From Face Recognition to Kinship Verification: An Adaptation Approach

Qingyan Duan[1], Lei Zhang[1]*, and Wangmeng Zuo[2]
[1]College of Communication Engineering, Chongqing University, Chongqing, China
[2]School of Computer Science and Technology, Harbin Institute of Technology, Harbin, China
qyduan@cqu.edu.cn, leizhang@cqu.edu.cn, wmzuo@hit.edu.cn

Abstract

*Kinship verification in the wild is a challenging yet interesting issue, which aims to determine whether two unconstrained facial images are from the same family or not. Most previous methods for kinship verification can be divided as low-level hand-crafted features based shallow methods and kin data only trained convolutional neural network (CNN) based deep methods. Worthy of affirmation, numerous work in vision get that convolutional features are discriminative, but bigger data dependent. A fact is that for a variety of data-limited vision problems, such as limited Kinship datasets, the ability of CNNs is seriously dropped because of overfitting. To this end, by inheriting the success of deep mining algorithms on face verification (e.g. LFW), in this paper, we propose a **C**oarse-to-**F**ine **T**ransfer (CFT) based deep kinship verification framework. As the idea implied, this paper tries to answer "is it possible to transfer a face recognition net to kinship verification?". Therefore, a supervised coarse pre-training and domain-specific ad hoc fine re-training paradigm is exploited, with which the kin-relation specific features are effectively captured from faces. Extensive experiments on benchmark datasets demonstrate that our proposed CFT adaptation approach is comparable to the state-of-the art methods with a large margin.*

1. Introduction

Human face carries with lots of individual information, and most human characteristics such as identity, age, gender, emotion etc. can be distinguished by facial images. Facial analysis has been widely studied in computer vision. In recent years, face recognition, that aims to discover the inherent identity-associated facial features, has witnessed a great achievement promoted by deep learning. The objective of face recognition is to identify who is the person in a given human facial image, while face verification tries to answer whether the two persons belong to the same person [14]. Also, the facial images can also reflect kin-relation, and it is challenging to recognize whether the two persons are from the same family. Therefore, an emerging topic, kinship verification, that aims to mining implicit kin-relation specific features, has been raised. It has many potential applications, such as missing children searching and social media mining, etc. [11]. In this work, the parent-child based kinship is studied, such as father-daughter, father-son, mother-daughter and mother-son. Some facial image pairs with kinship and no kinship have been shown in Figure 1, from which the difficulty of kin-relation discovery is shown.

Figure 1. Some samples positive (with kinship relation) and negative pairs (no kinship relation) from KinFaceW-I, KinFaceW-II, Cornell KinFace and UB KinFace, respectively. The first two rows are positive pairs and the last two rows are negative pairs. The kinship relation types from left to right are: father-daughter, father-son, mother-daughter and mother-son, respectively.

Recently, many algorithms have been proposed for kinship verification. Most of these work follow the technical routine from hand-crafted low-level feature extraction to large-margin metric learning. A representative work can be referred to as [11], in which a neighborhood repulsed metric learning (NRML) was proposed by learning a projection based metric with large margin and achieved the best per-

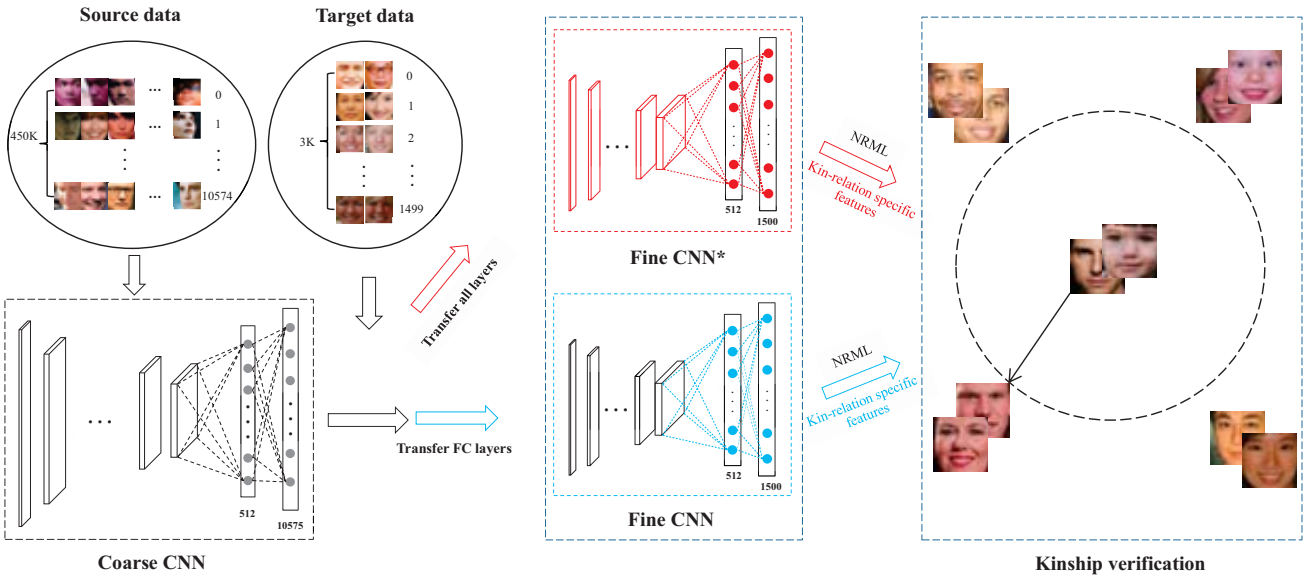


Figure 2. Pipeline of our proposed approach

formance on ensemble of hand-crafted low-level features. Although these work greatly promote kinship verification, they follow a conventional face recognition route that is kinship data dependent. Also, the hand-crafted feature extraction (e.g. LBP, HOG) is often used for general face analysis, but independent of kin-relation specific features. As a result, the implicit and abstract kinship information cannot be adequately represented [10]. Therefore, the kin-relation *specific* feature mining and discovery is still a challenging issue, which is also the focus of this paper to address.

