D. Tao, and B. Geng. Bregman divergence-based regularization for transfer subspace learning. *IEEE Transactions on Knowledge and Data Engineering*, 22(7):929–942, 2010.

[28] B. Smolka and K. Nurzynska. Power lbp: a novel texture operator for smiling and neutral facial display classification. *Procedia Computer Science*, 51:1555–1564, 2015.

[29] Y. Su, Y. Fu, Q. Tian, and X. Gao. Cross-database age estimation based on transfer learning. In *IEEE International Conference on Acoustics Speech and Signal Processing*, pages 1270–1273. IEEE, 2010.

[30] E. Tzeng, J. Hoffman, T. Darrell, and K. Saenko. Simultaneous deep transfer across domains and tasks. In *IEEE International Conference on Computer Vision*, pages 4068–4076, 2015.

[31] E. Tzeng, J. Hoffman, N. Zhang, K. Saenko, and T. Darrell. Deep domain confusion: Maximizing for domain invariance. *arXiv preprint arXiv:1412.3474*, 2014.

[32] S. Wu, M. Kan, Z. He, S. Shan, and X. Chen. Funnel-structured cascade for multi-view face detection with alignment-awareness. *Neurocomputing (under review)*, 2016.

[33] J. Yosinski, J. Clune, Y. Bengio, and H. Lipson. How transferable are features in deep neural networks? In *Advances in Neural Information Processing Systems*, pages 3320–3328, 2014.

[34] K. Zhang, Y. Huang, H. Wu, and L. Wang. Facial smile detection based on deep learning features. In *IAPR Asian Conference on Pattern Recognition*, pages 534–538. IEEE, 2015.

[35] K. Zhang, L. Tan, Z. Li, and Y. Qiao. Gender and smile classification using deep convolutional neural networks. In *IEEE Conference on Computer Vision and Pattern Recognition Workshops*, pages 34–38, 2016.