Deep learning, proposed by Hinton and Salakhutdinov [7], has become the most popular machine learning algorithms for discovering discriminative intermediate and high-level representations in a hierarchical manner [5]. In particular, convolutional neural network (CNN) have recently been shown to achieve great success in various computer vision tasks, such as face recognition [14, 21], object recognition, etc. All these achievements are attributed to massive labeled natural image data and large-scale parameters for high-level discriminative features representation. While compared with conventional machine learning, deep learning as a supervised approach, depends on large-scale data. Otherwise, overfitting may be encountered in small tasks. In face recognition task, the deep CNN model is generally trained on a large-scale constrained or unconstrained face database (i.e., CASIA Webface). Recently, CNNs have also been used for kinship verification [10, 20]. These work are closely related with this paper, but different in essence. These two work tend to train a shallow CNN model based on limited kinship data (several hundreds or thousands) for deep kinship facial feature representation. However, the ability of CNN is dropped due to data scarcity. Depth is an important aspect of CNN architecture [18], and with the deepen of CNN architecture the performance become better, but more parameters are needed. Therefore, overfitting in training will result in the singular kin-relation features.

The most straightforward method to solve the CNN based kinship verification issue is to prepare large-scale, structural and labeled kinship dataset. However, it is cost prohibitive and time consuming to structure a large kinship datasets with correctly annotated human facial images. To address this issue, in this paper, we propose a deep transfer learning paradigm for kinship verification, which is called Coarse-to-Fine Transfer (CFT). As the idea implies, we expect to inherit the success of face verification in LFW and object recognition in ImageNet into kinship verification, by pre-training a coarse CNN (cCNN) on a large-scale face recognition database (i.e., source data) and re-training a fine CNN (fCNN) model on the small-scale kinship database (i.e., target data). After integrating the cCNN and fCNN models together, a CFT network is formulated for kinship verification. In CFT framework, the cCNN is used to discover the generalized facial features and the fCNN is used to deep mining the kin-relation specific features. In order to further strengthen the discrimination of deep kin-relation features, the NRML [11] is used to seek a feature projection matrix, so that the deep kinship features can be projected into a kin-discriminated feature space. To our best knowledge, deep transfer learning has not been studied in kinship verification. Experimental results show that our proposed coarse-to-fine transfer method can well learn the transferable kin-relation specific features, and prove that it is feasi-

ble to transfer face recognition net to kinship verification.

The key contributions of this work are threefold.

- Different from previous methods in kinship verification, a much deeper CNN based coarse-to-fine transfer method is proposed to adequately extract deep kin-relation specific features based on cCNN and fCNN, which are more universal and discriminative for Kinship verification.

- To the best of our knowledge, it is the first time to exploit deep transfer learning for kinship verification. Our approach relaxes the kinship domain data, and train the fCNN based on a multi-class yet simple learning mechanism.

- Experimental comparison with shallow and deep learning methods demonstrate that our proposed CFT method is comparable to the state-of-the-arts. Further, our method is also shown to follow humans pace in Kinship verification, and closing the gap of human-machine performance.

2. Related Work

In this section, we review three closely related topics with this paper, including kinship verification, deep convolutional neural networks and transfer learning.

2.1. Kinship Verification

Kinship verification via facial image analysis is an challenging problem in computer vision. Since many researchers have investigated this problem, many kinds of learning based methods have been proposed. Those methods can be mainly divided into two categories: 1) feature-based [2, 4, 28] and 2) model-based [24]. The former aim to use general low-level feature descriptor to represent facial image. Existing feature representation approaches for kin-relation data include histogram of gradient (HOG) [4], scale-invariant feature transform (SIFT) [11, 27], local binary pattern (LBP) [11]. The ensemble of the above hand-crafted features demonstrates a superior kinship verification performance [11, 27]. Regardless of the hand-crafted features, data-driven and high-level CNN features [10, 20] have also been used for kinship verification, and show a significant progress compared with hand-designed features. The latter aim to learning an effective metric or model used to distinguish whether two face images are with kinship relation, such as neighborhood repulsed metric learning (NRML) [11], prototype-based discriminative feature learning (PDFL) [28], transfer subspace learning [15, 23] support vector machine (SVM) [28], large margin multi-metric learning [8], ensemble similarity learning (ESL) [32], and scalable similarity learning (SSL) [33].

Those previous works have achieved great progress over the challenging kinship verification. However, the common shortcoming is that the extracted image features are general representation of faces and lack of structural kin-relation meaning. To this end, the proposed deep CNN model based Coarse-to-Fine Transfer can be a competitive candidate for better insight of the implicit kin-relation characteristic inside the facial images.

2.2. Deep Convolutional Networks

Deep learning has shown its effectiveness in various computer vision tasks, such as face recognition and object recognition. CNN is an end-to-end supervised learning methods from pixel based images to the high-level semantic. The features from the bottom to top in the network architecture can be identified as low-level and high-level image representation. Several popular CNN models are summarized as follows. VGG [16], as a very deep CNN, achieved a great success in Large Scale Visual Recognition Challenge 2015. GoogLeNet [18] was proposed with deeper structure. A 152 layed ResNet with skip connection was also proposed for image recognition [6]. In face recognition, MTCNN [30] used the candidate CNNs to detect facial landmarks. A Deepface [19] was proposed a 3D-align. FaceNet [14] constructs a triplet-loss model and improve the face verification accuracy. The approach proposed in [26] can handle multi-modal face recognition with poses. Recently, the center-loss model proposed in [21] aims to obtain within-class separable features. Additionally, a very popular Faster R-CNN [13] has been proved to be very efficient and effective in object/pedestrain detection. All these studies in computer vision achieve surprisingly good performance, which motivates us to exploit deep convolutional networks for kinship verification.

Compared with conventional machine learning methods, there too many parameters needed to be learned in CNN, which, therefore, depends on a large-scale labeled database. Thus, it is not feasible to train a deeper CNN through small-scale kinship data. To this end, transfer learning is introduced in our method for learning the transferable features from large-scale domain data.

2.3. Transfer Learning

For statistical machine learning, the goal of learning is to obtain a classification model based on the well prepared training data, then the testing data with similar distribution can be predicted using the trained model. Nevertheless, an abundant annotated training database is difficult to acquire, and machine learning methods work under a common assumption that the training and testing data are drawn from the same feature space or distribution. This assumption may not hold in many real applications [12] due to the uncertainty of sampling conditions. Transfer learning aims to solve these above problems, by leveraging the large-scale, heterogenous domain data, and learn a powerful model for different tasks. Recently, transfer learning techniques have been applied successfully in computer vision applications. Wu and Dietterich [22] proposed to use both inadequate target domain data and plenty of low quality source domain

data for image classification. In [25], a novel feature space independent semi-supervised kernel matching method was proposed for domain adaptation. Latent sparse domain transfer (LSDT) [31] was proposed to solve domain adaptation and visual categorization of heterogeneous data.

Deep learning can be identified as a data-driven transfer learning technology, which is trained on large-scale heterogeneous data. Generally, the pre-trained deep models are used for another new domain independent of the training data. For example, in [9], a CNN model pretrained on ImageNet was used to predict poverty through satellite imagery by two-step transfer learning. Sun and Shetty investigated cross-domain transfer learning between video frames and web images by using pre-trained deep convolutional neural networks [17]. Esteva et al. [3] proposed to pre-train a CNN model on ImageNet, and fine-tune the model again using 13K clinical images. The superior performance has been compared against 21 board-certified dermatologists on biopsy-proven clinical images. The authors in [1, 29] have discussed the transferability of deep neural networks with extensive experiments by fine-tuning on a domain data.

Inspired by deep learning and transfer learning [1, 3, 9, 29], we propose a deep transfer model (CFT), which consists of a coarse CNN model (cCNN) and a fine CNN model (fCNN). The former is pre-trained on a large-scale face recognition database for coarse generalized facial features. The latter is re-trained on the small-scale Kinship database by using a 5-fold cross-validation strategy for structuring fine kin-relation specific features.

3. Coarse-to-Fine Transfer Approach

In our approach, kinship specific features are extracted based on a Coarse-to-Fine Transferring (CFT) paradigm. In order to obtain more discriminative and robust deep features, CFT is constructed based on a deep CNN architecture, consisting of a coarse CNN and a fine CNN. First, the cCNN is pre-trained on a large-scale facial image database as source data via face recognition based training mechanism. Then, the fCNN model is re-trained on small-scale kinship database based on cCNN via multi-class learning rule (two kinship images per class). For each facial image, cCNN is used for *generalized* facial features and fCNN is used for kin-relation *specific* characteristic features. Figure 2 shows the pipeline of our proposed transferring approach from face recognition net to kinship verification.

3.1. Coarse Convolutional Network (cCNN)

The proposed deep transfer model, that is based on convolutional neural network architecture, has over 7 million parameters. Similar to general CNN model, the proposed CFT model consists of convolutional layers, pooling layers, fully-connected layers and soft-max layer.

Numerous research has shown that the performance of CNN model is greatly attributed to the net depth [6, 18]. That is, the deeper the network is, the better the performance is. Therefore, we prefer using smaller convolution kernel rather than a bigger one, such that the network is deeper but without increasing the number of network parameters. For example, two 3×3 convolution kernel are used instead of one 5×5 kernel in our model. The convolution with ReLU nonlinearity is formulated as

$$\mathbf{Y}_j^l = \max(0, b_j^l + \sum_i \mathbf{W}_{ij}^l * \mathbf{X}_i^l) \tag{1}$$

where $\mathbf{X}_i^l$ and $\mathbf{Y}_j^l$ are the i-th input feature map and j-th output feature map in the l-th convolution layers, respectively. $\mathbf{W}_{ij}^l$ is the convolution kernel between the i-th input feature map and j-th output feature map. b_j^l is the bias of j-th output feature map, and $*$ denotes convolution operation. Note that the ReLU nonlinearity is used as activation function in each convolution layer, because it has been proved to have faster convergence and better stability than others.

After each two convolution layers, a pooling layer is followed for translation invariance, dimension reduction, and avoiding overfitting. In general, considering the ReLU activation results, the max-pooling layer is adopted, which is defined as

$$y_i^{(j,k)} = \max_{0 \le m,n \le s} \{x_i^{(j \cdot s+m, k \cdot s+n)}\} \tag{2}$$

where $y_i^{(j,k)}$ denotes the output of the i-th feature map in the location (j,k). Similarly, $x_i^{(j,k)}$ is the value of location (j,k) in the i-th feature map. The neighboring region size of max-pooling layer is 2×2.

In order to compare with other algorithms, the size of each input RGB image to cCNN is 64×64. Note that one pooling layer is deployed after two convolutional layers. The details of the proposed CFT network configuration for each CNN model are described in Table 1. The training process of cCNN follows a face recognition mechanism based on CASIA WebFace database (source data), which includes 494,414 unconstrained facial images of 10575 persons. After cleaning the data of very low quality, the final training data includes 452,720 images of 10575 persons. That is, the cCNN model is trained over 10,000 classes. The cCNN is proposed for data scarcity problem of Kinship verification, and transferred to another Kinface domain data by leveraging the fCNN introduced in the next section. The details for detection and alignment can be found in Section 3.2.

3.2. Fine Convolutional Network (fCNN)

The fCNN shares a similar network architecture and optimization algorithm with the cCNN. The difference lies in that the training data of cCNN is a large-scale face

Conv1	Pool1	Conv2	Pool2	Conv3	Pool3	Conv4	Pool4	Conv5	FC
conv11-32 conv12-64	max-2	conv21-64 conv22-128	max-2	conv31-96 conv32-192	max-2	conv41-128 conv42-256	max-2	conv51-160 conv52-320	FC1-512 FC2-10575

Table 1. CFT Configuration. The convolution layers are denoted as "conv⟨index⟩-⟨$No.$ of kernels⟩", the pooling layer parameters are denoted as "⟨pooling operator⟩-⟨grid size⟩", and the fully-connected layer parameters are denoted as "FC⟨index⟩-⟨$No.$ of outputs⟩"

database with about 500K labeled images over 10K classes (source data), while fCNN is trained on a small-scale kinface database (target data) with about 3K labeled images over 1K classes. However, fCNN still inherits the merit of cCNN for generalized facial feature (e.g. edges, corners, texture) and further achieves higher-level kin semantic feature. The fundamental reason why we propose deep transfer model is that for the existing kinship dataset, such as KinFaceW-I[1], KinFaceW-II[1], Cornell KinFace[2], and UB KinFace[3], the total number of the doublet or triplet positive pairs does not exceed 5K.

It is worth noting that, although both source domain and target domain are with facial images, there are essential differences in data protocol. The protocol of CASIA WebFace[4] data classification based, that is, the goal of face recognition is to identify who she/he is. However, the kinship verification aims to determine whether there is a kinship relation between two persons, and the label is 1 if yes, otherwise 0. As a result, the training protocol of target domain cannot be transferred to source domain for training the fCNN. To this end, we make a similar protocol on Kinship data with WebFace. Specifically, we select the positive pairs of parent-chid images and manually tag each positive pair as different label starting from digit 0. That is, for each positive parent-child pair, they are marked as the same identity. In this way, the target data and source data share a similar training protocol. During the fCNN training, a 5-fold cross-validation strategy is used. Therefore, the kin faces training data (4 folds) includes 3162 images of 1500 classes. For each kinface database except the UB KinFace data, 2 images per class are considered. UB KinFace is different from other three kinship datasets that it is constructed in triplet: children, young parents and old parents. The young parent and the old parent in each triplet are the same person but age different. Therefore, for UB KinFace, 3 images per class are considered. For performance evaluation, with 4 folds for training and the remaining 1 fold for testing, the average accuracy of 5-fold is reported.

Additionally, there are two ways for the fCNN training, and results in two associated algorithms CFT and CFT*.

- CFT: high-level kin-relation feature mining by training the fully-connected layers of fCNN on the kinship training data with convolutional layers frozen.

- CFT*: low-level and high-level feature mining by slightly training both convolutional layers together with fully-connected layers of fCNN on the kinship training data.

Notably, mini-batch Stochastic Gradient Descent (SGD) based error back propagation algorithm is used for training the proposed CFT network. The CFT is implemented by using Caffe and Python package.

3.3. Data Pretreatment: Face Detection and Alignment

For cCNN training on WebFace database and fCNN training on Kinship database, face detection is implemented by using MTCNN[5] facial point detection method proposed by [30], which detects five facial landmarks, including the two eye centers, the nose tip, and the two mouth corners. If the detection fails, we simply discard the image if it is from WebFace database, but use hand-crafted landmarks if it is from Kinship database. Then, the detected faces are globally aligned by using affine transformation based on the two eye centers and the mid-point between the two mouth corners. Finally, the aligned 64×64 facial images are obtained as the inputs of CNNs. For feature extraction of each kinface, the output of CFT is a kin-relation specific feature vector with length of 512, that is ready for verification.

4. Verification Metric

Verification metric is used to measure the similarity of two faces based on the CFT features. In this paper, we have applied the intrinsic and learning free Euclidean distance metric and an ad hoc NRML metric for kinship verification.

4.1. Learning Free Metric

The most intrinsic metric used to calculate the similarity of each two vectors x_1 and x_2 is Euclidean distance metric, which is learning free metric, and can be calculated as

$$d_{\mathrm{L}} = \sqrt{\sum_{k=1}^{n}(x_{1k} - x_{2k})^2} \tag{3}$$

where x_1 and x_2 are two samples with n-dimensional column vectors, respectively. Generally, the similarities of intra-class samples (with kin-relation) should be higher than inter-class samples (without kin-relation).

[1] http://www.kinfacew.com/download.html

[2] http://chenlab.ece.cornell.edu/people/ruogu/kin_verify.html

[3] http://www1.ece.neu.edu/ yunfu/research/Kinface/Kinface.htm

[4] http://www.cbsr.ia.ac.cn/english/CASIA-WebFace-Database.html

[5] https://github.com/kpzhang93/MTCNN_face_detection_alignment

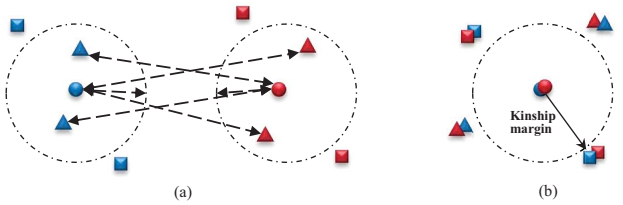


Figure 3. Intuitive illustration of NRML. (a) High-dimensional feature space. The data points in blue and red denote parents and children, respectively. The data points with kinship relation are denoted as circles. The data points in the neighborhood and non-neighborhood are denoted as triangles and squares, respectively. (b) The new NRML subspace, where a kinship margin is obtained.

4.2. Ad hoc Metric

Metric learning aims to structure an appropriate feature space, in which the structural difference between features can be measured. Neighborhood repulsed metric learning (NRML) [11] was proposed to seek a projected matrix, that can project the features from original space to a another space, where the intra-class samples are pulled as close as possible and the inter-class samples lying in a neighborhood are repulsed as far as possible. The idea of NRML is described in Figure 3. The model of NRML is formulated as

$$\begin{aligned} &\max_{n} J(W) = tr[W^T(H_1 + H_2 - H_3)W] \\ &\text{subject to} \quad W^T W = I, \end{aligned} \tag{4}$$

where $W^T W = I$ is a orthogonal constraint to restrict the scale of W, $H_1 \triangleq \frac{1}{NK}\sum_{i=1}^{N}\sum_{t_1=1}^{k}(x_i - y_{it_1})(x_i - y_{it_1})^T$, $H_2 \triangleq \frac{1}{NK}\sum_{i=1}^{N}\sum_{t_2=1}^{k}(x_{it_2} - y_i)(x_{it_2} - y_i)^T$, $H_3 \triangleq \frac{1}{N}\sum_{i=1}^{N}(x_i - y_i)(x_i - y_i)^T$, x_i and y_i are two m-dimensional column vectors of the ith parent-child kin pair, respectively, y_{it_1} represents the t_1th k-nearest neighbor of y_i and x_{it_2} denotes the t_2th k-nearest neighbor of x_i. The W can be obtained by solving an Eigen-decomposition problem. Then we use the W to project our CFT features into another space where cosine distance is calculated for Kinship verification.

5. Experiments

In this section, in order to demonstrate the effective of our proposed approach CFT, we experimented with four benchmark kinship datasets.

5.1. Datasets

In experiments, two kinds of databases are considered: large-scale CASIA WebFace data (500K) and small-scale KinFace data (4K). The KinFace data include four publicly available datasets, such as KinFaceW-I, KinFaceW-II [11], Cornell KinFace [4] and UB KinFace [23].

Method	F-S	F-D	M-S	M-D	Mean
Human A [28]	62.0	60.0	68.0	72.0	65.6
Human B [28]	68.0	66.5	74.0	75.0	70.9
MNRML [11]	72.5	66.5	66.2	72.0	69.6
MPDFL [28]	73.5	67.5	66.1	73.1	70.1
SMCNN [10]	75.0	**75.0**	68.7	72.2	72.7
DKV [20]	71.8	62.7	66.4	66.6	66.9
CFT	**79.5**	71.6	73.3	79.9	76.1
CFT*	78.8	71.7	**77.2**	**81.9**	**77.4**

Table 2. Accuracy of differen methods on KinFaceW-I

Method	F-S	F-D	M-S	M-D	Mean
Human A [28]	63.0	63.0	71.0	75.0	68.0
Human B [28]	72.0	72.5	77.0	80.0	75.4
MNRML [11]	76.9	74.3	77.4	77.6	76.5
MPDFL [28]	77.3	74.7	77.8	78.0	77.0
SMCNN [10]	75.0	**79.0**	78.0	**85.0**	**79.3**
DKV [20]	73.4	68.2	71.0	72.8	71.3
CFT	75.4	68.8	77.4	77.8	75.9
CFT*	**77.4**	76.6	**79.0**	83.8	**79.3**

Table 3. Accuracy of differen methods on KinFaceW-II

Method	Set 1	Set 2	Mean
MNRML [11]	66.8	67.3	67.1
MPDFL [28]	67.0	67.5	67.3
CFT	**70.3**	**74.3**	**72.3**
CFT*	66.5	64.5	65.5

Table 4. Accuracy of differen methods on UB KinFace

Method	MNRML [11]	MPDFL [28]	CFT	CFT*
Mean	71.6	71.9	**78.6**	78.3

Table 5. Accuracy of differen methods on Cornell KinFace

- Both KinFaceW-I and KinFaceW-II include four different types of kin relationships: father-son (F-S), father-daughter (F-D), mother-son (M-S) and mother-daughter (M-D). KinFaceW-I consists of 156, 134, 116, and 127 pairs, respectively. KinFaceW-II consists of 250 pairs for each relationship.

- Cornell KinFace contains totally 150 parent-child pairs.

- UB KinFace contains 200 triplets and each triplet is structured by child, young parent and old parent.

5.2. Experimental Setup

In experiments, the cCNN is first trained on WebFace, then the fCNN is trained on KinFace via 5-fold cross validation, and finally Euclidean metric and NRML metric are used for kinship verification.

We have compared our CFT method with four state-

Methods	KinFaceW-I				KinFaceW-II				UB		Cornell
	F-S	F-D	M-S	M-D	F-S	F-D	M-S	M-D	0-1	0-2	-
cCNN	76.6	66.8	72.4	77.2	75.2	67.0	76.2	73.8	70.3	66.3	73.7
CFT	**79.5**	71.6	73.3	79.9	75.4	68.8	77.4	77.8	**70.3**	**74.3**	**78.6**
CFT*	78.8	**71.7**	**77.2**	**81.9**	**77.4**	**76.6**	**79.0**	**83.8**	66.5	64.5	78.3

Table 6. Accuracy of cCNN and CFT on kinship databases

Methods	KinFaceW-I				KinFaceW-II				UB		Cornell
	F-S	F-D	M-S	M-D	F-S	F-D	M-S	M-D	0-1	0-2	-
ED based CFT	75.0	**88.4**	68.5	76.4	71.4	65.0	73.8	75.0	**71.3**	58.0	72.7
ED based CFT*	71.5	73.9	71.1	77.5	73.6	74.6	76.8	80.0	60.5	49.0	73.0
NRML based CFT	**79.5**	71.6	73.3	79.9	75.4	68.8	77.4	77.8	70.3	**74.3**	**78.6**
NRML based CFT*	78.8	71.7	**77.2**	**81.9**	**77.4**	**76.6**	**79.0**	**83.8**	66.5	64.5	78.3

Table 7. Accuracy of differen metric used on CFT

of-the-art methods in kinship verification, including two shallow learning methods such as MNRML [11] and MPDFL [28], and two deep learning methods such as SMCNN [10] and DKV [20]. Additionally, the performance comparison with human score [28] is also analyzed. Notably, for all algorithms, 5-fold cross-validation is used.

5.3. Comparison with Shallow Algorithms

The verification results of the proposed CFT (the fCNN' fully-connected layers are trained) and CFT* (all layers of fCNN are trained) on KinFaceW-I and KinFaceW-II have been shown in Table 2 and Table 3, respectively. The results of MNRML is copied from [11]. For fair comparison, the best result with feature ensemble of the compared methods are presented. As can be seen from these two tables, our proposed CFT methods show competitive performance.

Specifically, from the results listed in Table 2 and 3, we can observe that:

- The proposed CFT and CFT* methods consistently outperform state-of-the art face verification methods, i.e. MNRML and MPDFL based on feature ensemble and metric learning. Also, the effectiveness of high-level kin-relation semantic discovery has been demonstrated.

- The proposed CFT based methods also outperform the deep learning based face verification methods, i.e. SMCNN and DKV which are modeled on KinFace Only. The transferability of the proposed methods has been shown.

- By comparing our method with human knowledge on the KinFaceW-I and KinFaceW-II, the results show that CFT methods achieve even better performance than human.

- By comparing CFT with CFT*, we get that CFT* with all layers trained shows a superiority to CFT with Only fully-connected layers trained. That is, kin-relation specific features are associated with low-level and high-level layers on KinFaceW-I and KinFaceW-II datasets.

- By comparing Table 2 with Table 3, the performance on KinFaceW-II seems to be better than KinFaceW-I. This may be due to that the data sampling for each pair is from the same scene with different size. Interestingly, this phenomenon does not happen in CFT methods. The reason is that we use a massive WebFace data for transfer learning, and a large prior knowledge about faces have been captured. Thus, the performance dose not strongly depend on KinFace self, and the robustness of our proposed CFT is shown.

Further, the experimental results on UB KinFace and Cornell KinFace have been shown in Table 4 and Table 5, respectively. In Table 4, two subsets is constructed from the UB KinFace database: Set 1 (200 children and young parents image pairs) and Set 2 (200 children and old parents image pairs). From these tables, we can see that the proposed CFT methods outperform state-of-the-art MNRML and MPDFL 5 and 7 percentage, respectively. The superiority of the proposed methods are demonstrated. Notably, for UB KinFace, the CFT* fails, this may be caused by the triplet structure of UB. When constructing labels, the young-old parents with age difference are also used, which is not suitable for CFT*. Besides, it is noteworthy that, different from KinFaceW-I and KinFaceW-II, CFT with Only fully-connected layers trained shows a superiority to CFT with all layers trained. So that, the kin-relation specific features are only associated with high-level layers on UB KinFace and Cornell KinFace datasets.

5.4. Comparison with Deep Algorithms

In Table 2 and Table 3, DKV and SMCNN, as kinship verification methods, are compared. The proposed method outperform them with the following considerations.

- For DKV, stacked auto-encoder (SAE) network instead of CNN is constructed, in which each auto-encoder is trained and stacked together by throwing the decoder part. Additionally, the hand-crafted LBP features instead of raw pixel images are feeded as inputs. Therefore, the high-level

kin-relation specific features cannot be captured.

- For SMCNN, a similarity metric based convolutional structures were designed as CNN does, which contains 8 layers: 4 convolutional layers, 3 pooling layers and 1 fully-connected layer. First, the CNN was trained by using all KinFace training samples and then fine-tuned with image pairs in different kin relation. That is, SMCNN was trained twice with the same KinFace data, and therefore over-fitting may be caused. Additionally, SMCNN depends on KinFace data used, which results in that it cannot be adapted to another KinFace data. From Table 2 and 3, we observe that the performance difference of SMCNN between KinFaceW-I and KinFaceW-II is very significant.

- For CFT, the performance is significantly improved, which demonstrates that transfer learning with a deeper model in coarse-to-fine manner from a large-scale facial images can effectively enhance the kin-relation feature mining ability. The overfitting and robustness to unconstrained Kinship verification can be improved. The following experiments provide further evidence of this view.

5.5. Comparison With Coarse CNN

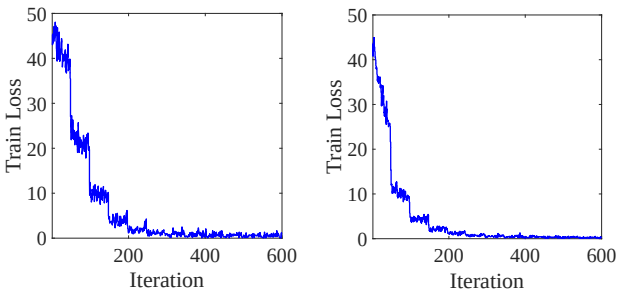
The comparisons on four kinship datasets with coarse CNN (trained over WebFace Only) are presented in Table 6, from which we observe that features extracted from coarse CNN are effective. This demonstrates the feasibility of transfer learning from source domain to target domain. However, CFT methods that integrate a fine CNN show a significant improvement over the coarse CNN. This shows that a small-scale kinship domain data is also important for extracting the high-level deep kin-relation specific features. Therefore, the proposed Coarse-to-Fine transfer learning paradigm is effective.

5.6. Comparison of Different Metrics

For the extracted CFT features, the learning free Euclidean distance (ED) and neighborhood replused metric learning (NRML) have been considered in verification. The results on four datasets are shown in Table 7. We observe that the ad hoc NRML metric outperforms the intrinsic ED metric. This demonstrates that it is useful to design appropriate large margin metric learning method for improving the proposed CFT method. This also motivates us to combine deep learning and conventional learning for more surprising results.

5.7. Convergence

The convergence in fCNN training with two different types (CFT vs. CFT*) is shown in Figure 4, respectively. It is obvious that the convergence rate for all layers training is faster than only fully-connected layer training. This demonstrates that low-level layers are also useful in helping kin-relation specific feature mining. However, for both methods, a good convergence can be achieved.



(a) CFT (b) CFT*

Figure 4. Convergence curve of coarse-to-fine transfer

6. Conclusion

In this paper, we propose a Coarse-to-Fine Transfer Learning method for kinship verification, which is motivated by the transferable advantage of facial knowledge and the kinship data scarcity problem in training a deep model. We aim to explore the learning problem of small-scale data by leveraging another large-scale domain data. Specifically, in CFT, two deep CNN model including cCNN and fCNN are exploited for capturing the high-level and discriminative kin-relation specific semantic features. The coarse cCNN model is trained by leveraging a large-scale face recognition database (i.e. CASIA WebFace) and used for generalized low-level facial features but with weak kin-relation. Then, the fine fCNN model is trained by using a small-scale domain data (i.e. KinFace) based on the cCNN model with two types: Transferring fully-connected layers only (CFT) and transferring all layers (CFT*). Finally, the neighborhood replused metric learning (NRML) is used for verification based on CFT features. Extensive experiments demonstrate that the proposed CFT show the best performance is comparable to the state-of-the-art kinship verification methods. In future, structured deep transfer model such as GAN-based instead of *data-driven* transfer will be studied.

Acknowledgements

This work was supported by the National Science Fund of China under Grants (61771079, 61401048) and the Fundamental Research Funds for the Central Universities (No. 106112017CDJQJ168819).

References

[1] Y. Bengio. Deep learning of representations for unsupervised and transfer learning. *ICML Unsupervised and Transfer Learning*, 27:17–36, 2012.

[2] H. Dibeklioglu, A. Ali Salah, and T. Gevers. Like father, like son: Facial expression dynamics for kinship verification. In *ICCV*, 2013.

[3] A. Esteva, B. Kuprel, R. A. Novoa, J. Ko, S. M. Swetter, H. M. Blau, and S. Thrun. Dermatologist-level classification of skin cancer with deep neural networks. *Nature*, pages 115–118, 2017.

[4] R. Fang, K. D. Tang, N. Snavely, and T. Chen. Towards computational models of kinship verification. In *ICIP*, 2010.

[5] X. Glorot, A. Bordes, and Y. Bengio. Domain adaptation for large-scale sentiment classification: A deep learning approach. In *ICML*, 2011.

[6] K. He, X. Zhang, S. Ren, and J. Sun. Deep residual learning for image recognition. In *CVPR*, 2015.

[7] G. E. Hinton and R. R. Salakhutdinov. Reducing the dimensionality of data with neural networks. *Science*, 313(5786):504–507, 2006.

[8] J. Hu, J. Lu, J. Yuan, and Y.-P. Tan. Large margin multi-metric learning for face and kinship verification in the wild. In *ACCV*, 2014.

[9] N. Jean, M. Burke, M. Xie, W. M. Davis, D. B. Lobell, and S. Ermon. Combining satellite imagery and machine learning to predict poverty. *Science*, 353(6301):790, 2016.

[10] L. Li, X. Feng, X. Wu, Z. Xia, and A. Hadid. Kinship verification from faces via similarity metric based convolutional neural network. In *ICIAR*, 2016.

[11] J. Lu, X. Zhou, Y.-P. Tan, Y. Shang, and J. Zhou. Neighborhood repulsed metric learning for kinship verification. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 36(2):331–345, 2014.

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

[13] S. Ren, K. He, R. Girshick, and J. Sun. Faster r-cnn: Towards real-time object detection with region proposal networks. *Advances in neural information processing systems*, pages 91–99, 2015.

[14] F. Schroff, D. Kalenichenko, and J. Philbin. Facenet: A unified embedding for face recognition and clustering. In *CVPR*, 2015.

[15] M. Shao, S. Xia, and Y. Fu. Genealogical face recognition based on ub kinface database. In *CVPRW*, 2011.

[16] K. Simonyan and A. Zisserman. Very deep convolutional networks for large-scale image recognition. *arXiv preprint arXiv:1409.1556*, 2014.

[17] C. Sun, S. Shetty, R. Sukthankar, and R. Nevatia. Temporal localization of fine-grained actions in videos by domain transfer from web images. In *ACM International Conference on Multimedia*, pages 371–380, 2015.

[18] C. Szegedy, W. Liu, Y. Jia, and P. Sermanet. Going deeper with convolutions. In *CVPR*, 2014.

[19] Y. Taigman, M. Yang, M. Ranzato, and L. Wolf. Deepface: Closing the gap to human-level performance in face verification. In *CVPR*, 2014.

[20] M. Wang, Z. Li, X. Shu, and J. Wang. Deep kinship verification. In *IEEE International Workshop on Multimedia Signal Processing*, pages 1–6, 2015.

[21] Y. Wen, K. Zhang, Z. Li, and Y. Qiao. A discriminative feature learning approach for deep face recognition. *Computers & Operations Research*, 47(9):11–26, 2016.

[22] P. Wu and T. G. Dietterich. Improving svm accuracy by training on auxiliary data sources. In *Proceedings of the twenty-first international conference on Machine learning*, page 110, 2004.

[23] S. Xia, M. Shao, and Y. Fu. Kinship verification through transfer learning. In *IJCAI*, 2011.

[24] S. Xia, M. Shao, J. Luo, and Y. Fu. Understanding kin relationships in a photo. *IEEE Transactions on Multimedia*, 14(4):1046–1056, 2012.

[25] M. Xiao and Y. Guo. Feature space independent semi-supervised domain adaptation via kernel matching. *IEEE Transactions on Pattern Analysis & Machine Intelligence*, 37(1):54–66, 2014.

[26] C. Xiong, X. Zhao, D. Tang, and K. Jayashree. Conditional convolutional neural network for modality-aware face recognition. In *ICCV*, 2015.

[27] H. Yan, J. Lu, W. Deng, and X. Zhou. Discriminative multimetric learning for kinship verification. *IEEE Transactions on Information forensics and security*, 9(7):1169–1178, 2014.

[28] H. Yan, J. Lu, and X. Zhou. Prototype-based discriminative feature learning for kinship verification. *IEEE Transactions on cybernetics*, 45(11):2535–2545, 2015.

[29] 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.

[30] K. Zhang, Z. Zhang, Z. Li, and Y. Qiao. Joint face detection and alignment using multi-task cascaded convolutional networks. *IEEE Journal of Solid-State Circuits*, 23(99):1161–1173, 2016.

[31] L. Zhang, W. Zuo, and D. Zhang. Lsdt: Latent sparse domain transfer learning for visual adaptation. *IEEE Transactions on Image Processing A Publication of the IEEE Signal Processing Society*, 25(3):1177–1191, 2016.

[32] X. Zhou, Y. Shang, H. Yan, and G. Guo. Ensemble similarity learning for kinship verification from facial images in the wild. *Information Fusion*, 32:40–48, 2016.

[33] X. Zhou, H. Yan, and Y. Shang. Kinship verification from facial images by scalable similarity fusion. *Neurocomputing*, 197:136–142, 2016.